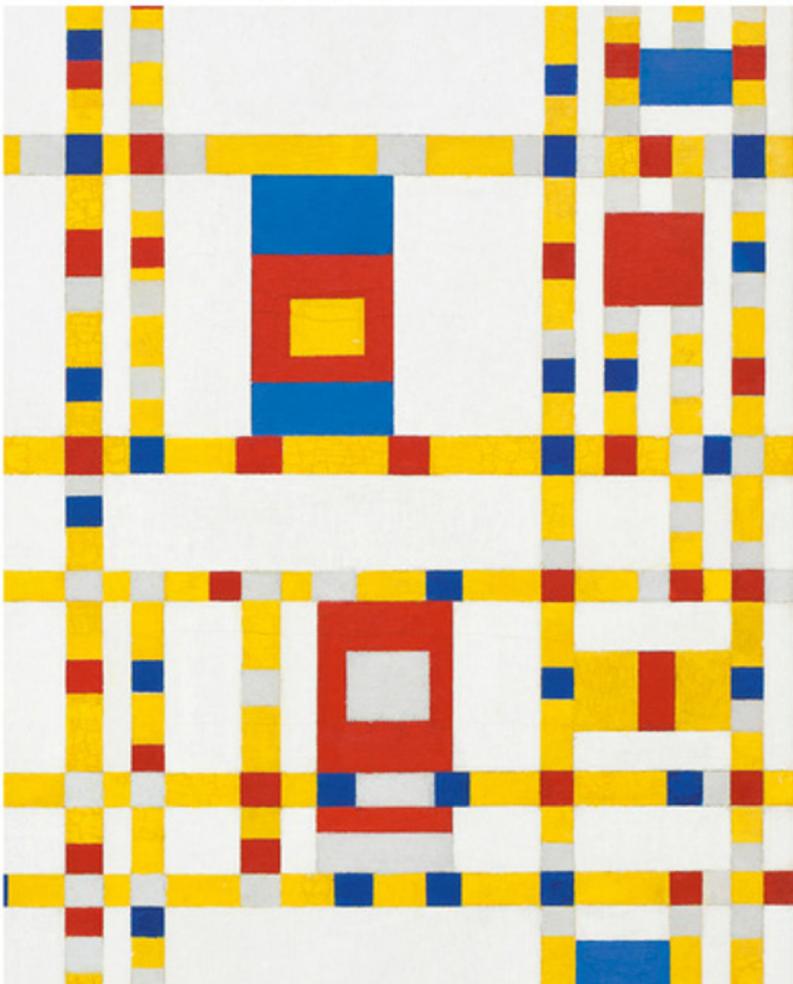


The Handbook of Dialectology



Edited by

**Charles Boberg,
John Nerbonne, and Dominic Watt**

WILEY Blackwell

The Handbook of Dialectology

Blackwell Handbooks in Linguistics

This outstanding multi-volume series covers all the major subdisciplines within linguistics today to offer a comprehensive survey of linguistics as a whole. To see the full list of titles available in the series please visit www.wiley.com/go/linguistics-handbooks

- The Handbook of Child Language*
Edited by Paul Fletcher & Brian MacWhinney
- The Handbook of Phonological Theory, Second Edition*
Edited by John A. Goldsmith, Jason Riggle, & Alan C. L. Yu
- The Handbook of Contemporary Semantic Theory*
Edited by Shalom Lappin
- The Handbook of Sociolinguistics*
Edited by Florian Coulmas
- The Handbook of Phonetic Sciences, Second Edition*
Edited by William J. Hardcastle & John Laver
- The Handbook of Morphology*
Edited by Andrew Spencer & Arnold Zwicky
- The Handbook of Japanese Linguistics*
Edited by Natsuko Tsujimura
- The Handbook of Linguistics*
Edited by Mark Aronoff & Janie Rees-Miller
- The Handbook of Contemporary Syntactic Theory*
Edited by Mark Baltin & Chris Collins
- The Handbook of Discourse Analysis*
Edited by Deborah Schiffrin, Deborah Tannen, & Heidi E. Hamilton
- The Handbook of Language Variation and Change, Second Edition*
Edited by J. K. Chambers & Natalie Schilling
- The Handbook of Historical Linguistics*
Edited by Brian D. Joseph & Richard D. Janda
- The Handbook of Language, Gender, and Sexuality, Second Edition*
Edited by Susan Ehrlich, Miriam Meyerhoff, & Janet Holmes
- The Handbook of Second Language Acquisition*
Edited by Catherine J. Doughty & Michael H. Long
- The Handbook of Bilingualism and Multilingualism, Second Edition*
Edited by Tej K. Bhatia & William C. Ritchie
- The Handbook of Pragmatics*
Edited by Laurence R. Horn & Gregory Ward
- The Handbook of Applied Linguistics*
Edited by Alan Davies & Catherine Elder
- The Handbook of Speech Perception*
Edited by David B. Pisoni & Robert E. Remez
- The Handbook of the History of English*
Edited by Ans van Kemenade & Bettelou Los
- The Handbook of English Linguistics*
Edited by Bas Aarts & April McMahon
- The Handbook of World Englishes*
Edited by Braj B. Kachru, Yamuna Kachru, & Cecil L. Nelson
- The Handbook of Educational Linguistics*
Edited by Bernard Spolsky & Francis M. Hult
- The Handbook of Clinical Linguistics*
Edited by Martin J. Ball, Michael R. Perkins, Nicole Müller, & Sara Howard
- The Handbook of Pidgin and Creole Studies*
Edited by Silvia Kouwenberg & John Victor Singler
- The Handbook of Language Teaching*
Edited by Michael H. Long & Catherine J. Doughty
- The Handbook of Language Contact*
Edited by Raymond Hickey
- The Handbook of Language and Speech Disorders*
Edited by Jack S. Damico, Nicole Müller, & Martin J. Ball
- The Handbook of Computational Linguistics and Natural Language Processing*
Edited by Alexander Clark, Chris Fox, & Shalom Lappin
- The Handbook of Language and Globalization*
Edited by Nikolas Coupland
- The Handbook of Hispanic Sociolinguistics*
Edited by Manuel Díaz-Campos
- The Handbook of Language Socialization*
Edited by Alessandro Duranti, Elinor Ochs, & Bambi B. Schieffelin
- The Handbook of Intercultural Discourse and Communication*
Edited by Christina Bratt Paulston, Scott F. Kiesling, & Elizabeth S. Rangel
- The Handbook of Historical Sociolinguistics*
Edited by Juan Manuel Hernández-Campoy & Juan Camilo Conde-Silvestre
- The Handbook of Hispanic Linguistics*
Edited by José Ignacio Hualde, Antxon Olarrea, & Erin O'Rourke
- The Handbook of Conversation Analysis*
Edited by Jack Sidnell & Tanya Stivers
- The Handbook of English for Specific Purposes*
Edited by Brian Paltridge & Sue Starfield
- The Handbook of Spanish Second Language Acquisition*
Edited by Kimberly L. Geeslin
- The Handbook of Chinese Linguistics*
Edited by C.-T. James Huang, Y.-H. Audrey Li, & Andrew Simpson
- The Handbook of Language Emergence*
Edited by Brian MacWhinney & William O'Grady
- The Handbook of Korean Linguistics*
Edited by Lucien Brown & Jaehoon Yeon
- The Handbook of Speech Production*
Edited by Melissa A. Redford
- The Handbook of Contemporary Semantic Theory, Second Edition*
Edited by Shalom Lappin & Chris Fox
- The Handbook of Classroom Discourse and Interaction*
Edited by Numa Markee
- The Handbook of Narrative Analysis*
Edited by Anna De Fina & Alexandra Georgakopoulou
- The Handbook of English Pronunciation*
Edited by Marnie Reed & John M. Levis
- The Handbook of Discourse Analysis, Second Edition*
Edited by Deborah Tannen, Heidi E. Hamilton, & Deborah Schiffrin
- The Handbook of Bilingual and Multilingual Education*
Edited by Wayne E. Wright, Sovicheth Boun, & Ofelia García
- The Handbook of Portuguese Linguistics*
Edited by W. Leo Wetzel, João Costa, and Sergio Menuzzi
- The Handbook of Linguistics, Second Edition*
Edited by Mark Aronoff and Janie Rees-Miller
- The Handbook of Translation and Cognition*
Edited by John W. Schwieter and Aline Ferreira
- The Handbook of Dialectology*
Edited by Charles Boberg, John Nerbonne, and Dominic Watt

The Handbook of Dialectology

Edited by

*Charles Boberg, John Nerbonne,
and Dominic Watt*

WILEY Blackwell

This edition first published 2018
© 2018 John Wiley & Sons, Inc.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by law. Advice on how to obtain permission to reuse material from this title is available at <http://www.wiley.com/go/permissions>.

The right of Charles Boberg, John Nerbonne, and Dominic Watt to be identified as the authors of the editorial material in this work has been asserted in accordance with law.

Registered Offices

John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

Editorial Office

9600 Garsington Road, Oxford, OX4 2DQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Limit of Liability/Disclaimer of Warranty

While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

Library of Congress Cataloging-in-Publication Data

Names: Boberg, Charles, editor. | Nerbonne, John A., 1951– editor. | Watt, Dominic James Landon, editor.

Title: The handbook of Dialectology / edited by Charles Boberg, John Nerbonne, Dominic Watt.

Description: First edition. | Hoboken, NJ : John Wiley & Sons, Inc., [2017] |

Series: Blackwell Handbooks in linguistics | Includes index.

Identifiers: LCCN 2017017090 (print) | LCCN 2017029760 (ebook) | ISBN 9781118827598 (pdf) |
ISBN 9781118827581 (epub) | ISBN 9781118827550 (cloth) | ISBN 9781119068419 (pbk.)

Subjects: LCSH: Dialectology—Handbooks, manuals, etc. | Language and languages—Variation—Handbooks, manuals, etc.

Classification: LCC P367 (ebook) | LCC P367 .H46 2017 (print) | DDC 417/.2-dc23

LC record available at <https://lccn.loc.gov/2017017090>

Cover image: © 2015. Digital image, The Museum of Modern Art, New York/Scala, Florence

Cover design: Wiley

Set in 9.5/11.5pt Palatino by SPi Global, Pondicherry, India

Contents

List of Contributors	viii
Introduction CHARLES BOBERG, JOHN NERBONNE, AND DOMINIC WATT	1
Section 1: Theory (section editor: Dominic Watt)	17
Section Introduction DOMINIC WATT	
1 Dialectology, Philology, and Historical Linguistics RAYMOND HICKEY	23
2 The Dialect Dictionary JACQUES VAN KEYMEULEN	39
3 Linguistic Atlases WILLIAM A. KRETZSCHMAR, JR.	57
4 Structural Dialectology MATTHEW J. GORDON	73
5 Dialectology and Formal Linguistic Theory: The Blind Man and the Lame FRANS HINSKENS	88
6 Sociodialectology TORE KRISTIANSEN	106
7 Dialectometry HANS GOEBL	123
8 Dialect Contact and New Dialect Formation DAVID BRITAIN	143
9 Dialect Change in Europe—Leveling and Convergence PETER AUER	159
10 Perceptual Dialectology DENNIS R. PRESTON	177
11 Dialect Intelligibility CHARLOTTE GOOSKENS	204
12 Applied Dialectology: Dialect Coaching, Dialect Reduction, and Forensic Phonetics DOMINIC WATT	219

Section 2: Methods (section editor: John Nerbonne)	233
Section Introduction	
JOHN NERBONNE	
13 Dialect Sampling Methods	241
RONALD MACAULAY	
14 The Dialect Questionnaire	253
CARMEN LLAMAS	
15 Written Dialect Surveys	268
J.K. CHAMBERS	
16 Field Interviews in Dialectology	284
GUY BAILEY	
17 Corpus-Based Approaches to Dialect Study	300
BENEDIKT SZMRECSANYI AND LIESELOTTE ANDERWALD	
18 Acoustic Phonetic Dialectology	314
ERIK R. THOMAS	
19 Computational Dialectology	330
WILBERT HEERINGA AND JELENA PROKIĆ	
20 Dialect Maps	348
STEFAN RABANUS	
21 Identifying Regional Dialects in On-Line Social Media	368
JACOB EISENSTEIN	
22 Logistic Regression Analysis of Linguistic Data	384
JOHN C. PAOLILLO	
23 Statistics for Aggregate Variationist Analyses	400
JOHN NERBONNE AND MARTIJN WIELING	
24 Spatial Statistics for Dialectology	415
JACK GRIEVE	
Section 3: Data (section editor: Charles Boberg)	435
Section Introduction	
CHARLES BOBERG	
25 Dialects of British and Southern Hemisphere English	439
KEVIN WATSON	
26 Dialects of North American English	450
CHARLES BOBERG	
27 Dialects of German, Dutch, and the Scandinavian Languages	462
SEBASTIAN KÜRSCHNER	
28 Dialects of French	474
DAMIEN HALL	
29 Dialects of Italy	486
TULLIO TELMON	

30	Dialects of Spanish and Portuguese JOHN M. LIPSKI	498
31	Dialects of the Slavic Languages VLADIMIR ZHOBOV AND RONELLE ALEXANDER	510
32	Dialects of Arabic ENAM AL-WER AND RUDOLF DE JONG	523
33	Dialects in the Indo-Aryan Landscape ASHWINI DEO	535
34	Dialects of Chinese CHAOJU TANG	547
35	Dialects of Japanese TAKUICHIRO ONISHI	559
36	Dialects of Malay/Indonesian ALEXANDER ADELAAR	571
	Index	582

List of Contributors

Alexander Adelaar

The University of Melbourne,
Australia

Ronelle Alexander

University of California, Berkeley, CA, USA

Enam Al-Wer

University of Essex, Colchester, UK

Lieselotte Anderwald

University of Kiel, Germany

Peter Auer

Department of German Linguistics,
University of Freiburg,
Germany

Guy Bailey

University of Texas Rio Grande Valley,
Texas, USA

Charles Boberg

McGill University, Montreal, Canada

David Britain

Department of English, University of Bern,
Switzerland

J.K. Chambers

University of Toronto, Canada

Ashwini Deo

The Ohio State University, USA

Jacob Eisenstein

Georgia Institute of Technology,
USA

Hans Goebel

Department of Romance Studies,
University of Salzburg, Austria

Charlotte Gooskens

Center for Language and Cognition,
University of Groningen, Netherlands

Matthew J. Gordon

Department of English, University of
Missouri, Columbia, Missouri, USA

Jack Grieve

Aston University, UK

Damien Hall

Newcastle University, Newcastle, UK

Wilbert Heeringa

Fryske Akademy, Netherlands

Raymond Hickey

Institute for Anglophone Studies,
University of Duisburg and Essen,
Germany

Frans Hinskens

Meertens Instituut (KNAW) & Vrije
Universiteit Amsterdam, Netherlands

Rudolf de Jong

University of Leiden, Netherlands-Flemish
Institute in Cairo

Jacques Van Keymeulen

Department of Linguistics – Dutch,
Ghent University, Belgium

William A. Kretzschmar, Jr.

Department of English, University of Georgia, Athens, Georgia, USA

Tore Kristiansen

Nordic Research Institute, University of Copenhagen, Denmark

Sebastian Kürschner

Catholic University of Eichstätt-Ingolstadt, Germany

John M. Lipski

Pennsylvania State University, State College, PA, USA

Carmen Llamas

University of York, UK

Ronald Macaulay

Pitzer College, USA

John Nerbonne

University of Groningen, Netherlands & University of Freiburg, Germany

Takuichiro Onishi

National Institute for Japanese Language and Linguistics, Tokyo, Japan

John C. Paolillo

Indiana University Bloomington, USA

Dennis R. Preston

Department of English, Oklahoma State University, Stillwater, Oklahoma, USA

Jelena Prokić

University of Marburg, Germany

Stefan Rabanus

University of Verona, Italy

Benedikt Szmrecsanyi

University of Leuven, Belgium

Chaoju Tang

University of Electronic Science and Technology of China, Chengdu, Sichuan, P.R.China

Tullio Telmon

Università degli Studi di Torino, Torino, Italy

Erik R. Thomas

North Carolina State University, USA

Kevin Watson

University of Canterbury, Christchurch, New Zealand

Dominic Watt

Department of Language and Linguistic Science, University of York, UK

Martijn Wieling

University of Groningen, Netherlands

Vladimir Zhobov

Sofia University, Sofia, Bulgaria

Introduction

CHARLES BOBERG, JOHN NERBONNE, AND
DOMINIC WATT

DIALECTOLOGY is the study of *dialect*, or regional variation in language, a subfield of linguistics. This handbook presents a comprehensive survey of that subfield, including the theory of dialect variation; the methods of collecting, analyzing, and interpreting dialect data; and the facts of dialect variation in many of the world’s most widely spoken languages. Before proceeding with our survey, we offer by way of introduction the following reflections on some of the most basic issues in the field, as well as an explanation of the approach we have taken in planning this book and an outline of what is to follow.

1 The Origins of Dialect Variation and the Status of Dialectology

Dialect differences are caused by two forces operating in tandem: language change and the expansion of speech communities. Language change is of course a constant, on-going process in all speech communities: one of the axioms of historical linguistics is that all languages change all the time. As long as communities remain small, language changes are adopted or rejected by the community as a whole, or show only social differentiation. When a speech community expands sufficiently across a territory, however, the network of interpersonal communication that diffuses changes among its members is disrupted: sheer distance, or physical barriers like mountains and bodies of water—and sometimes also cultural, economic, or social divisions—make it impossible for change to diffuse evenly across the entire community. Eventually, an accumulation of undiffused or partially diffused changes causes community members in one region to recognize that people in other regions speak a different version of their language: what we would call a dialect.

Given enough time, this process of differentiation can cause dialects to diverge to the point where they are no longer wholly mutually intelligible, in which case we begin calling them separate but historically related languages. Such divergence lies at the heart of how historical linguists conceive of the development of families of related languages, like the Indo-European languages spoken across most of Europe and the Americas today, which hypothetically began their individual existence as dialects of a common ancestral language. In other cases, dialect differences can persist in a stable relationship for centuries, without leading to language divergence, or can decline and disappear, as the communication barriers that produced them are overcome by social or technological change. All normal languages, except those spoken in single, restricted locations, display regional variation and have

always done so: accounts of dialect differences are as old as written language itself, appearing two millennia ago in Ancient Greece and China. Given its universality, dialect variation should be seen as a fundamental aspect of human language and dialectology an important branch of linguistics, the scientific study of language. A linguistics that did not include dialectology would be incomplete.

Languages vary in many ways: across time and space, as just discussed, as well as across social categories. Today, dialectology is often seen as part of a larger sub-discipline of linguistics dealing with all of these types of variation, collectively called *language variation and change* (see, e.g., Chambers and Schilling 2013, another handbook in this series). This integrative approach reflects the many ways in which these types of variation have been shown to interact, first brought into clear focus in the work of William Labov (see below). Much of the variation we observe in speech communities is in fact the synchronic manifestation of diachronic processes, or changes in progress: newer forms, before being uniformly adopted, compete for dominance with older forms, in patterns that reflect an intersection of regional and social influences. Nevertheless, as difficult as it can be to isolate regional from other types of variation, the primary focus of this book will be on regional variation.

2 Defining Dialects

We shall begin our discussion of regional variation with just this problem, by exploring the meaning of the word *dialect*, which cannot be properly understood without reference to social variation as well. As linguistic variation arises in speech communities, it usually reflects social differences: different ways of speaking, like different ways of dressing or eating or having fun, come to be associated with groups arrayed on a socio-economic hierarchy involving wealth, power, education, ethnic or social identity, and other factors. Varieties of speech associated primarily with social groups are properly called *sociolects* rather than *dialects* and are the main focus of the allied subfield of *sociolinguistics*, but this type of variation also has an important place in dialectology, since regional varieties of a language—the definition of *dialects* given above—often develop social attributes. In particular, one variety, usually that spoken by the social, economic, and political élite in a nation's capital city or other great metropolis, normally comes to be seen as the “correct” form of the language. In many cases, this evaluation is shared not only by its own speakers, who use it as a symbol and even a justification of their higher social position, but also by others in the community, who accept that their own speech is by comparison inferior, or “incorrect.” Because of its perceived social superiority, the élite variety is promoted to the status of a regional or national “standard” variety, which is preferred or even required in domains like broadcasting, education, government, journalism, the law, literature, liturgy, and science. It often serves these functions not only in its city or region of origin but across the entire linguistic territory, at higher social levels. This establishes a nationwide *diglossia* between the pan-regional “standard” variety, which comes to be seen not as just another dialect but as the unmarked form of the language itself (for instance, the form taught to foreigners who want to learn the language), and the regionally restricted and socially inferior “dialects,” which continue to be the language of everyday life for peasants or farmers in the countryside and for factory workers and trades people in the towns and cities. Rural and urban dialects often receive distinct social evaluations. Rural dialects are frequently seen as quaint and musical, if also unsophisticated and somewhat comic, and are associated with idyllic notions of traditional country life. Urban dialects are more often seen as lazy, ignorant, and linguistically and morally degenerate, since they are associated (at least in many middle-class minds) with the social problems of the lower-class sections of large cities.

An amusing instantiation of this ideology can be seen in the animated adaptation of Kenneth Grahame's children's story *The Wind in the Willows* that was made in the 1980s for Thames Television in the U.K. Though all of the characters are animated figures of animals, the heroes of the story, Rat, Mole, and Badger, speak with subtly different versions of standard British English, or "Received Pronunciation"; the sympathetic minor characters, like a plainspoken otter and a benign cow, have rural, West Country dialects, but the local gang of criminals, the weasels, are given working-class dialects from London ("Cockney") and the urban industrial North. That said, the great fool of the piece, Mr. Toad, the lord of the local manor and a sort of upper-class twit, has the poshest accent of all, reminding us that the correspondence between high-class speech and positive social attributes is not always simple or direct (indeed, not only fools but cads and villains often have upper-class accents in popular entertainment). Nevertheless, the fact that this is a children's program—and a delightful and brilliantly produced one at that, it should be admitted—emphasizes the extent to which dialect ideologies are inculcated in children at a young age by schools, media, and other institutions.

Even more problematic than negative attitudes about dialects is the transfer of such attitudes to the speakers themselves: people who speak what some think of as "lazy" or "ignorant" dialects are thought of as lazy or ignorant themselves, a stereotype that can be used to justify denying them educational, occupational, or social opportunities. Conversely, speakers of standard varieties may be given unfair advantages in the same contexts, a fact that has encouraged many ambitious people from working-class social backgrounds to try to "improve" their speech, often with measurable benefits. This, indeed, is the main justification for teaching standard varieties in schools, whose main purpose is to maximize the socio-economic opportunities of their students. Defenders of the exalted status of standard varieties might argue that they are, in fact, democratizing (or at least meritocratizing) instruments, since they can be learned in school or by other means, thereby conferring socio-economic benefits on the ambitious and becoming a symbol of individual achievement rather than of inherited privilege. Sociolinguists have argued passionately—and correctly—that these notions of superior and inferior dialects are based purely on social prejudice rather than linguistic fact, but they have proven very difficult to dislodge from popular culture, persisting at both ends of the social spectrum (for a critical look at the concepts of "standard" versus "dialect" in English, see Milroy and Milroy (1999) and the contributors to Bex and Watts (1999)).

Not all "dialects," of course, are socially stigmatized, at least not by general consensus. Many non-standard dialects, if they lose points on the "status" dimension that governs access to the most prestigious schools and jobs, gain them on the "solidarity" dimension: their speakers are perceived as friendlier, more attractive, more relaxed, funnier, or more honest than speakers of the standard variety, if not more suitable as surgeons or bank presidents. Other non-standard dialects may be generally disparaged by people outside their own region or social group but are the focus of intense local pride within it. Speakers of these dialects often have a correspondingly negative view of the standard variety and its speakers: as Fischer (1958: 56) observed half a century ago in the pioneer of sociolinguistic studies, "A variant which one man uses because he wants to seem dignified another man would reject because he did not want to seem stiff." Still other non-standard dialects are valued even by speakers of the standard variety as genuinely beautiful or cultured, even if inappropriate for some of the domains reserved to the standard variety.

Moreover, not all regional differences are socially marked. It is easy to think of variables in North American English, for example, that appear to be purely regional, with no common perception that one variant is more correct than the other. This is often true of lexical variation, which juxtaposes forms like *see-saw* and *teeter-totter*, both meaning a tilting board that children play on, or *cottage* and *cabin*, terms for a rural summer vacation home, or *pop* and

soda, generic terms for non-alcoholic carbonated beverages, without social prejudice. Many regional phonological variables, too, seem to lack social symbolism: Americans as a whole have no opinion on whether it is correct to pronounce pairs of words like *cot* and *caught*, or *stock* and *stalk*, differently, as in large sections of the eastern half of the country, or the same, as in most of the western half (where opinions exist, they relate to the phonetic qualities of the vowels involved, not the presence of phonemic contrast). Grammatical variation, by contrast, is more frequently aligned with social factors: everyone in the United States, as well as in other English-speaking countries, knows that “double negatives” and lack of “standard” subject-verb agreement are “wrong” and that those who use them mark themselves as lower-class, a message continually reinforced by schools and other institutions. Potential interactions with social factors, then, are an important aspect of dialect study.

If national standard varieties of languages coexist in a diglossic relationship with dialects of those languages, they also, in the case of multinational languages like English, French and Spanish, coexist with other national standards. In this context, such “standard varieties” are themselves “dialects.” In some cases, as between fellow ex-colonies like the United States and Canada or many Latin American countries, these relationships are fairly egalitarian, with national differences viewed as purely regional rather than social. In other cases, as between ex-colonies and their former colonizers, unequal sociolinguistic status can persist long after political independence. A general equality between the standard varieties of British and American English has now, after two centuries, come to be accepted by many English-speakers, including those in second-language education. Few people in the United States today would consider shifting toward British standards when reading the news on television or teaching English to foreigners. This equality, however, reflects the enormous size, power, and prestige of the United States, which has clearly surpassed that of the mother country. By contrast, relations between standard European French and the ex-colonial varieties of French spoken in Canada or other parts of the former French empire are still more hierarchical, with varieties closer to the European standard preferred in broadcasting and second-language teaching, for example. In some cases, opinion about such matters is regionally divided: while many in Spain consider *castellano*, the standard variety of Iberian Spanish based on the dialect of Castile, a global standard, Latin Americans are less likely to accept this notion and the form of Spanish taught in the United States most commonly follows a Mexican rather than Castilian standard, for instance in failing to preserve the Castilian distinction between *s* and *z* (*casa*, “house,” versus *caza*, “hunts”).

While dialects can differ at every level of structure—phonetic, phonological, morphological, lexical, syntactic, semantic, and so on—the term *dialect* is often used in a complementary relation with another term, *accent*, whereby *dialect* means differences in grammar and lexicon, while *accent* is restricted to phonological and especially phonetic differences, such as the quality of vowel sounds (as in the exhaustive survey of “accents of English” compiled by Wells 1982). This distinction takes on an important social dimension in Britain, for instance, where a three-level structure of language variation was traditionally observed: the national élite, particularly those educated at Oxford and Cambridge Universities, spoke “standard” British English with a “standard” or non-regional accent known as “Received Pronunciation,” regardless of where they lived (at least within England—the Celtic “nations” were to some extent exempt from this standard and had their own regional standards); the urban middle class spoke “standard English” with a regional “accent,” differing from the élite “standard” only in pronunciation, especially of vowel sounds; and the working class, urban and rural, spoke regional “dialect,” with non-standard grammar and lexicon, which also implied a marked regional “accent.” These social distinctions have recently been waning, with a decline in élite use of some traditionally prestigious features now seen as unattractively snobbish and a deliberate promotion of regional accents in domains like national broadcasting (e.g., on the BBC) where they were not previously accepted. Nonetheless, to a large

extent this differentiation can still be heard today and might also be argued to apply increasingly to the United States, where a non-regional “General American English” is pushing out local speech patterns in many regions (see Chapter 26 for examples).

A particularly problematic issue in defining *dialect* has been its taxonomic relation to the term *language*, the latter supposedly comprising a set of mutually intelligible dialects: if two people speak differently but can understand one another, they are speaking dialects of the same language; if they cannot understand one another, they are speaking different languages. It has often been pointed out that popular and even academic ideas about classifying varieties as languages or dialects reflect non-linguistic factors, like political boundaries and cultural history, as much as strictly linguistic criteria of mutual intelligibility. The stock examples in this discussion include, on the one hand, Mandarin and Cantonese, which many people think of as dialects of a single Chinese language but which are not mutually intelligible in speech (see Tang, this volume); and on the other hand, Hindi and Urdu, spoken in India and Pakistan respectively, which many people think of as separate languages but which are in fact largely mutually intelligible, separated more by an international boundary and by the cultural and religious affiliations of their speakers than by any marked linguistic divergence (see Deo, this volume). Europe, too, includes many instances of political boundaries creating and reinforcing “language” differences across what were once gradually shifting continua of local dialects, such as those between Germany and the Netherlands (see Kürschner, this volume), Spain and Portugal (see Lipski, this volume) or parts of the former Soviet Union (see Zhobov and Alexander, this volume); Italy presents a particularly complex blend of regional “dialects” and “languages” that are all offshoots of Latin (see Telmon, this volume).

Mutual intelligibility is itself a hazy concept, of course, involving not the binary distinction implied by the terms *language* and *dialect* but a cline or scale of linguistic similarity. At one end of the scale, we find cases of minimal regional difference with unrestricted mutual intelligibility, as between the major national standard varieties of English or Spanish: middle-class people in London and Los Angeles, or Madrid and Mexico City, recognize clear differences in each other’s speech but have very little difficulty understanding one another, if any at all. At the other end of the scale, we find complete unintelligibility, as between English and Arabic or Mandarin. In the middle, however, are many degrees of partial intelligibility. Some of these involve varieties that differ markedly from the most widely recognized international standards, such as the types of English spoken in Glasgow, Belfast, Appalachia, Jamaica, Singapore, or Nigeria. Others involve closely related “languages,” such as the Scandinavian or Romance languages, which began their histories as dialects of a common ancestral language and still retain a large common grammar and vocabulary, but have since drifted far enough apart to make mutual comprehension difficult, especially in speech. In many of these cases, moreover, the partial intelligibility that does exist is not symmetrical: Danes understand Swedes better than Swedes understand Danes and Portuguese speakers can generally make out more Spanish than vice versa. Intelligibility can be affected by non-linguistic factors like education, exposure, and the comparative social status and population sizes of the languages and cultures involved, as much as by purely linguistic matters like sound change or vocabulary replacement. These problematic issues will be reprised in several chapters of this book (see, i.a., Gooskens on dialect intelligibility).

3 The Origins and Development of Dialectology

From issues surrounding the nature and definition of dialects, let us now turn to a brief review of the history of dialectology, setting the stage for the chapters that follow. While it certainly has precedents in other places and earlier times, the modern “western” tradition of dialectology began in Europe in the nineteenth century. For many of its earliest practitioners,

dialect study was a hobby: an entertaining pastime for self-taught philologists with an interest in cultural history and folkways and a romantic conception of rural life, then very much in fashion (seen also in the literature, music, and painting of the period). Some early dialect collectors, for instance, were parish priests or schoolteachers, who had both a measure of formal education and a strong connection to the local communities they served. Henry Higgins, the fictitious dialect phonetician parodied by George Bernard Shaw in his 1913 play *Pygmalion*, though based partly on real-life phonetician and philologist Henry Sweet (1845–1912), comes off more as a gentleman of leisure with eccentric interests than as the sort of person we would recognize today as a professional academic or serious scholar. Early interest in dialect was given extra urgency by a genuine concern, not altogether unjustified, that rural culture would soon be irretrievably altered or lost as industrialization and urbanization increased. This “curatorial” approach to dialect study sought to record as much of traditional rural speech as possible before it was too late, not unlike the efforts of modern linguists to record and study the thousands of indigenous and minority languages whose vitality is now threatened by digital technology and cultural globalization. Dialects also came to be seen as entertaining by the growing urban bourgeoisie of the nineteenth century, who enjoyed tittering at rustic stereotypes presented in theatres and music halls or in novels. Many of the greatest writers of the time, like the Brontë sisters, Dickens, and Hardy, filled their novels with passages of dialect, not only as a creative device, adding realism to their rural or working-class characters, but as comic entertainment for their largely urban, middle-class readers (a tradition that continues today in film and television).

Over time, however, the subject also developed a more serious, academic side. As will be recounted in several subsequent chapters (see also the general accounts of the development of dialectology in Petyt 1980, Francis 1983, or Chambers and Trudgill 1998), serious academic study of dialects began in Germany, where Georg Wenker and his colleagues carried out a postal survey of dialect variation across the German-speaking territory of Europe, starting in the 1870s (Wenker *et al.* 1927–1956). Wenker asked school teachers to translate a set of 40 sentences into local dialect, as they observed it in their communities; he then collected these records and compiled them in a *Deutscher Sprachatlas*, or “German language atlas,” showing where each form was found and how one region differed from another. This effort was closely followed by the *Atlas linguistique de la France*, published by Jules Gilliéron in several volumes over the first decade of the twentieth century. The French study took a different approach to data collection: it was based instead on face-to-face interviews with dialect speakers carried out in the field, using a standard questionnaire administered by a trained fieldworker (Gilliéron and Edmont 1902–1920).

The ultimate goal of these projects, like many that came after them, was to produce a *dialect atlas* (see Kretzschmar, this volume): a collection of maps showing the regional distribution of linguistic variants—sets of alternate words, pronunciations, or grammatical forms—over a given territory (usually the territory covered by speakers of a single language, or a subdivision of that territory). On these maps, symbols or transcriptions indicated the variants occurring in each location and lines called *isoglosses* could be drawn to divide spatial distributions of variants or mark the outer limit of a distribution; bundles of these isoglosses were taken to indicate major dialect boundaries. This aspect of dialectology is also known by the term *dialect geography*. As the name implies, dialect geographers used their maps to develop geographic interpretations of the spatial distribution of dialect forms, such as the role of barriers to communication, like mountain ranges, in preventing the diffusion of variants and thereby creating dialect divisions, or of channels of communication, like rivers and roads, in encouraging diffusion over wider areas; they also turned to information on cultural and settlement history in their efforts to explain the location of dialect boundaries.

Alongside dialect geography, an allied tradition of *dialect lexicography* also emerged in the nineteenth century, which involved the production of dictionaries of dialect words and

phrases, with definitions, examples, and usage notes, recorded in list form rather than on maps (see Van Keymeulen, this volume). At the turn of the twentieth century in England, for instance, Alexander Ellis published his records of dialect pronunciation, collected two decades earlier (Ellis 1890), and the English Dialect Society produced an *English Dialect Dictionary*, compiled by Joseph Wright (1898–1905).

As the field evolved, the interests of many dialectologists expanded beyond dialect variation itself to include connections with other aspects of language study. For instance, some dialectologists became involved in a debate with linguistic historians over the nature of language change. A group of nineteenth-century historical linguists known as the Neogrammarians had proposed that *sound change*—gradual shifts in the pronunciation of sounds found in sets of words—was a regular and exceptionless process that operated rather like the physical laws of natural science (Osthoff and Brugmann 1878). Systematic sound changes, gradually transforming all of the instances of a given sound simultaneously, were held to be responsible for the linguistic diversification of speech communities. Over thousands of years, this process had given rise to families of “genetically” related languages. These families could be modeled as a tree, whose trunk represented the original “proto-language” and whose branches represented the innovations that distinguished each sub-family and, ultimately, each individual language. The primary nineteenth-century example of this was the Indo-European language family and its Italic, Celtic, Germanic, Slavic, and other branches, which were readily observable in contemporary Europe, with a fascinating extension to Iran and northern India. For instance, the initial /p/ sound of the hypothetical ancestral Indo-European language, evident in Latin words like *pater*, *pe(di)s*, and *piscis*, had become an /f/ in the Germanic languages, producing German *Vater*, *Fuß*, and *Fisch*, or English *father*, *foot*, and *fish*, against French *père*, *pied*, and *poisson* or Italian *padre*, *piede*, and *pesce*, among dozens of other examples. Such regular similarities, called *systematic correspondences*, served as the basis for reconstructing the sound system and vocabulary of the now-extinct proto-language (the three correspondences just mentioned have been reconstructed respectively as **pH₂ter*, **pōds/ped-*, and **pisk-*). Notwithstanding the apparently solid evidence of the Neogrammarians, other linguists saw their theory as an extreme view, which idealized the process of sound change and ignored a great deal of contradictory evidence. Some apparent exceptions to sound change patterns could be resolved by refinements to already formulated rules, but others were more tenacious, perhaps reflecting factors like borrowing, dialect mixture, or social pressures, which were seen as external to the mechanism of sound change but nevertheless posed problems for the theory.

As dialectologists began their survey work, the Neogrammarians initially hoped that the collection of data on traditional rural dialects, which were thought to be free of the complicating impurities of urban speech communities and standardized literary languages, would prove their theory right, by showing systematic and regular application of sound changes. When these data began to be analyzed, however, dialectologists found that they often revealed glaring exceptions to the hoped-for patterns of regular change. In some villages, a mixture of changed and unchanged forms was found, suggesting that some changes, at least, were irregular, affecting some instances of a sound but not others, and that the basic unit of phonological change was the word, not the phoneme, or sound. A classic illustration comes from Dutch: Kloeké (1927) found that local forms of *house* and *mouse*, which both had long /u:/ in Proto-Germanic and should have followed parallel developments according to Neogrammarian theory (as they did in English), displayed different sounds in some Dutch towns, a direct contradiction of the regularity of sound change (see Bloomfield (1933: 328–331) for an influential discussion of these data).

In response to the Neogrammarian dictum that sound change is regular and suffers no exceptions, dialectologists therefore advanced their own, opposite slogan, that “every word has its own history,” apparently denying any sort of regularity to sound change. In its French

form, *chaque mot a son histoire*, this view is usually attributed to Gilliéron (see Gilliéron and Roques (1912)), but it goes further back to Hugo Schuchardt in the nineteenth century and perhaps as far as Grimm (1819: XIV), who says, "...jedes Wort hat seine Geschichte und lebt sein eigenes Leben" ("every word has its history and lives its own life"). The dispute over the regularity of sound change produced a deep cleft between what would become two separate traditions of linguistic thought. The dialectologists accused the Neogrammarians of ignoring the complexity of actual data in their efforts to attain higher levels of generalization and theoretical abstraction, while the Neogrammarians accused the dialectologists of obsessing over minutiae and variability for their own sake, like stamp collectors, without addressing questions of broader scientific significance. This rift is still observable today, in the division between formal theoretical linguistics, which is in some ways the heir of Neogrammarian philology (with other important influences, like the work of Saussure), and the field of language variation and change, including much of modern historical linguistics, sociolinguistics, and dialectology, which carries on the more skeptical or at least more empirical and data-oriented viewpoint of the nineteenth-century dialectologists.

While formal theoretical approaches to the study of language, such as the structuralism of the mid-twentieth century and the generative school of the late twentieth century, came to dominate modern academic linguistics, especially in eastern North America, a robust tradition of work on language variation and change, including dialectology, also continued to thrive, even if it was increasingly sidelined in many prestigious linguistics programs at major universities. By the 1930s, the French method of interviewing dialect speakers in the field and making meticulous records of their speech that could later be transformed into maps was extended to North America by Hans Kurath, who produced a *Linguistic Atlas of New England*, intended to be the first of several regional dialect atlas projects that would eventually cover the entire continent (Kurath *et al.* 1939–1943). This project was sadly never completed but has nevertheless produced a great deal of data and a tradition of work that continues today. In addition to the original New England atlas, the major published atlases of American English now cover the Middle and South Atlantic states (McDavid *et al.* 1980), the Gulf states (Pederson, McDaniel, and Adams 1986–1993) and the Upper Midwest (Allen 1973–1976). Following World War II, the *Survey of English Dialects* published maps of dialect variation across England (Orton and Dieth 1962–1971; see also Upton and Widdowson 1996), and in the 1960s a second major American dialectology project, the *Dictionary of American Regional English*, was undertaken, which is now complete (Cassidy and Hall 1985–2012; see also Carver 1987). Traditional dialect survey work has also continued in many other countries across the globe, as attested in the chapters of Section 3 of this volume.

Dialectology received a new stimulus in the 1960s from the work of William Labov in the closely allied field of *sociolinguistics*, which investigates relationships between linguistic variation and social structure and identity. One of the main concerns of early dialectology, as mentioned above, had been the effects of urbanization, mass education, and other forms of social change on traditional rural dialects, which were feared to be disappearing. A priority of many dialectologists was therefore to collect and study records of these dialects before they were lost. The best exemplars of traditional dialects were thought to be older rural men with minimal formal education and long family histories in the region, who were consequently favored as informants. Comparatively little interest was taken in other types of speakers, who were seen as less representative of "pure dialect," or in cities, which were seen to offer nothing more than chaotic mixtures of modified regional dialects brought in by migrants from the surrounding countryside, or working-class urban varieties that were seen as linguistically and morally corrupt. By the 1960s, these assumptions no longer seemed justified. A new generation of sociolinguists sought to base their descriptions and theories of linguistic variation on the speech of the majority of the population. In the United States, Britain, and other western nations, this majority now lived in cities, where it

comprised not only old men of local stock but women, young people, recent migrants, and a wide range of social classes and ethnic groups, including those who spoke varieties stigmatized as debased, indolent, and ugly.

When Labov began to study urban speech communities, starting with New York City, he found that they displayed not the chaotic dialect mixture dismissed by some dialectologists as uninformative but *orderly heterogeneity*, a pattern in which the probability of occurrence of competing linguistic variants, such as standard and non-standard pronunciations or grammatical forms, depends on a complex yet systematic interplay of many different factors (Labov 1972). These factors included social attributes of speakers, like age, sex, and social class, as well as speech style, or the social context of speech (the identities of interlocutors, the setting and topic of conversation, etc.). Labov's focus on correlations between linguistic and social variables and on shifting frequencies of variants has earned this type of sociolinguistics the names *correlational* or *quantitative* or *variationist sociolinguistics*. Some have called it *urban dialectology* (e.g., Chambers and Trudgill 1998: 54), though of course the sort of variation that Labov and others have studied in major cities can also be found in small towns and rural communities (as Labov himself did on the island of Martha's Vineyard, off the Massachusetts coast), if on a smaller scale reflective of their narrower range of social diversity. Once the focus of interest shifted from a curatorial mission to preserve obsolescent traditional speech varieties to a more objective interest in how language reflects social identity, systematic variability could be found in any speech community; subsequent studies in British cities like Norwich (Trudgill 1974a) and Glasgow (Macaulay 1977) clearly demonstrated that this property of speech communities was not unique to New York City or to the United States. Moreover, communities can be compared simultaneously in the regional and social dimensions: they differ both one from another and within themselves, so that regional comparisons have to take local, community-internal variability into account. For instance, regional divergence may be greater at certain social levels, or among particular ethnic groups. This hybrid approach has been called *social* or *socio-dialectology* (e.g., by Rona 1976; see also Kristiansen, this volume).

Labov's contributions to modern dialectology go beyond shifting the focus to cities. He also pioneered the use of acoustic phonetic analysis to make detailed and reliable measures of vowel quality (Labov, Yaeger, and Steiner 1972; see also Thomas, this volume), which could be used to track the progress and distribution of sound changes that were continually modifying the pronunciation of urban dialects in contemporary American English. This work produced a new hybrid subfield normally called *socio-phonetics*; despite this label, the variation measured and analyzed by these techniques was as much regional as social.

Finally and perhaps most importantly, following on the insights of several predecessors (Weinreich 1954; Moulton 1960, 1962), Labov sought to re-establish connections between dialect study and theoretical linguistics, particularly structural phonemics (see Gordon, this volume). By framing his investigations of linguistic variation in terms of major questions of linguistic theory, such as resolving the Neogrammarian controversy discussed above (Labov 1981), or explaining what kinds of linguistic elements typically get transferred between dialects in contact situations and what kinds need to be learned by children from their parents (Labov 2007), Labov hoped to end what he saw as the intellectual isolation of dialectology, thereby augmenting its scientific value and stature. At the most fundamental level, he argued persuasively that the development of linguistic theory should not be divorced from the close study of data on how language is actually used by real people in real communicative contexts, and must give a satisfactory account of the variability found in these data, rather than dismissing it as optional rule application. He further believed that the study of dialect, in turn, could profit from new insights provided by reference to concepts and questions in general theoretical linguistics, as shown by Chambers (1973) and discussed here by Hinskens (this volume). Though not all dialectologists today feel the need to engage with questions of general linguistic theory, Labov's work has illuminated

many opportunities for those who do; it has transformed the modern study of language variation and change, producing a whole new tradition of dialect study best represented by the *Atlas of North American English* (Labov, Ash, and Boberg 2006), the journal *Language Variation and Change*, and the annual *New Ways of Analyzing Variation* (N WAV) conference, now a major venue for the latest research in dialectology as well as sociolinguistics.

Modern dialectology has seen other advances as well. Where traditional dialectologists had to draw their maps by hand, a laborious and necessarily selective process, today's practitioners benefit from a growing array of computer cartography tools, which support new insights into the regional distribution of dialect variants (see Rabanus, this volume; also Wikle 1997; Lameli, Kehrein, and Rabanus 2010). Another group of scholars has been concerned with developing objective measures of dialect difference that rise above anecdotal accounts and avoid selective analysis, a subfield known as *dialectometry*, which has incorporated sophisticated and rigorous quantitative methods from other social sciences and from statistics and computer engineering (Goebl and Nerbonne and Wieling, this volume; also Goebl 1982; Kretzschmar 1996; Boberg 2005; Nerbonne and Kleiweg 2007; Nerbonne and Kretzschmar 2006; Nerbonne 2009; Grieve, Speelman, and Geeraerts 2011).

Still other dialectologists have focused on the nature of the borders between dialect regions (Watt, Llamas, and Johnson 2010; Watt and Llamas 2014). These have often been observed to be "fuzzy," involving *transition zones* in which the features of neighboring dialects are commingled, with a gradual shift from one dialect to the next, rather than sharp, with a sudden and easily perceptible change in speech at a specific location. Related to these topics are studies of the rise of dialect continua across stretches of terrain (Heeringa and Nerbonne 2001), of contact between dialects (Trudgill 1986), of the transitional forms that arise from this contact (Britain 1997; Chambers and Trudgill 1998), of the spatial diffusion of linguistic elements from one dialect to another (Trudgill 1974b; Callary 1975; Bailey *et al.* 1993; Auer, Hinskens, and Kerswill 2005) and of the rise of new dialects created by migration and dialect mixture (Kerswill and Williams 2005).

Another recent trend in dialect study has been to turn from production to perception, by examining what ordinary people think about the dialect diversity that surrounds them (Preston, this volume; Preston and Long 1999); how dialect differences interfere with cross-dialectal intelligibility (Labov and Ash 1997); or how listeners categorize speakers by dialect region and which features these categorizations rely on (Clopper and Pisoni 2004).

Finally, dialectology, like all fields of study, has not been immune to the recent influence of the internet, which presents new opportunities (and challenges) for data collection and analysis, dramatically increasing both the quantity of data available to researchers and the speed at which these data can be collected and analyzed (e.g., Eisenstein, O'Connor, Smith, and Xing 2010; Grieve, Asnaghi, and Ruette 2013). Technological advances have also contributed to the creation and analysis of large, searchable corpora of data, as discussed by Szemrecsanyi and Anderwald (this volume), allowing conclusions about variation and change to be drawn from ever larger sets of data. In particular, internet searches make possible the rapid collection of vast quantities of data on regional variation in ordinary language—most commonly written language but also speech—as opposed to language deliberately collected for the purposes of study. This is a potentially transformative change that minimizes the gap between dialectology and the variation it attempts to study.

4 The Present and Future State of Dialectology

Studies of diffusion, which includes the spread of features not only from one region to another but from one social group to another, give rise to the question of whether the advent of mass education, personal mobility, and instant communication in modern, industrialized

nations threatens the very survival of dialects, echoing the original concerns of nineteenth-century dialectologists (Britain 2009; Kristiansen 1998). Insofar as many traditional rural and non-standard urban dialects are now declining or disappearing, this implies a gradual contraction of the subject matter of dialectology, which might suggest a pessimistic view of the future of the field. On the other hand, many older dialects are sustained by a strong force of local identity that prevents their decline. Even in fully industrialized or post-industrial countries like Britain and Germany, some distinctive regional dialects, like those in the North of England (e.g., Tyneside, Yorkshire, or Lancashire English) or the South of Germany (e.g., Bavarian or Swabian German), continue to be spoken enthusiastically by millions of people and show little sign of disappearing anytime soon, even if their features are constantly modified by contact with standard and non-local speech. In other cases, like that of Denmark, which saw traditional dialects virtually disappear over the twentieth century (Kristensen and Thelander 1984; Pedersen 2003), local identity manifests itself today in subtler forms of variation, sustaining small regional differences in an otherwise homogeneous supra-regional or national type of speech that diffuses from cosmopolitan centers (Kristiansen 1998). Boberg (2005, 2008, 2010) observes a similarly fine-grained yet tenacious regional differentiation, which we might call *micro-variation*, in Canadian English, which is otherwise reputed to be remarkably homogenous over the country's vast territory.

Moreover, if we look beyond traditional dialects, we find a proliferation of new dialects constantly emerging. This is true in many large urban centers, for example, whose populations are increasingly diverse: distinct ethnic and cultural subgroups in cities like London (Cheshire *et al.* 2011), Berlin (Wiese 2009), Stockholm (Kotsinas 1988) and Copenhagen (Quist 2000) mark their emerging social identities in linguistic as well as other ways, though intra-community variation of this kind is strictly speaking more a concern of sociolinguistics than of dialectology.

An even more important source of new dialects, however, is the most consequential linguistic development in the world today: the rise of English as a global language. This phenomenon, which has generated a whole new subfield of language variation and change called *World Englishes*, with its own conferences and journals, fundamentally involves the creation and development of dozens of new dialects of English (Crystal 2003; Schneider 2007; Kachru, Kachru, and Nelson 2009; Melchers and Shaw 2013). These were spoken at first in countries with historical ties to the former British Empire, where many native yet distinctive varieties of English are to be found, usually in multilingual settings (e.g., India, Ghana, Nigeria, Hong Kong, and Singapore; the "outer circle" of Kachru (1985)). Today, they are increasingly flourishing in countries with no such connection, where English has no historical status but is now used as a lingua franca for intercultural and international communication (Kachru's "expanding circle"), especially in domains such as advertising, digital media, diplomacy, high technology, international commerce, international sport, popular entertainment, post-secondary education, scientific research, and of course tourism. In these contexts, new, non-native varieties of English, which may become native varieties in the future, exhibit distinctive features that reflect the native languages with which they coexist.

In Europe, for instance, where English has become the *de facto* working language of the European Union and is widely learned as a second language in primary school by children across the continent, the point at which future generations will regard Dutch English, Swedish English, even German English as legitimate dialects of English, spoken by bilingual populations, seems less remote every year (after all, Irish, Scottish, and Welsh English, now universally accepted as dialects of English, also began as second-language varieties, spoken by Britain's Celtic populations). Assuming this scenario persists (which should not be taken for granted), a similar evolution of Russian English, Chinese English, and Japanese English dialects may not be far behind. Once these originally second-language varieties become semi-native, mutually intelligible regional types of English, they enter the legitimate domain of dialectology; from this perspective, the future of the field looks bright (if only for English dialectology!).

In short, the field of dialectology has grown and adapted in many ways and continues to respond to a changing environment today. While it no longer holds the central position in linguistic science that it enjoyed in the nineteenth century, it nevertheless remains a dynamic and relevant sub-discipline that continues to produce new scholarly work and attract new generations of students. From its origins in Europe it has now spread across the globe, with dialect studies available or in progress on languages spoken in every region of the world, as seen in Section 3 of this volume. Yet, despite all of these changes, most dialectologists today continue to focus on the central questions that gave rise to the field over a century ago:

1. How do languages vary across the territories in which they are spoken?
2. What are the common patterns in this variation, including the linguistic constraints that govern it, viewed across different languages?
3. How do settlement history, topography, social patterns, urbanization, and other non-linguistic factors explain the spatial distribution of linguistic features?
4. What is the nature of the transitions or boundaries between spatial distributions?
5. How do innovative features spread across new territory?
6. Is regional variation receding, stabilizing, or increasing over time?

Despite their long history, all of these questions remain relevant today, as they are addressed with the new methods described above and with new data, both from new communities and from previously studied communities that continue to change.

5 Rationale and Plan of This Book

Given the recent expansion and diversification of dialectological scholarship reviewed above, contemporary students of dialect at all levels of expertise now face a significant challenge. They must keep up with technical and theoretical advances in a wide range of different sub-disciplines, as well as with a constantly growing body of data on dialect variation in a wide range of languages. Moreover, as the demands of assimilating all of this new information grow heavier, it becomes more difficult—yet no less important—to maintain an intellectual connection between contemporary research and the scholarly achievements of the past. Foundational work should always be taught and re-taught as an underpinning for modern research, but also critically re-evaluated in light of new information and alternative, innovative thinking.

In light of these challenges, the field has become far too large for even the most senior and widely experienced scholar to have more than a passing acquaintance with all of its various sub-divisions, let alone for the junior scholar or beginning student who wishes to progress beyond the surveys available in short introductory monographs suitable for undergraduate courses. Yet those with an interest in developing a broad knowledge of dialectology, or in having access to such knowledge on an as-needed basis, have been faced with making their own surveys of a very large and in some cases inaccessible corpus of materials, which is simply not available in many libraries. Dialect atlases in particular are expensive and space-consuming luxuries found only in large or specialized collections, while much of the original work on non-English dialects, particularly that written before the late twentieth century, has not been translated out of its original languages and cannot be read by most English-speaking students even where it is available.

The present volume therefore seeks to provide both experienced practitioners and their apprentices with an overview of the field of dialectology—past and present—comprising three main aspects of the topic: principal theoretical approaches, methodological traditions, and sets of data. This trio of topics provides the main organizational basis for the book, reflected in the

three sections of the table of contents, each comprising 12 chapters. Because dialectology is and always has been fundamentally a data-driven field, committed to empirical investigation more than to theoretical speculation, or rather to basing the development of theory firmly on competently collected and analyzed sets of data, the methods of dialectology and the data of dialectology are just as important to any review of the field as the various aspects of dialectological theory. Accordingly, whereas the book's first section gives a detailed account of the historical and contemporary development of dialectological thinking, including crucial concepts like the dialect dictionary, the dialect atlas, and the various interfaces with other areas of linguistics and non-linguistic sciences discussed above, the second section is concerned entirely with methodological matters—how dialect data are collected and analyzed—and the third with the data themselves, illustrated with descriptive overviews of dialect variation in the world's most widely spoken languages and language families, particularly those that have produced the richest traditions of dialect study. Primary editorial responsibility for each of these sections was assigned to one of the three co-editors of the volume: Watt oversaw the section on theory, Nerbonne that on method, and Boberg that on data. Beyond this co-written general introduction, each section editor provides a specialized introduction to his section, which introduces and discusses the chapters it contains in more detail than is possible or appropriate here.

REFERENCES

- Allen, Harold B. 1973–1976. *The Linguistic Atlas of the Upper Midwest*. 3 vols. Minneapolis: University of Minnesota Press.
- Auer, Peter, Frans Hinskens, and Paul Kerswill (eds). 2005. *Dialect Change: Convergence and Divergence in European Languages*. Cambridge, UK: Cambridge University Press.
- Bailey, Guy, Tom Wikle, Jan Tillery, and Lori Sand. 1993. Some patterns of linguistic diffusion. *Language Variation and Change* 5: 359–390.
- Bex, Tony and Richard J. Watts (eds). 1999. *Standard English: The Widening Debate*. London: Routledge.
- Bloomfield, Leonard. 1933. *Language*. Chicago: University of Chicago Press.
- Boberg, Charles. 2005. The North American Regional Vocabulary Survey: Renewing the study of lexical variation in North American English. *American Speech* 80/1: 22–60.
- Boberg, Charles. 2008. Regional phonetic differentiation in Standard Canadian English. *Journal of English Linguistics* 36/2: 129–154.
- Boberg, Charles. 2010. *The English Language in Canada: Status, History and Comparative Analysis*. Cambridge, UK: Cambridge University Press.
- Britain, David. 1997. Dialect contact and phonological reallocation: 'Canadian Raising' in the English Fens. *Language in Society* 26: 15–46.
- Britain, David. 2009. One foot in the grave? Dialect death, dialect contact, and dialect birth in England. *International Journal of the Sociology of Language* 2009: 196–197.
- Callary, Robert E. 1975. Phonological change and the development of an urban dialect in Illinois. *Language in Society* 4: 155–169.
- Carver, Craig M. 1987. *American Regional Dialects: A Word Geography*. Ann Arbor, MI: University of Michigan Press.
- Cassidy, Frederic Gomes, and Joan Houston Hall (eds). 1985–2012. *Dictionary of American Regional English*. Cambridge, MA: Harvard University Press.
- Chambers, J.K. 1973. Canadian raising. *Canadian Journal of Linguistics* 18: 113–135.
- Chambers, J. K. and Natalie Schilling (eds). 2013. *The Handbook of Language Variation and Change, 2nd Edition*. Wiley-Blackwell.
- Chambers, J.K. and Peter Trudgill. 1998. *Dialectology*, 2nd ed. Cambridge, UK: Cambridge University Press.
- Cheshire, Jenny, Paul Kerswill, Sue Fox, and Eivind Torgersen. 2011. Contact, the feature pool and the speech community: The emergence of Multicultural London English. *Journal of Sociolinguistics* 15/2: 151–196.
- Clopper, Cynthia G., and David B. Pisoni. 2004. Some acoustic cues for the perceptual categorization of American English regional dialects. *Journal of Phonetics* 32: 111–140.

- Crystal, David. 2003. *English as a Global Language*. Cambridge, UK: Cambridge University Press.
- Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. *Proceedings of Empirical Methods on Natural Language Processing*, 2010: 1277–1287.
- Ellis, Alexander John. 1890. *English Dialects, Their Sounds and Homes: Being an Abridgment of the Author's Existing Phonology of English Dialects*. London: K. Paul, Trench, Trübner & Co.
- Fischer, John L. 1958. Social influences on the choice of a linguistic variant. *Word* 14: 47–56.
- Francis, W. Nelson. 1983. *Dialectology: An Introduction*. London: Longman.
- Gilliéron, Jules, and Edmond Edmund. 1902–1920. *Atlas linguistique de la France*. Paris: Champion.
- Gilliéron, Jules, and Mario Roques. 1912. *Étude de géographie linguistique*. Paris: Champion.
- Goebl, Hans. 1982. *Dialektometrie; Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie*. Vienna: Verlag der Österreichischen Akademie der Wissenschaften.
- Grieve, Jack, Costanza Asnaghi, and Tom Ruette. 2013. Site-restricted web searches for data collection in regional dialectology. *American Speech* 88/4: 413–440.
- Grieve, Jack, Dirk Speelman, and Dirk Geeraerts. 2011. A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change* 23: 193–221.
- Grimm, Jacob. 1819. *Deutsche Grammatik, 1. Theil*. Göttingen: Dieterich.
- Heeringa, Wilbert, and John Nerbonne. 2001. Dialect areas and dialect continua. *Language Variation and Change* 13: 375–400.
- Kachru, Braj B. 1985. Standards, codification, and sociolinguistic realism: The English language in the outer circle. In Randolph Quirk and Henry Widdowson (eds), *English in the World: Teaching and Learning of Language and Literature* (Cambridge, U.K.: Cambridge University Press), pp. 11–30.
- Kachru, Braj B., Yamuna Kachru, and Cecil L. Nelson (eds). 2009. *The Handbook of World Englishes*. Malden, MA: Blackwell Publishing Limited.
- Kerswill, Paul, and Ann Williams. 2005. New towns and koineization: Linguistic and social correlates. *Linguistics* 43: 1023–1048.
- Kloeke, G.G. 1927. *De Hollandsche expansie in de zestiende en zeventiende eeuw en haar weerspiegeling in de hedendaagsche Nederlandsche dialecten*. Den Haag: Martinus Nijhoff.
- Kotsinas, Ulla-Britt. 1988. Immigrant children's Swedish—a new variety? *Journal of Multilingual and Multicultural Development* 9/1–2: 129–140.
- Kretzschmar, William A., Jr. 1996. Quantitative areal analysis of dialect features. *Language Variation and Change* 8: 13–39.
- Kristensen, Kjeld, and Mats Thelander. 1984. On dialect levelling in Denmark and Sweden. *Folia Linguistica* 18(1–2): 223–246.
- Kristiansen, Tore. 1998. The role of standard ideology in the disappearance of the traditional Danish dialects. *Folia Linguistica* 32/1–2: 115–129.
- Kurath, Hans, Miles Hanley, Bernard Bloch, and Guy S. Loman, Jr. 1939–1943. *Linguistic Atlas of New England*. Providence: Brown University Press.
- Labov, William. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.
- Labov, William. 1981. Resolving the Neogrammarian controversy. *Language* 57: 267–308.
- Labov, William. 2007. Transmission and diffusion. *Language* 83/2: 344–387.
- Labov, William, Malcah Yaeger, and Richard Steiner. 1972. *A Quantitative Study of Sound Change in Progress*. Philadelphia: U.S. Regional Survey.
- Labov, William, and Sharon Ash. 1997. Understanding Birmingham. In Cynthia Bernstein, Thomas Nunnally and Robin Sabino (eds), *Language Variety in the South Revisited* (Tuscaloosa, AL: University of Alabama Press), 508–573.
- Labov, William, Sharon Ash, and Charles Boberg. 2006. *The Atlas of North American English: Phonetics, Phonology and Sound Change*. Berlin: Mouton de Gruyter.
- Lameli, Alfred, Roland Kehrein, and Stefan Rabanus (eds). 2010. *Language and Space: An International Handbook of Linguistic Variation*. Vol. 2: *Language Mapping*. Berlin: Walter de Gruyter.
- Macaulay, Ronald K.S. 1977. *Language, Social Class, and Education: A Glasgow Study*. Edinburgh: Edinburgh University Press.
- McDavid, Raven I., et al. 1980. *Linguistic Atlas of the Middle and South Atlantic States*. Chicago: University of Chicago Press.
- Melchers, Gunnel, and Philip Shaw. 2013. *World Englishes*. London: Routledge.
- Milroy, James, and Lesley Milroy. 1999. *Authority in Language: Investigating Standard English*. London: Routledge.

- Moulton, William G. 1960. The short vowel systems of Northern Switzerland: A study in structural dialectology. *Word* 16: 155–182.
- Moulton, William G. 1962. Dialect geography and the concept of phonological space. *Word* 18: 23–32.
- Nerbonne, John. 2009. Data-driven dialectology. *Language and Linguistics Compass* 3/1: 175–198.
- Nerbonne, John, and Peter Kleiweg. 2007. Toward a dialectological yardstick. *Journal of Quantitative Linguistics* 14: 148–166.
- Nerbonne, John, and William A. Kretzschmar, Jr. 2006. Progress in dialectometry: Toward explanation. *Literary and Linguistic Computing* 21: 387–397.
- Orton, Harold, and Eugen Dieth. 1962–1971. *Survey of English Dialects: Basic Materials*. Leeds: E. J. Arnold & Son.
- Osthoff, Hermann, and Karl Brugmann. 1878. *Morphologische Untersuchungen auf dem Gebiete der indogermanischen Sprachen*. Leipzig: S. Hirzel.
- Pedersen, Inge Lise. 2003. Traditional dialects of Danish and the de-dialectalization 1900–2000. *International Journal of the Sociology of Language* 159: 9–28.
- Pederson, Lee, Susan L. McDaniel and Carol M. Adams (eds). 1986–1993. *Linguistic Atlas of the Gulf States*. 7 vols. Athens, GA: University of Georgia Press.
- Petyt, K.M. 1980. *The Study of Dialect*. London: Andre Deutsch.
- Preston, Dennis R., and Daniel Long. 1999. *Handbook of Perceptual Dialectology*. Amsterdam: Benjamins.
- Quist, Pia. 2000. Ny københavnsk ‘multietnolekt’: Om sprogbrug blandt unge i sprogligt og kulturelt heterogene miljøer [A new Copenhagen ‘multi-ethnolect’: Language use among adolescents in linguistically and culturally heterogeneous settings]. *Danske Talesprog* 1: 143–211.
- Rona, José Pedro. 1976. The social dimension of dialectology. *International Journal of the Sociology of Language* 9: 7–22.
- Schneider, Edgar W. 2007. *Postcolonial English: Varieties around the World*. Cambridge, U.K.: Cambridge University Press.
- Trudgill, Peter. 1974a. *The Social Differentiation of English in Norwich*. Cambridge, UK: Cambridge University Press.
- Trudgill, Peter. 1974b. Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. *Language in Society* 3: 215–246.
- Trudgill, Peter. 1986. *Dialects in Contact*. Oxford, UK: Blackwell.
- Upton, Clive, and J.D.A. Widdowson. 1996. *An Atlas of English Dialects*. Oxford, UK: Oxford University Press.
- Watt, Dominic, and Carmen Llamas (eds). 2014. *Borders and Identities*. Edinburgh: Edinburgh University Press.
- Watt, Dominic, Carmen Llamas and Daniel Ezra Johnson. 2010. Levels of linguistic accommodation across a national border. *Journal of English Linguistics* 38: 270–289.
- Weinreich, Uriel. 1954. Is a structural dialectology possible? *Word* 10: 388–400.
- Wenker, Georg, Ferdinand Wrede, Walther Mitzka, and Bernhard Martin. 1927–1956. *Der Sprachatlas des deutschen Reichs*. Marburg: Elwert.
- Wells, J.C. 1982. *Accents of English*. Cambridge, U.K.: Cambridge University Press.
- Wiese, Heike. 2009. Grammatical innovation in multiethnic urban Europe: New linguistic practices among adolescents. *Lingua* 119: 782–806.
- Wikle, Tom. 1997. Quantitative mapping techniques for displaying language variation and change. In Cynthia Bernstein, Thomas E. Nunnally and Robin Sabino (eds.), *Language Variety in the South Revisited* (Tuscaloosa, AL.: University of Alabama Press), 417–433.
- Wright, Joseph (ed.). 1898–1905. *The English Dialect Dictionary*, 6 vols. Oxford: Oxford University Press.

Section 1 – Theory Introduction

DOMINIC WATT

The first group of chapters in this book is devoted to matters of theory. We will, on the one hand, consider theories that have been elaborated within the field of dialectology itself. On the other, we will evaluate the contributions made to dialectology by theories from other domains of linguistics. Some of the latter set of theories have had a significant impact on the ways in which we conceptualize the notion of dialect, as well as how we describe dialects and classify them into superordinate groupings with respect to the languages of which dialects are said to be subspecies. We will see how dialectology and philology have contributed in their turn to theories of language, even in the case of theories whose exponents do not acknowledge, or may not even be aware of, that contribution. The following chapters will make it clear that linguistics as we recognize it today would probably be markedly different had so many paths in synchronic and diachronic language study not first been broken by scholars who devoted their careers to the study of dialect variation and relatedness, initially among European languages but later among languages from all the inhabited continents.

It has sometimes been claimed that dialectology is pre-theoretical, or atheoretical. Although the assertion that any sort of dialectology could be carried out in a theoretical vacuum could not be easier to counter, is not hard to see why the belief persists. Until fairly recently, the professed mission of many dialectologists was simply to describe and catalog linguistic phenomena, rather than to try also to explain the genesis of these phenomena and the mechanisms by which they are transmitted from speaker to speaker across space and time. It was seen by many practitioners of dialectology to be more pressing in the immediate term to observe speech and language as they were actually being used out there in the world than it was to seek profound truths about the nature of language from more abstract philosophical perspectives. Translating this stance into action was to a significant degree motivated by a growing awareness and concern that the traditional dialects were vanishing, and that a failure to document them before they disappeared would mean irrevocably walling up a window onto the past. The prospect of imminently losing access to traditional forms of speech that could link us more directly to the world of our distant ancestors provided a powerful spur. It inspired scholars, individually or in teams, to go to huge lengths to set down, systematically and scrupulously, the rich detail of lexical, phonetic, phonological, morphological, and syntactic variation in the speech and language of people who were otherwise marginalized because of their lack of education, sophistication, or “breeding.” Dialectology gave, as never before, a voice to untutored rural dwellers, those seen as uncorrupted by modernity and urban life, and whose language was as untainted as it could be by the effects of the pressure to conform to institutionally imposed linguistic standards.

The erstwhile dialectological focus on the *what*, rather than the *why* and *how*, might be likened to large-scale data-gathering programs in disciplines such as zoology, botany, or astronomy. Cataloging the diversity of insects, flowering plants, galaxies, or exoplanets that have not yet been named or classified is a vital first step in understanding how the systems in which those entities operate are structured and how they function. It would be rash to assume that our theories of how those systems work are watertight until we have sampled the universe of variation as exhaustively as our finite resources will permit. Yet there are linguistic theorists who would argue that from the descriptive point of view, we essentially know all we need to know about variability in certain languages, and that dialect diversity in those languages is in any case largely irrelevant to the central enterprise of the field. It is perfectly valid, they would argue, to make pronouncements about the grammatical properties of an entire language by examining just one of its dialects—which is almost always the standard variety, if one exists—and not consulting any speakers at all. Why would one go to all the trouble of asking other people about their language, if one can use one's own native-speaker intuitions as a source of data, or get the information one needs from published sources?

Dialectology, according to views of the above kind, has little to offer “serious linguistics.” The dialectologists’ preoccupation with regional and social variation in language production has been viewed as a harmless enough trait, but spending even part of a career studying this aspect of performance (which I have heard dismissed, in paraphrase of John McCarthy’s aphorism, as mere “froth on the surface of language”) has hitherto generally not been thought a worthy pursuit for the theoretician. The proper subject matter of linguistics, on this view, is the set of abstract principles that govern how sentences are constructed, or the constraints that determine how phonological units such as syllables or feet or tonemes may be strung together. The focus on lexis in many dialectological surveys bolstered a perception that dialectologists and philologists were not really concerned with phenomena in the phonological or grammatical domains, and given that in many cases, the questionnaires designed to elicit dialect lexis dwelt on terminology pertaining to occupations such as agriculture, animal husbandry, or traditional arts and crafts, it is easy to see why many researchers working in other subareas of linguistics formed the impression that the methods and underlying conceptual framework of dialectology had stalled decades earlier. Accusations of “golden ageism” were also frequently leveled at dialectologists, probably owing to a tendency in some quarters for ideologies concerning the supposed purity or uninterrupted lineages of traditional dialects to bias scholars in such a way that they downplayed the influence of relevant societal factors on the historical trajectories of languages and their dialects (e.g., marginalizing the evidence suggesting that geographical and social mobility were considerably more prevalent in former centuries than we often assume, such that we tend to believe today that mobility is a modern phenomenon; see e.g., Long 2013; Whyte 2000; also Milroy 1992).

The criticisms of dialectology discussed above are not wholly without foundation. But the field has moved on a great deal in recent decades. New theoretical insights and methodologies from cognate disciplines such as sociolinguistics, human geography, and social and evolutionary psychology (e.g., Buchstaller and Alvanides 2013; Cohen 2012) are helping contemporary dialectology to flourish, and the findings of studies in archaeology and population genetics (e.g., Heggarty *et al.* 2005; Winney *et al.* 2012), along with innovations in mapping and ‘Big Data’ analysis techniques (e.g., Huang *et al.* 2015; Grieve, this volume), are being fused with dialectological data in ways that mutually strengthen all of the contributing disciplines. It is of benefit to linguists of every stripe that new theories of language evolution, acquisition, structure, and use are crystalizing out of the interplay among these diverse approaches to human characteristics and behaviors.

We commence the Theory section of this volume by taking the long view of dialectology’s emergence from the earlier philological tradition, and how historical linguistics simultaneously grew as a parallel offshoot, with all three disciplines cross-fertilizing one another.

In his chapter, Raymond Hickey argues that the boundaries of dialectology are a matter of terminological convention rather than the result of the field having intrinsically sharp edges, and that the scope of dialectology is in any case subject to change as interconnections with other disciplines are forged. Similar perspectives are offered by Jacques Van Keymeulen in his discussion of the history and current status of the dialect dictionary (Ch. 2). The advent in recent years of searchable large-scale multimedia repositories of dialectal data has transformed the dialect dictionary from a static, cut-and-dried record of painstaking scholarship carried out over years or decades, to a dynamic “living” resource that can be regularly updated with new research findings and data gathered from the general public using innovative methods such as crowdsourcing apps (e.g., Goldman *et al.* 2014).

The potential for producing graphical depictions of data held in textual or numerical form by harnessing the power of mapping techniques, in particular contemporary cartographical software, is explored by Bill Kretzschmar, whose chapter (Ch. 3) deals with the evolution and future of the linguistic atlas. Theoretical claims in support of Kretzschmar’s contention that speech and language represent complex, non-linear, scale-free fractal systems become ever more testable in the era of prodigiously large datasets of the kind collected for linguistic atlas projects. Several of the themes Kretzschmar develops are reprised in the fourth chapter, on structural dialectology, by Matthew Gordon. In his exposition of the influences that the structuralist tradition in linguistics and dialectology have had upon one another, Gordon points to the design properties of the *Atlas of North American English* (ANAE; Labov, Ash, and Boberg 2006), arguing that ANAE instantiates many of the precepts of the American and European structuralist approaches to phonology. For example, ANAE organizes vowels into subclasses (long/short, ingliding/upgliding, and so forth), which are then used as a means of bundling North American regional dialects into supercategories on the basis of how vowel types are realized phonetically and how their qualities alter in response to homeostatic pressures of the sort that regulate chain shifting (see further Docherty and Watt 2001; Gordon 2013).

The impact upon dialectology made by the revolution in linguistic thinking brought about by the Chomskyan generativist paradigm that largely supplanted the American structuralist school in the 1950s (see relevant entries in Allan 2013; but cf. Salmons and Honeybone 2015) is examined in detail by Frans Hinskens (Ch. 5). Hinskens is at pains to point out that the relationship between dialectology and “formal linguistic theory” is by no means a one-way street; by analogy with the fable of the blind man and the lame, who pool their complementary faculties so as to overcome their individual disabilities, Hinskens contends that mainstream linguistics owes as much to dialectology as dialectology does to mainstream linguistics. Measures to ensure the continuing healthy symbiosis of dialectology with another of its sister fields—this time variationist sociolinguistics—are endorsed by Tore Kristiansen in his chapter on sociodialectology, this field of inquiry being the product of what happened, as he puts it, “when dialectology moved from the countryside into town.” In seeking to understand the mechanisms by which the embedding of innovative forms in linguistic structure takes place, Kristiansen believes we must integrate the explanatory frameworks developed within the Labovian sociolinguistic paradigm with those that have come down to us via other tributaries to the stream of research that has emerged from more than two centuries of scholarly interest in dialect variation.

One strand in this braid, to echo Kristiansen’s riverine metaphor, is the variety of dialectometry practiced by Hans Goebel (Ch. 7), for which he suggests the title “atlantometry” owing to the fact that the statistical modeling methods he employs are based exclusively on dialect atlas data. The geolinguistic focus of the early European dialectologists is hardly accidental, when one considers how energetically the new nation states went about surveying their territories and standardizing the ways in which the observations they made were to be quantified. Figuring out how to handle the multidimensional richness of dialect data would present

its own challenges, of course. From our contemporary perspective, it is only too easy to underestimate the hurdles that the early dialectometrists had to overcome in terms of bootstrapping into existence a set of metrics that would permit them to capture linguistic distance in a principled, systematic, and consistent fashion.

Even if dialects were stable and unchanging, the tasks faced by the atlantometrist or any other investigator of dialect would be daunting enough. But, as David Britain points out in his chapter on dialect contact and new dialect formation, new dialects arise all the time. The role of contact between speakers of different dialects in the emergence of new subvarieties of a language has been the subject of considerable debate in recent times, for example in terms of the extent to which identity factors might bias interactants toward or away from the adoption of certain linguistic behaviors (cf. Trudgill 2008; Hickey 2013). Although divergence between dialects is frequently observed and does not presuppose an absence of contact between the speakers of the varieties in question (e.g., Auer, Hinskens, and Kerswill 2005), it is also clear that human beings seem to be possessed, as Britain has it, of an overwhelming “urge to converge” linguistically. This predisposition is played out on a continent-sized stage and over more than a millennium of history in the European contexts explored by Peter Auer (Ch. 9), whose discussion focuses chiefly on continental Germanic languages, but applies in large part also to members of the other European language families. The grand themes of leveling and convergence are illustrated time and again in dialectological and sociolinguistic studies of language varieties across Europe, such that it becomes unviable to dispute Auer’s assertion that a march toward linguistic uniformity is the major defining characteristic of the European dialects over the course of the last century.

What is surprising is that in many parts of the world non-linguists cleave quite so strongly to their beliefs in the persistence of dialect forms that are either extinct or shortly to become so. Tapping into these beliefs and attitudes is the goal of perceptual dialectology (also known as “folk linguistics”), the field of which Dennis Preston (Ch. 10) has become the most prominent champion. Close scrutiny of what laypeople have to say about the language forms they and their fellow speakers use—or are believed to use—can tell us a good deal about how languages are changing or are liable to change in the future, insofar as the metalanguage of non-linguists often yields insights into their proclivity to adopt or reject incoming changes or to retain traditional ones. Laypeople may justify their views about dialects and accents in terms of how readily the speech patterns of other speakers and speaker groups can be understood, a habit which, as Charlotte Gooskens observes in Chapter 11, is no mere matter of casual academic interest. It may help language planners to decide whether a dialect is treated as a distinct language or not, with all the implications this decision might have for the rights and opportunities of minority populations. It is also true to say that stigmatizing certain language varieties for their (ostensibly) low intelligibility can have very tangible knock-on consequences for the variety’s speakers, who may suffer disadvantage as a result of negative attitudes toward their vernaculars held by those in positions of power in the educational or legal systems. In the final chapter in this section (Ch. 12), I deal further with perception issues, but this time in the context of strategies that speakers might apply if they wish to disguise their regular speech patterns. Leading listeners to believe that one’s feigned dialect or accent is authentic is generally a perfectly benign deception—we expect this ability to be well-developed in professional actors or individuals who have received elocution lessons, for instance—but where the aim is to gull listeners into thinking they are hearing a different person talking (e.g., so as to obtain money fraudulently, or to impersonate a murder victim so as to mislead others into thinking the victim is still alive) the law may view the attempt as a criminal offence.

Essays on other applications of dialectological data could easily fill a separate volume, and it is encouraging to see that publishing projects of just this type are underway as part of the current drive to expand and enrich the already very sizeable body of literature on dialect variation. It is our hope that the chapters in this Handbook will quickly come to represent an invaluable and authoritative resource for all involved in these new dissemination initiatives.

REFERENCES

- Allan, Keith, ed. 2013. *The Oxford Handbook of the History of Linguistics*. Oxford: Oxford University Press.
- Auer, Peter, Frans Hinskens, and Paul Kerswill, eds. 2005. *Dialect Change: Convergence and Divergence in European Languages*. Cambridge: Cambridge University Press.
- Buchstaller, Isabelle, and Seraphim Alvanides. 2013. "Employing geographical principles for sampling in state of the art dialectological projects." *Journal of Linguistic Geography*, 1: 96–114.
- Cohen, Emma. 2012. "The evolution of tag-based cooperation in humans: The case for accent." *Current Anthropology*, 53(5): 588–616.
- Docherty, Gerard J., and Dominic Watt. 2001. "Chain shifts." In *The Concise Encyclopedia of Sociolinguistics*, edited by Rajend Mesthrie, 303–307. Amsterdam: Pergamon/Elsevier Science.
- Goldman, Jean-Philippe, Adrian Leemann, Marie-José Kolly, Ingrid Hove, Ibrahim Almajai, Volker Dellwo, and Steven Moran. 2014. "A crowdsourcing smartphone application for Swiss German: Putting language documentation in the hands of the users." *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, Reykjavik: 3444–3447.
- Gordon, Matthew J. 2013. "Investigating chain shifts and mergers." In *The Handbook of Language Variation and Change*, 2nd edn., edited by Jack K. Chambers, and Natalie Schilling, 203–219. Oxford: Wiley-Blackwell.
- Heggarty, Paul, April McMahon, and Robert McMahon. 2005. "From phonetic similarity to dialect classification: A principled approach." In *Perspectives on Variation: Sociolinguistic, Historical, Comparative*, edited by Nicole Delbecque, Johan van der Auwera, and Dirk Geeraerts, 43–91. Amsterdam: Mouton de Gruyter.
- Hickey, Raymond, ed. 2013. *The Handbook of Language Contact*. Oxford: Wiley-Blackwell.
- Huang, Yuan, Diansheng Guo, Alice Kasakoff, and Jack Grieve. 2015. "Understanding U.S. regional linguistic variation with Twitter data analysis." *Computers, Environment and Urban Systems*. Available online 31 December 2015, ISSN 0198-9715, <http://dx.doi.org/10.1016/j.compenvurbsys.2015.12.003>.
- Labov, William, Sharon Ash, and Charles Boberg. 2006. *The Atlas of North American English: Phonetics, Phonology and Sound Change*. Berlin: Mouton de Gruyter.
- Long, Jason. 2013. "The surprising social mobility of Victorian Britain." *European Review of Economic History*, 17(1): 1–23.
- Milroy, James. 1992. *Linguistic Variation and Change: On the Historical Sociolinguistics of English*. Oxford: Blackwell.
- Salmons, Joseph, and Patrick Honeybone. 2015. "Structuralist historical phonology: Systems in sound change." In *The Oxford Handbook of Historical Phonology*, edited by Patrick Honeybone, and Joseph Salmons, 32–46. Oxford: Oxford University Press.
- Trudgill, Peter. 2008. "Colonial dialect contact in the history of European languages: On the irrelevance of identity to new-dialect formation." *Language in Society*, 37(2): 241–254.
- Whyte, Ian D. 2000. *Migration and Society in Britain, 1550–1830*. London: Palgrave Macmillan.
- Winney, Bruce, Abdelhamid Boumertit, Tammy Day, Dan Davison, Chikodi Echeta, Irina Evseeva, Katarzyna Hutnik, Stephen Leslie, Kristin Nicodemus, Ellen C. Royrvik, Susan Tonks, Xiaofeng Yang, James Cheshire, Paul Longley, Pablo Mateos, Alexandra Groom, Caroline Relton, D. Tim Bishop, Kathryn Black, Emma Northwood, Louise Parkinson, Timothy M Frayling, Anna Steele, Julian R. Sampson, Turi King, Ron Dixon, Derek Middleton, Barbara Jennings, Rory Bowden, Peter Donnelly, and Walter Bodmer. 2012. "People of the British Isles: Preliminary analysis of genotypes and surnames in a UK-control population." *European Journal of Human Genetics*, 20: 203–210.

1 Dialectology, Philology, and Historical Linguistics

RAYMOND HICKEY

1.1 Introduction

The term “dialect” is understood today to refer to a geographically delimited form of language. The purpose of the present chapter is to trace the history of this meaning of the word and to outline the rise of dialectology, which is the historical study of dialects in this sense (Fisiak ed. 1988). Furthermore, this study seeks to set dialectology in relation to the disciplines of philology (Turner 2014; Momma 2015: 1–27) and historical linguistics. These latter two are closely related in that the former fed into the latter. Indeed, the modern discipline of linguistics arose at the end of the eighteenth and beginning of the nineteenth centuries out of earlier concerns of philology, which is the study of the textual records of languages.

The etymology of “dialect” can be traced back to Classical Greek, in which the word διαλέκτος originally referred to discourse, conversation, or way of speaking, and later came to mean a regional variety of a language. It is this last meaning that initiated the modern understanding of the word (the older meaning of “investigative discussion” can still be recognized in the term “dialectic”). However, one cannot say that once the meaning of “regional variety” was established one had a usage similar to that today. The essential difference is that nowadays “dialect” stands in a contrastive relationship to “standard,” a form of language favored in the public domain and employed in compiling official documents in a country. The reference to “country” is important here: the modern sense of “standard,” with all its prescriptive connotations, is essentially an artifact of modern nation states. Thus, the often negative connotations of dialect did not hold until the notion of a preferred form of language arose, a form that enjoyed preference in writing, education, and public speaking. How early this preference occurred historically is difficult to say with certainty. True, there were historical constellations of language varieties in which one was used more than others. This applied in the Hellenistic period of Greek (roughly three centuries before the beginning of the common era), when the dialect of Attica (including the city of Athens) was used widely as a koiné or common form of language in the eastern Mediterranean (Woodard 2008). In England, during the later Old English period, the language of the West Saxon region was employed in written documents (Gneuss 1972), such as religious or legal texts, and thus enjoyed a similar status to Attic Greek in ancient Greece. But in neither case did later attributes of standard forms of language apply, above all codification and prescriptivism, which involved the censure of dialect forms of the same language.

To trace the history of dialects and their study, that is, dialectology, one should distinguish three aspects of this complex: (i) awareness of dialects, (ii) attitudes toward dialects, and (iii) the description and study of dialects. These aspects stand in chronological order: first awareness arises, and mention of dialects is found in the textual record. Somewhat later, attitudes toward dialects seem to have developed. In the Western world these are invariably negative, with mention made of a preferred form of the language in question. Later still, one finds descriptions of dialects, usually of one particular language with which an author has a specific connection, either by birth or interest.

1.2 Dialect Awareness and Attitudes

An awareness of dialect differences in England goes back at least to the Middle Ages: Geoffrey Chaucer used Northern English (Hickey 2015) for the purpose of character portrayal in *The Reeve's Tale* (Tolkien 1934; Wales 2006, 75). The north/south dichotomy is referred to by later authors on language, notably George Puttenham, who, in his *The Arte of English Poesie* (1589), states his preference for “our Southerne English,” which is the “usual speech at court and that of London and the shires lying about London within lx. myles and not much aboue” (Mugglestone 2007, 9). One of the earliest listings of dialect areas was made by Alexander Gil in his celebrated *Logonomia Anglica* (1619). On discussing the main features of the different dialectal areas, he mentions the northern lack of rounding in *beath* “both,” and the northern forms *sal* “shall,” *sud* “should,” *fula* “follow,” and *briks* “breeches.” There was also an awareness of the Englishes spoken in the Celtic regions: Shakespeare in the “Four Nations Scene” in *Henry V* portrays the speech of English, Welsh, Scottish, and Irish characters.

In France, the primary north/south dialect division was also characterized as early as 1284, by the poet Bernat d'Auriac. Here the forms of the keyword “yes” are essential, and have even resulted in the names of two large parts of France—Languedoc and Languedoel—the former referring to the region south of the River Loire, the latter to that north of the Loire (which later developed into modern French).

There would seem to have been two attitudes to English dialects in the early modern period. One was neutral and the other decidedly in favor of southern speech. John Hart (d. 1574) spoke of “the flower of the English tongue,” referring to the language of the London court. The more neutral attitude is seen in dictionaries of the time, for example, William Bullokar (1616): “So every country hath commonly in divers parts thereof some difference of language, which is called the Dialect of that place,” a view echoed by Thomas Blount (1656): “Dialect is a manner of speech peculiar to some part of a Country or people, and differing from the manner used by other parts or people, yet all using the same Radical Language, for the main or substance of it.” But the great lexicographers of the eighteenth century—Johnson, Sheridan, Walker—showed no interest in regional variation, meaning that dialects were excluded from the emerging ideology of a standard in English.

1.3 The Description of Dialects

Early descriptions of dialects differ from later studies in that they largely consist of dialect words. The gathering of such words and their publication as lists has a long tradition in England and other European countries. The most significant early English work is John Ray's *A Collection of English Words not Generally Used* (1674). Ray states that “in many places, especially of the North, the Language of the common people, is to a stranger very difficult to be understood” (Preface *To the Reader*), and was hence motivated to record northern words.

In some dictionaries, northern words that were not current in the south of England are given. For instance, John Palsgrave's *Lesclarcissement de la Langue Francoise* (1530) mentions words such as *sperre* "to shut" and *that ylke day* "that same day" that are representative of "the northern language" (Palsgrave 1530, fo. CCC. lxviii; see also Ruano-García 2010, 109–128; Stein 1997). Another significant work in this respect is the unpublished compilation by Bishop White Kennett (1660–1728) titled the *Etymological Collections of English Words and Provincial Expressions* (1690s, MS Lansdowne 1033).

1.4 The Antiquarian Tradition

Kennett's collection is representative of a genre of early dialect studies in which the ultimate origin and the subsequent history of English are of interest. This type of work is part of an antiquarian tradition that arose in the early modern period and has continued into modern times, albeit usually without the wild claims for the genesis of languages that were typical of early antiquarian works. The writers were well-meaning amateurs, often members of the clergy or military. An example of the latter was the English army officer Charles Vallancey (1721–1812), whose interest in dialects led him to compile a glossary of the archaic dialect of Forth and Bargy, Co. Wexford, in the southeast of Ireland (Hickey 2007, 66–84).

1.5 Dialects in the Age of Prescriptivism

The lack of academic concern with dialects prior to the nineteenth century could be attributed to the absence of a scientific framework for the study of language in general. However, in the eighteenth century one recognizes a deliberate neglect of regional features in English (Beal 2010a), and indeed severe condemnation of all language traits that do not correspond to "standard" usage, whatever the latter might mean. More neutral attitudes are visible in the detailed entries offered in some seventeenth-century English dictionaries. As we saw above, William Bullokar and Thomas Blount gave definitions of dialect that look very objective to modern readers. Bullokar's entry continues:

... [I]n England the Dialect or manner of speech of the North, is different from that in the South, and the Western dialect differing from them both. The Grecians had five especial Dialects: as in The property of speech in Athens: 2 in Ionia; 3. In Doris; 4. In Eolia: and 5. that manner of speech which was generally used of them all.

(Bullokar 1616 [no pagination]).

Blount's definition is yet more comprehensive:

In *England*, the Dialect, or manner of speech in the North, is different from that in the South; and the Western differs from both. As in this example: At *London* we say, *I would eat more cheese if I had it*, the Northern man saith, *Ay sud eat mare cheese gin ay hader*, and the Western man saith, *Chud ee'at more chiese on chad it: Chud ee'at more cheese un ich had it*. The Grecians had five especial Dialects ... So every Country commonly hath in diverse parts of it some difference of language, which is called the Dialect or Subdialect of that place. In Italy, there are above eight several dialects or Subdialects...

(Blount 1656 [no pagination]).

Blount is remarkable in that he gives examples to illustrate different dialects of English. However, he is not followed by others. Somewhat later (1676), Elisha Coles published *An English Dictionary* in which he sees dialects as "Logick, speech; also a particular Propriety or

Idiom in the same speech," with no reference to regions whatever. Some authors do at least specify that dialects are found in different parts of a country, for example, John Kersey, whose *Dictionarium Anglo-Britannicum* (1708) defines dialect as "a Propriety of manner of Speech in any Language, peculiar to each several Province or Country." Reference to region is also found in Thomas Spence (Beal 1999) who says of dialect that it is "A polite manner of speaking, or diversity made in any language by the inhabitants in any part of the country where it is spoken; stile; speech" (Spence 1775 [no pagination]).

As noted earlier, the great lexicographers of the eighteenth century ignore dialect's regional essence, at most referring to dialects of classical Greek. Instead, they concentrate on its meaning as a manner of expression. Samuel Johnson offers the following in his authoritative *Dictionary of the English Language* (1755):

DIALECT

1. The subdivision of a language; as the Attic, Doric, Ionic, Æolic dialects.
2. Stile; manner of expression.

When themselves do practise that whereof they write, they change their dialect; and those words they shun, as if there were in them some secret sting. Hooker, b.v.s. 22.

3. Language; speech.

Later language commentators and prescriptivists, above all, Thomas Sheridan (1719–1788) and John Walker (1732–1807), were content to adopt and repeat Johnson's definition. Neither Sheridan nor Walker had time for dialectal variation, which directly conflicted with their standard English ideology; hence, their derisory comments on the regional speech of Britain and Ireland.

Besides the prescriptivism of authors like Sheridan and Walker, another motivation is recognisable in the eighteenth-century neglect of dialect. Consider this statement in George Puttenham's *The Arte of English Poesie*: "After a speach is fully fashioned to the common vnderstanding, & accepted by consent of a whole countrey & nation, it is called a language." This is the view of language as a unifying factor, in Puttenham's case among the different regions of Britain. Here we have a very early reference to a "national language," a notion picked up by later authors such as the Scot James Buchanan in his 1766 *Essay Towards Establishing a Standard for an Elegant and Uniform Pronunciation of the English Language Through the British Dominions*. Attention to dialects, let alone their valorisation, would not have been reconcilable with the desire for a "national language."

By the close of the eighteenth century, notions of national language and standard language would seem to have merged, at least for many authors. The variation which was to be suppressed was regional (Beal 2010b), with the parallel promotion of the English of southeast England. As the variation was principally phonetic, one's accent came to indicate one's relative standardness as a speaker (Beal 2004), which remains the case today (Beal 2008).

1.6 The Denigration of Dialects

After the early eighteenth century the assessment of pronunciation appears to have changed (Beal 2010b). While Defoe in the 1720s could remark non-judgmentally on the attitude of Northumbrians to features of their pronunciation, after the mid-eighteenth century comments are far more critical. A vocabulary was adopted by authors on language that is condemnatory of all features that were not part of received southeastern English usage.

"Vulgar" is a censorious epithet used to describe a variety disapprovingly. The term was very common in the eighteenth and nineteenth centuries in evaluative treatments of

language like Savage's *The Vulgarisms and Improprieties of the English Language* (1833). However, before the eighteenth century "vulgar" simply meant "of the people" (cf. Latin *vulgaris* "common people"). John Walker is particularly keen to specify what he thinks merits the label "vulgar." For instance, given that provincial speakers had to look to the capital for phonetic guidance, any "vulgarisms" used by Londoners are especially to be condemned. In his *Critical Pronouncing Dictionary* he lists "faults of the Londoners," who, "as they are the models of pronunciation to the distant provinces, ought to be the more scrupulously correct" (1791, xii).

The early and late modern periods saw increasing divergence of English pronunciation and spelling, not solely as a result of the Great Vowel Shift (Pyles and Algeo 2004, 170–173). Several other developments contributed to this divergence, for example, the lowering and unrounding of short [u] to [ʌ] in the STRUT lexical set and vowel lengthening in BATH words. Many changes in this period resulted in homophony and hence led to distinctions in spelling which did not correspond to pronunciation differences (e.g., the merging in southeastern English of the TERM and TURN sets to a rhotacised schwa, which then simplified solely to schwa). But many dialects retained this distinction, which led to the stigmatisation of these varieties.

Twentieth-century scholars revisited the issue of social stigma (Lippi-Green 1997) both for those dialects with an ethnic basis—for example, African American English (Rickford 1999) or Chicano English (Fought 2006)—and those on the periphery of industrialised societies, for example, the "Ocracoke Brogue" of North Carolina's Barrier Islands (Wolfram and Schilling-Estes 2004). The educational implications of the stigma experienced by speakers was often a primary concern (see Wolfram, Adger and Christian 1999; Baugh 2004).

1.7 From Philology to Linguistics

In scholarly literature the term "philology" has two similar but distinct meanings. The first is the study of older texts, whereas the second is the comparison of older stages of languages. The second meaning usually has the longer designation "comparative philology." This scholarly activity was common in the nineteenth century, when the family relationships among languages, chiefly Indo-European ones, were reconstructed on the basis of older textual records.

Hale (2007) states in his handbook of historical linguistics that he sees philology as the scrutiny and analysis of historical artefacts, which do not represent language but are imperfect windows on language. He claims that "Philology is responsible for establishing the attributes of a text, many of which may be relevant for subsequent linguistic analysis" (2007, 21), and continues, "[t]here are two goals, related to one another, of this enterprise: to understand the linguistic structures present in the text itself (let's call this the 'local' goal) and to understand the structures, entities, and processes which made the grammar of the 'composer' of the text (let's call this the 'ultimate' goal)" (2007, 23).

In the present chapter it is the second meaning of "philology," which is used. In this sense, "comparative philology" is synonymous with historical linguistics as practised throughout the nineteenth century. Because at the time the concern of linguists was overwhelmingly with members of the Indo-European language family, the field is commonly known as Indo-European studies (German *Indogermanistik*). It arose in the late eighteenth century, triggered by the work of Sir William Jones (1746–1794), who insisted on the relatedness of the Indo-European languages, then as now one of the major language families of the world and certainly the best researched. Jones was followed by others such as Rasmus Rask (1787–1832) and Jacob Grimm (1785–1863), who established the science of comparative philology at the beginning of the nineteenth century. It was the dominant school of linguistics until the advent of structuralism at the turn of the twentieth century.

1.8 Features of Indo-European Studies and Comparative Philology

1. The dissociation of linguistics and philosophy
2. The establishment of a sound foundation for etymology
3. The abandonment of attempts to prove putative relationships between Hebrew (the language of the Old Testament) and various European languages
4. The development of a descriptive apparatus for phonetics

In the latter half of the nineteenth century the Neogrammarian hypothesis was developed by a group of young linguists (the *Junggrammatiker*, lit. “new grammarians”) working in Leipzig. They assumed that language change proceeds gradually on a phonetic level, affecting all input sounds simultaneously. The Neogrammarians’ confidence in their assessment of sound change was fuelled by additional discoveries, most notably that by Karl Verner (1846–1896). Verner gave a satisfactory account—now known as Verner’s Law—of the apparent irregularity in many word forms in Germanic, which had been a concern since the days of Jacob Grimm. Subsequently, belief in the regularity of sound change, tempered by analogy, was fully established. The theoretical underpinnings of the Neogrammarian hypothesis were provided by Hermann Paul in his seminal *Prinzipien der Sprachgeschichte* (1886), in particular the much-debated view that sound laws were exceptionless.

The manner in which the Indo-European languages are assumed to have divided is envisaged by the *Stammbaum* “family tree” metaphor, a notion introduced by August Schleicher (1821–1868). A *Stammbaum* is an inverted tree, with branchings from top to bottom. At the top is Proto-Indo-European, and at the bottom the individual languages of the various branches. The tree representation has also been used to show the interrelationship of dialects.

Indo-European studies/comparative philology involved comparing cognate forms from genetically-related (usually Indo-European) languages with a view to reconstructing the proto-language from which the others were thought to have derived. This allowed scholars to trace superficially different forms back to a single (generally unattested) form. For instance, English *heart*, German *Herz*, Latin *cordia*, and Greek *kardia* can be shown to derive regularly from an Indo-European root **kerd*. The same principle was assumed to be possible when investigating dialects: comparing dialectal forms helped scholars in reconstructing earlier stages of languages, often because key forms missing from more standard varieties were attested in dialects. For example, the northern word *thole* “suffer, endure” is a continuation of Old English *þolian* (cf. German *dulden*), although the form does not exist in present-day standard English.

Indo-European studies/comparative philology was initially a German endeavor, but over the course of the nineteenth century scholars from other countries, for example, Scandinavian countries, the Netherlands, and France, became involved. In England, the main scholar was Henry Sweet (1845–1912), the author of books on phonetics and the history of English in general. He also developed a system of phonetic transcription, the *Romic Alphabet*, an important precursor of the International Phonetic Alphabet primarily promoted by Paul Passy (1859–1940) and a necessary instrument for the documentation of dialects.

1.9 The Dawn of Modern Dialectology: The Beginnings of a New Discipline

The systematic study of dialects began in the latter half of the nineteenth century, although as we have seen there is a long history of observation of dialect differences prior to this. The linguistic analysis of dialect variation is associated with the rise of historical linguistics,

which led to publications such as Walter Skeat's overview of historical dialects (Skeat 1912). But it was the activity of scholars dedicated to documenting traditional dialects in danger of dying out, which was to prove more relevant for later dialectology. Investigations of dialects usually produced maps on which isoglosses were drawn. An isogloss is a line separating the occurrence of two different but related forms, for example, the isogloss which separates the occurrence of [ʊ] (in the north of England) from [ʌ] (in the south of England) in the STRUT lexical set (Wells 1982, 131–133), or that separating the regions in which non-prevocalic /r/ is present (the “rhotic” areas) versus absent (“non-rhotic” areas, which account for practically the whole of England other than the southwest and a small area of Lancashire). While isoglosses seem useful and present a neat picture of sound distributions, they only apply to speakers of traditional dialects and even then cannot depict the co-occurrence and statistical distribution of variants in transitional regions, let alone do justice to relevant factors such as age, gender, class, rural/urban divisions, and so on (all of which would be relevant to the documentation of a feature such as word-initial /h/ in English).

Nonetheless, isoglosses were an integral part of surveys of traditional dialects in the twentieth century before the advent of sociolinguistics. The prime example of such a survey in England is the comprehensive *Survey of English Dialects* (SED), initiated by Harold Orton (1898–1975), who had studied under Joseph Wright and Henry Wyld. He was appointed professor at Leeds after World War II, and together with Eugen Dieth (Zurich) supervised the SED. The survey involved over 1,000 questions per questionnaire, covering pronunciation, grammar, and vocabulary on topics such as farm life, nature, household matters, weather, health, and social activities. It appeared between 1962 and 1978 with the *Basic Material* being published as four volumes containing informants' responses to interview questions. Further interpretative volumes were published based on the SED's findings, for example, Kolb's *Phonological Atlas of the Northern Regions* (1966), *A Word Geography of England* (Orton and Wright 1974), and *The Linguistic Atlas of England* (Orton, Sanderson, and Widdowson 1978). Further works based on SED data are Upton, Sanderson, and Widdowson (1987), Viereck and Ramisch (1991), and Upton and Widdowson (1996).

Despite the collaborative work that characterizes the SED, there were also key individuals working in dialectology. In England, the most prominent was Joseph Wright (1855–1930), who set dialect study on a new footing in the early twentieth century. Wright studied in Heidelberg and Leipzig and at these centers came into contact with leading linguists of the day. Later he accepted a professorship at Oxford. He is now known for two works, the *English Dialect Dictionary* (5 vols., 1898–1905) and the sixth volume of this work, his *English Dialect Grammar*, all of which are still consulted today, although the coverage is incomplete. Wright's predecessor, Alexander Ellis (1814–1890), was to become one of the foremost phoneticians and dialectologists of his day, and is remembered for his five-volume *On Early English Pronunciation* (1868–1889).

In Europe there were also early pioneers of dialectology. The main work of Jules Gilliéron (1854–1926), a French linguist who was instrumental in the development of modern dialectology and areal linguistics, was a multi-volume atlas of French dialects produced at the beginning of the twentieth century. Gilliéron sent out trained fieldworkers to conduct interviews and record data using a consistent phonetic notation. One of Gilliéron's fieldworkers, Edmond Edmont (1849–1926), conducted no fewer than 700 interviews across France between 1896 to 1900, using questionnaires involving over 1,000 items. The results of his observations, chiefly drawn from male informants, together with results from Gilliéron and his other assistants, were subsequently published between 1902 and 1910 as the *Atlas linguistique de la France*. In Germany, similarly pioneering work was carried out by Georg Wenker (1852–1911). In 1876, he began sending out questionnaires of some 40 sentences to over 1,200 schoolmasters across the north of Germany, asking them to provide equivalents of words in their local dialect. Over a decade he received about 45,000 completed questionnaires. Wenker

transferred the information to maps that in 1881 were published as *Sprachatlas des Deutschen Reiches* “Linguistic Atlas of the German Empire,” covering north and central Germany. Wenker continued gathering questionnaires, and in 1926 the first volume of the *Deutscher Sprachatlas*, based largely on Wenker’s data, was published under the editorship of Ferdinand Wrede. Another notable German is Wilhelm Doegen (1877–1967), who had an interest in recording minority languages and dialects. Doegen studied phonetics in Berlin and later under Henry Sweet in Oxford, where he increased his knowledge of English and the anglophone world. He also became a member of the International Phonetic Association. Doegen’s original recordings of English dialect speakers were destroyed during World War II, but shellac copies survived and in the 1990s the Humboldt University in Berlin started digitizing this material to form the Berliner Lautarchiv corpus.

An exception to the general orientation of traditional dialectology and a precursor of modern studies of language and society is Louis Gauchat (1866–1942), a French-speaking Swiss scholar who in 1905 published a study of language use in the Alpine town of Charmey. His recognition that young people used different pronunciations from older ones, and that females led in the use of new variants, that is, are the vanguard in change, anticipated many of the insights of sociolinguistics as it developed in the 1960s and 1970s (Labov 1972).

1.10 Dialect Societies and Materials

Societies for the study of dialects arose in the nineteenth century in parallel to the activities of scholars. In England, the *English Dialect Society* was founded by Walter Skeat and lasted from 1873 to 1896, after which it was dissolved voluntarily. In this relatively short timespan the society published some 80 works on the dialects of England.

In America, a similar institution was founded in 1889. The *American Dialect Society*, mainly dedicated to the study of the English language in North America, published (and still publishes) the academic journal, *American Speech*, which has successfully adapted to modern developments in linguistics.

Journals dedicated solely to dialectology have sometimes had a precarious existence. The Belgian journal, *Orbis*, began in 1952 with a focus on dialectology, but went into sharp decline in the 1980s. The recently founded *Journal of Linguistic Geography* (2013–) is an online journal published by Cambridge University Press. Journals with a broader remit, mainly those that deal with variation from a contemporary sociolinguistic perspective (e.g., *English World-Wide* (1980–), *World Englishes* (1981–), and *Language Variation and Change* (1989–)), have been more successful.

The nineteenth century also saw the publication of dialect dictionaries, often dedicated to specific regions of a country. The North of England is the subject of Brockett’s *Glossary of North Country Words* (1825, 1846), whereas more restricted locales are treated in works such as Dinsdale’s *Glossary of Provincial Words Used in Teesdale* [Co. Durham] (1849), Nodal and Milner’s *Glossary of the Lancashire Dialect* (1875–1882), and Dickinson’s *Glossary of Words and Phrases pertaining to the Dialect of Cumberland* (1878–1881). Other dictionaries have a broader scope, for example, Pickering’s *Vocabulary or Collection of Words and Phrases which have been supposed to be peculiar to the United States of America* (1816), whereas some consist of extractions of dialect words from more general works (Wakelin 1987; Görlich 1995), for example, Axon’s *English Dialect Words in the Eighteenth Century as Shown in the Universal Etymological Dictionary of Nathaniel Bailey* (1883).

The twentieth century saw comprehensive dialect dictionaries attempting complete coverage of a country or clearly delimited region (Penhallurick 2009). The five-volume *Dictionary of American Regional English* (DARE), compiled under the supervision of Frederick Cassidy and Joan Hall at the University of Wisconsin and published between 1985 and 2012 by

Harvard University Press, gives complete coverage of regional vocabulary in the United States. The comprehensive *Dictionary of Newfoundland English* covers dialect vocabulary in Newfoundland, Canada. It was compiled by George Story, William Kirwin, and John Widdowson, and first published in 1982.

The list of dictionaries could be extended considerably if those dealing with a single anglophone country were to be included (see Hickey 2014). For instance, the *Dictionary of Canadianisms on Historical Principles* (1967) is a major lexicographical work compiled under the supervision of Walter Avis (1919–1979). In 2006, a comprehensive revision was initiated at the University of British Columbia as the project DCHP-2, which contains much lexical material of dialect origin in the British Isles.

1.11 Dialect Studies

Monographs on dialects come in various guises. Apart from popular literature on local dialects there is academic literature, which can be for a general audience, for example, Trudgill (1990), Hughes, Trudgill, and Watt (2012), or for scholars, basically all other studies. Some studies are in a more traditional mode, for example, Brook (1978 [1963]), Petyt (1980), Wakelin (1972, 1977), Kirk, Sanderson, and Widdowson (1985), or Kolb *et al.* (1979). Other works have taken the insights of modern linguistics on board, for example, Milroy and Milroy (eds, 1993), Kortmann *et al.* (2004), and Dossena and Lass (eds, 2009). One can also mention the studies of forms of American English found in Wolfram and Schilling-Estes (2006) as well as, in a more popular vein, Wolfram and Ward (2005).

1.12 Data Collection Methods

The initial source of data for dialectology was the wordlist, a collection of words supposedly peculiar to the speech of a region. This plotted a trajectory for dialect studies that was characterized by lexical issues. The words sought were frequently those concerning traditional lifestyles (farming, crafts, and domestic issues in rural life). To glean such data, researchers devised lexical questionnaires containing questions like “What do you call the animal which builds dams in streams and rivers?” or, during interviews, “What do you call this part of the body?” with the interviewer touching his/her knee. The limitations of such methods are obvious, and clinging to them spelt oblivion for many dialect studies.

The techniques of modern sociolinguistics (Podesva and Sharma, eds. 2013), above all the rapid anonymous interview promoted by William Labov (Labov 1966), were often adopted to circumvent the observer’s paradox (speakers’ alteration of their speech while under observation by the linguist). But apart from short stretches of speech used for phonetic analysis, this method was not very suitable. For better or worse, informants were usually aware that they were being interviewed for a survey. Indeed, the increased attention paid to ethical issues demanded that the purpose of a survey be revealed to potential informants in advance. When collecting syntactic data, furthermore, much longer stretches of speech in which constructions might occur are necessary, and so anonymous recording would be impractical.

Other data collection issues moved centre-stage during the later twentieth century. One is randomness: for a survey to be representative, all speakers in a community must in principle have the same chance of being selected for a survey. This principle has been followed in surveys such as Telsur, which formed the basis for the *Atlas of Northern American English* (Labov, Ash, and Boberg 2005). Another is the issue of speech registers, which speakers have at their disposal. It may be of specific concern for a survey to determine how speakers alter their

speech as a function of formality. To capture this, some dialect studies record people speaking freely, reading a text passage, then a wordlist, to see where linguistic features fall on a cline of formality.

Parallel to these concerns, alternative methods of interfacing with informants were trialled. Interviews in groups generally led to a relaxation in speech style. The withdrawal of the interviewer, with informants being recorded on their own, offered another means of avoiding the observer's paradox, albeit with new issues of reliability and control arising. These elicitation techniques were chiefly employed for the collection and later analysis of syntactic data (Buchstaller and Corrigan 2011; Walker 2013), which requires a considerable amount of informal material (Schilling 2013).

1.13 Accessibility of Data

Once data have been collected, the issue of presentation to and accessibility for the public is addressed. Traditionally, dialect material has been presented in print, often in several volumes of maps with entries for speakers illustrating specific dialect forms (see Wagner 1958–1964). Increasingly, such information is being presented in a digital format. Some surveys have combined print material with a CD-ROM/DVD, for example, Labov, Ash, and Boberg (2006), often containing clickable active maps linked to sound files associated with speakers from specific locations (see Hickey (2004) as an example). Websites dedicated to dialects/varieties act as sources of general information, for example, the *International Dialects of English Archive* at <http://www.dialectsarchive.com>, whereas others are more research-based, for example, *Variation and Change in Dublin English* (<http://www.uni-due.de/VCDE>).

1.14 Dialectology and General Linguistics

Dialect geography built upon the Neogrammarian hypothesis that sound change was regular, that is, rule-governed and exceptionless. Although the Neogrammarians' claims concerning the rules of sound change are substantially correct, dialect situations are more complex and reveal that sound changes are not always exceptionless. Much discussion surrounding this issue has taken place (Labov 1981), with the level of language on which a feature is located playing a role, as well as the attitudes of speakers to incipient or ongoing change that can lead to their promoting or disfavoring changes.

Already by the nineteenth century the issue of how dialect features spread spatially was an issue addressed by scholars. Schleicher's tree model (see above) regarded dialect diversity as arising through a process of binary branching. A later view is the wave theory developed by Johannes Schmidt around 1870. This sees language changes as spreading out from a center like concentric waves in water.

In the twentieth century, with the focus on urban rather than rural forms of language, new conceptions of feature spread arose. The cascade model of diffusion regards changes as spreading from one urban centre to another without affecting the intervening countryside. An example is the spread of TH-fronting to urban centers around England, which are distant from London, without the intervening rural areas being affected. The size factor seems to be important, with larger cities adopting changes before smaller ones (Britain 2012). There would appear to be some instances of spread in the opposite direction, as captured by the term "counterhierarchical diffusion" (the opposite of what usually happens), whereby features spread from rural to urban settings. An instance of such spread would be *fixin' to*, which has been adopted into urban areas of Oklahoma.

1.15 Structuralism and Generativism

The heyday of American structuralism, between the 1930s and 1950s, saw some attempts to apply its principles to dialectology. The best-known of these is probably Weinreich (1954), in which the author argues for a “diasystem,” a superordinate level of structure above individual dialects, which would account for their perceived structural similarities (Weinreich 1954, 389–390). However, this notion proved untenable for many dialects, which developed as separate systems in geographically distinct areas, despite having common historic origins.

The ghost of a unifying underlying structure to dialects was not easily put to rest. In the 1960s and early 1970s attempts were made, for instance by Brian Newton investigating Greek dialects (Newton 1976) or by Martín Ó Murchú examining Irish dialects (Ó Murchú 1969), to show that the assumptions of generative linguistics could help explain dialect-relatedness. This strand of research was not very fruitful, however, and was discontinued. Ultimately, the historical relatedness of dialects was accepted as the source of present-day similarities, and abstractions across synchronic varieties were disfavored.

1.16 Dialectometry

Addressing the question of dialect relations without assuming a single superstructure continues via the approach known as “dialectometry” (Szmrecsanyi 2013). Essentially a European approach to dialectology—employed, for instance, in the analysis of Romance languages (Goebl 1982) or Dutch (Nerbonne and Heeringa 2010; Heeringa and Nerbonne 2013)—it uses numerical classification methods to analyze the apparent relatedness of dialects, and to measure the “distance” between them. The proponents of dialectometry highlight its ability to quantify dialect differences and to offer a measure of language change (Nerbonne 2003), whereas its critics see in its deterministic and mechanistic analysis of dialect variation a relative neglect of sociolinguistically determined variation.

Another approach to the grouping of dialects has become available with the increase in computers’ computational power. This makes use of phenograms, which are graphic representations of the structural similarities across groups of dialects (or languages; see Brato and Huber (2012) for a typical instance of such grouping based on African Englishes). Phenograms do not take the history of forms into account, only their synchronic manifestations. Thus, in the analysis of Spanish varieties discussed in Heggarty, McMahon, and McMahon (2005), Spanish in Madrid and in Bogotá show considerable similarities, but it is not clear whether this is due to accidentally shared developments or to continuity of the original Spanish input to Colombia (Heggarty, McMahon, and McMahon 2005, 85).

1.17 The Rise of New Dialects

Many European languages experienced diversification as a result of colonialism in the key period from 1600 to 1900. New forms of Portuguese, Spanish, Dutch, French and English arose outside Europe. This development led several scholars to consider the processes by which new dialects of established languages arise and acquire specific profiles. There are two main models in this field: Trudgill’s “New Dialect Formation” (NDF; Trudgill 2004, 2008) and Schneider’s “Dynamic Model” (DM), see Buschfeld, Hoffmann, Huber, and Kautzsch (eds. 2014).

NDF is viewed as a historical process whereby a new focused variety arises from a series of dialect inputs, for example, in late nineteenth-century New Zealand. Trudgill postulates the following stages: (1) rudimentary leveling, (2a) extreme variability, (2b) further leveling, (3) focusing. Thus new dialect formation has as its beginning a mixture of dialects, and as its

endpoint a single new dialect. In New Zealand, new dialect formation followed the initial immigration of speakers from different regions of the British Isles. This was a process of dialect mixture from which, over just a few generations, a focused variety arose that was then uniform and distinct from other existing varieties of English. Whereas the progression from input to output is uncontroversial, the question of just what input features survived into the later focused variety has been a matter of debate. Trudgill's stance is deterministic: the quantitative representation of features across speakers of input dialects (given in percentages) determines whether they become part of the output, with an appeal to linguistic markedness to explain the survival of minority variants such as schwa in, for example, *trusted*. If a feature was used by more than 50% of the English, Scottish, and Irish communities of early anglophone New Zealand, then it survived. For this to have worked, early anglophone New Zealand society would have had to be uniform, with contact among all speakers. Trudgill did not consider the status of immigrants and disputed the role of social factors for the young in following generations, for example, the fact that New Zealand was a British colony and hence, southeastern English features would have been favored by later generations; he also rejected any embryonic identity function for the combination of features that emerged in the later, focused variety (Hickey 2003). Additionally, there is no evidence that in a scenario where sociolinguistic factors apparently played no role the quantitative occurrence of a feature across the early communities would determine its survival. It might well be that in such a situation, if it ever obtained, the survival of features might be random.

The Dynamic Model was devised by the Austrian-German linguist Edgar Schneider to account for the development of English in former British colonies. It stresses the manner in which overseas varieties of English have evolved in specific ecologies and strives to account for which combinations of features have emerged. The model stresses the essential interaction of social identities and linguistic forms, the nature of which accounts in large measure for the profiles of post-colonial Englishes. Contact occupies a central position in Schneider's model, both between dialects present among settlers, as well as between English speakers and indigenous-language speakers in various colonial locations. Contact-induced change produced differing results depending on the social and demographic conditions under which it occurred, that is, on the local ecology, and on its linguistic triggers (e.g., code-switching, code-alternation, bilingualism, or non-prescriptive adult language acquisition).

The model was first presented in Schneider (2003) and later in more detailed form in Schneider's (2007) monograph *Postcolonial English*. It assumes that former colonies underwent various stages, which can lead ultimately to the development and differentiation of independent endonormative varieties of English, though this stage has not been reached in all cases. Schneider also proposes that there is a shared underlying process—a unilateral causal implication—driving the formation of postcolonial Englishes, as follows: sociohistorical background > identity of early groups > sociolinguistic conditions of communication and contact > resulting features of the emerging post-colonial variety.

Schneider identifies five stages in the development of post-colonial Englishes: Phase 1: foundation—dialect mixture and koineisation (for locations with multiple dialect inputs); Phase 2: exonormative stabilization—a “British-plus” identity for the English-speaking residents when the colony is established and has secured its position *vis-à-vis* the home country, mostly England (or the United States in the case of the Philippines); Phase 3: nativization involving the emergence of local patterns, often associated with political independence or the striving for this; Phase 4: endonormative stabilization, for example, “national self-confidence,” with codification, usually soon after independence; and Phase 5: differentiation—the birth of new dialects, internal developments now linked to internal socioethnic distribution processes. Further issues considered in Schneider's model include the distinction of settler and indigenous strands in the early stages of new varieties, the impact of accommodation, and the importance of identity formation.

1.18 Conclusion

Where does dialect study stand today? The scholarly investigation of dialects (Maguire, McMahon, and Dedić 2010) is as dynamic as ever, as evidenced by the present volume and by many recurring conferences such as *Methods in Dialectology*, and the orientation of research has changed considerably since the advent of modern sociolinguistics in the mid-twentieth century (Shorrocks 2000, 2001). The old style, in which older rural males formed the focus, has been abandoned completely, and is only referenced nowadays where a certain dialect shows nothing but literature of this type. Some of the stock components of older dialectology, such as the isogloss (see above), are no longer viewed as particularly useful as they rest on older, less inclusive conceptions of language variation.

The scope of dialectology has increased manifoldly with its exact extent resting ultimately on terminology. If dialectology is taken to encompass urban, sociolinguistic investigations—many authors speak of “social dialects” (Wolfram and Fasold 1974) or “urban dialects”—then its scope is wide indeed. But this would represent a weakening of the original focus of dialectology comparable to the overextended use of “pidgin” and “creole,” which is often found in the literature. In the sense of the linguistic study of regional forms of language, dialectology has matured considerably in the past half century and has proved its ability to adopt and incorporate insights from neighboring fields in linguistics. Examples are the compilation of corpora (Anderwald and Szemrešcanyi 2009) and the digitisation of existing literature, for example, Wright’s 1910 *English Dialect Dictionary* (Markus 2009; Markus, Upton, and Heuberger 2010), or the application of methods from the “language variation and change” paradigm (Chambers and Schilling 2013), which evaluates the social determinants of microvariation in speech communities. The ability of dialectology to be enriched by such inputs amply proves its vitality and robustness as a linguistic discipline in its own right.

REFERENCES

- Anderwald, Lieselotte, and Benedikt Szemrešcanyi. 2009. “Corpus linguistics and dialectology.” In *Corpus Linguistics: An International Handbook*, edited by Anke Lüdeling and Merja Kytö, 1126–1139. Berlin: de Gruyter.
- Baugh, John. 2004. “Ebonics and its controversy.” In *Language in the USA: Themes for the Twenty-First Century*, edited by Edward Finegan and John Rickford, 305–18. Cambridge: Cambridge University Press.
- Beal, Joan. 1999. *English Pronunciation in the Eighteenth Century: Thomas Spence’s ‘Grand Repository of the English language’*. Oxford: Oxford University Press.
- Beal, Joan. 2004. “‘Marks of disgrace’: Attitudes to non-standard pronunciation in eighteenth-century English pronouncing dictionaries.” In *Methods and Data in English Historical Dialectology*, edited by Marina Dossena and Roger Lass, 329–349. Frankfurt: Peter Lang.
- Beal, Joan. 2008. “‘Shamed by your English?’ The market value of a good pronunciation.” In *Perspectives on Prescriptivism*, edited by Joan Beal, Carmela Nocera, and Massimo Sturiale, 21–40. Frankfurt: Peter Lang.
- Beal, Joan. 2010a. *An Introduction to Regional Englishes: Dialect Variation in England*. Edinburgh: Edinburgh University Press.
- Beal, Joan. 2010b. “Prescriptivism and the suppression of variation.” *Eighteenth-Century English: Ideology and Change*, edited by Raymond Hickey, 21–37. Cambridge: Cambridge University Press.
- Blount, Thomas 1656. *Glossographia; or a Dictionary, interpreting all such Hard Words*. London: Printed by The Newcomb. Reprinted in 1969 by The Scolar Press (Menston).
- Brato, Thorsten, and Magnus Huber. 2012. “English in Africa.” In *Areal Features of the Anglophone World*, edited by Raymond Hickey, 161–185. Berlin: de Gruyter.

- Britain, David. 2012. "Diffusion". In *English Historical Linguistics*, edited by Alexander Bergs, and Laurel Brinton, 2013–2043. Berlin: de Gruyter.
- Brook, George. 1978 [1963]. *English Dialects*, 3rd ed. London: Deutsch.
- Buchstaller, Isabelle, and Karen Corrigan. 2011. "How to make intuitions succeed: Testing methods for analysing syntactic microvariation." In *Analysing Variation in English*, edited by Warren Maguire, and April McMahon, 30–48. Cambridge: Cambridge University Press.
- Bullockar, John. 1616. *An English Expositor: Teaching the Interpretation of the Hardest Words in our Language*. London: John Legatt.
- Buschfeld, Sara, Thomas Hoffmann, Magnus Huber, and Alexander Kautzsch, eds. 2014. *The Evolution of Englishes: The Dynamic Model and Beyond*. Amsterdam: Benjamins.
- Chambers, Jack, and Natalie Schilling, eds. 2013. *Handbook of Language Variation and Change*, 2nd ed. Oxford: Wiley-Blackwell.
- Dossena, Marina, and Roger Lass, eds. 2009. *Studies in English and European Historical Dialectology*. Bern: Peter Lang.
- Ellis, Alexander. 1868–1889. *On Early English Pronunciation* (5 vols.). London: Philological Society.
- Fisiak, Jacek, ed. 1988. *Historical Dialectology: Regional and Social*. Berlin: de Gruyter.
- Fought, Carmen. 2006. *Language and Ethnicity*. Cambridge: Cambridge University Press.
- Gauchat, Louis. 1905. "L'unité phonétique dans le patois d'une commune." In *Aus Romanischen Sprachen und Literaturen: Festschrift Heinrich Morf*, edited by Ernest Bovet, Ernst Brugger, Wilhelm Degen, Arturo Farinelli, Adolf Fluri, Louis Gauchat, Jakob Jud, Jules Jeanjaquet, Emil Keller, Martha Langkavel, Marie Johanna Minckwitz, Kaethe Schirmacher, Ernst Tappolet, and Louis Betz, 175–232. Halle: Niemeyer.
- Gil, Alexander. 1619. *Logonomia Anglicana*.
- Gneuss, Helmut. 1972. "The origin of standard Old English and Æthelwold's school at Winchester." *Anglo-Saxon England* 1: 63–68.
- Goebl, Hans. 1982. *Dialektometrie: Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie*. Vienna: Österreichische Akademie der Wissenschaften.
- Görlach, Manfred. 1995. "Dialect lexis in early modern dictionaries." In *New Studies in the History of English*, edited by Manfred Görlach, 82–127. Heidelberg: Winter.
- Hale, Mark. 2007. *Historical Linguistics: Theory and Method*. Oxford: Blackwell.
- Heeringa, Wilbert, and John Nerbonne. 2013. "Dialectometry." In *Language and Space: An International Handbook of Linguistic Variation*, Vol. 3 – Dutch, edited by Frans Hinskens, and Johan Taeldeman, 624–645. Berlin: de Gruyter.
- Heggarty, Paul, April McMahon, and Robert McMahon. 2005. "From phonetic similarity to dialect classification." In *Perspectives on Variation: Sociolinguistic, Historical, Comparative*, edited by Nicole Delbecque, Johan van der Auwera, and Dirk Geeraerts, 43–91. Berlin: de Gruyter.
- Hickey, Raymond. 2003. "How do dialects get the features they have? On the process of new dialect formation." In *Motives for Language Change*, edited by Raymond Hickey, 213–239. Cambridge: Cambridge University Press.
- Hickey, Raymond. 2004. *A Sound Atlas of Irish English*. Berlin: de Gruyter.
- Hickey, Raymond. 2007. *Irish English: History and Present-Day Forms*. Cambridge: Cambridge University Press.
- Hickey, Raymond. 2014. *A Dictionary of Varieties of English*. Malden, MA: Wiley-Blackwell.
- Hickey, Raymond. 2015. *Researching Northern English*. Amsterdam: Benjamins.
- Hughes, Arthur, Peter Trudgill, and Dominic Watt. 2012. *English Accents and Dialects: An Introduction to Social and Regional Varieties of British English*, 5th ed. London: Hodder Education.
- Kirk, John, Stewart Sanderson, and John Widdowson, eds. 1985. *Studies in Linguistic Geography: The Dialects of English in Britain and Ireland*. London: Croom Helm.
- Kolb, Eduard, Beat Glauser, Willy Elmer, and Renate Stamm. 1979. *Atlas of English Sounds*. Bern: Francke.
- Kortmann, Bernd, Tanja Herrmann, Lukas Pietsch, and Susanne Wagner. 2004. *A Comparative Grammar of British English Dialects: Agreement, Gender, Relative Clauses*. Berlin: de Gruyter.
- Labov, William. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.
- Labov, William. 1981. "Resolving the Neogrammarian controversy." *Language* 57: 267–308.
- Labov, William. 1966. *The Social Stratification of English in New York City*. Cambridge: Cambridge University Press.
- Labov, William, Sharon Ash, and Charles Boberg. 2006. *Atlas of North American English: Phonetics*,

- Phonology, and Sound Change.* Berlin: de Gruyter.
- Lippi-Green, Rosina. 1997. *English with an Accent: Language, Ideology, and Discrimination in the United States.* London: Routledge.
- Maguire, Warren, April McMahon, and Dan Dediu. 2010. "The past, present, and future of English dialects: Quantifying convergence, divergence, and dynamic equilibrium." *Language Variation and Change* 22: 69–104.
- Markus, Manfred. 2009. "Joseph Wright's *English Dialect Dictionary* and its sources." In *Current Issues in Late Modern English*, edited by Ingrid Tieken-Boon van Ostade, and Wim Van der Wurff, 263–282. Bern: Peter Lang.
- Markus, Manfred, Clive Upton, and Reinhard Heuberger, eds. 2010. *Joseph Wright's English Dialect Dictionary and Beyond.* Bern: Peter Lang.
- Milroy, James, and Lesley Milroy, eds. 1993. *Real English: The Grammar of the English Dialects in the British Isles.* London: Longman.
- Momma, Haruko. 2015. *From Philology to English Studies Language and Culture in the Nineteenth Century.* Cambridge: Cambridge University Press.
- Mugglestone, Lynda. 2007. *Talking Proper: The Rise and Fall of the English Accent as Social Symbol*, 2nd ed. Oxford: Oxford University Press.
- Nerbonne, John. 2003. "Introducing computational techniques in dialectometry." *Computers and the Humanities* 37(3): 245–255.
- Nerbonne, John, and Wilbert Heeringa. 2010. Measuring dialect differences. In *Language and Space: An International Handbook of Linguistic Variation*, Vol. 1 – Theories and Methods, edited by Peter Auer, and Jürgen Schmidt, 550–567. Berlin: de Gruyter.
- Ó Murchú, Mairtín. 1969. "Common core and underlying representations." *Eriu* 21: 42–75.
- Penhallurick, Robert. 2009. "Dialect dictionaries." In *The Oxford History of English Lexicography*, Vol. 2, edited by Anthony Cowie, 290–313. Oxford: Oxford University Press.
- Petyt, Malcolm. 1980. *The Study of Dialect: An Introduction to Dialectology.* London: Longman.
- Podesva, Robert, and Devyani Sharma, eds. 2013. *Research Methods in Linguistics.* Cambridge: Cambridge University Press.
- Puttenham, George 1589. *The Arte of English Poesie.* London: Richard Field.
- Pyles, Thomas, and John Algeo. 2004. *The Origins and Development of the English Language*, 5th ed. New York: Harcourt Brace Jovanovich.
- Ray, John 1674. *A Collection of English Words not Generally Used.* London: Printed for Christopher Wilkinson. Reprinted in 1969 by The Scolar Press (Menston).
- Rickford, John. 1999. *African American Vernacular English: Features, Evolution, Educational Implications.* Oxford: Blackwell.
- Ruano-García, Javier. 2010. *Early Modern Northern English Lexis: A Literary Corpus-Based Study.* Bern: Peter Lang.
- Schilling, Natalie. 2013. "Surveys and interviews." In *Research Methods in Linguistics*, edited by Robert Podesva, and Devyani Sharma, 96–115. Cambridge: Cambridge University Press.
- Schneider, Edgar. 2003. "The dynamics of New Englishes: From identity construction to dialect birth." *Language* 79(2): 233–281.
- Schneider, Edgar. 2007. *Postcolonial English: Varieties around the World.* Cambridge: Cambridge University Press.
- Shorrocks, Graham. 2000. "Purpose, theory and method in English dialectology: Towards a more objective history of the discipline." In *Debating Dialect: Essays on the Philosophy of Language*, edited by Robert Penhallurick, 84–105. Cardiff: University of Wales Press.
- Shorrocks, Graham. 2001. "The dialectology of English in the British Isles." In *History of the Language Sciences*, Vol. 2, edited by Sylvain Auroux, Ernst Koerner, Hans-Josef Niederehe, and Kees Versteegh, 1553–1562. Berlin: de Gruyter.
- Skeat, Walter. 1912. *English Dialects from the Eighth Century to the Present Day.* New York: Kraus Reprint Co.
- Stein, Gabriele. 1997. *John Palsgrave as a Renaissance Linguist: A Pioneer in Vernacular Language.* Oxford: Clarendon.
- Szmrecsanyi, Benedikt. 2013. *Grammatical Variation in British English Dialects: A Study in Corpus-Based Dialectometry.* Cambridge: Cambridge University Press.
- Szmrecsanyi, Benedikt. 2012. *Grammatical Variation in British English Dialects: A Study in Corpus-Based Dialectometry.* Cambridge: Cambridge University Press.
- Tolkien, John R. R. 1934. "Chaucer as a Philologist: *The Reeve's Tale.*" *Transactions of the Philological Society* 33(1): 1–70.
- Trudgill, Peter. 1990. *The Dialects of England.* Oxford: Blackwell.
- Trudgill, Peter. 2004. *New Dialect Formation: The Inevitability of Colonial Englishes.* Edinburgh: Edinburgh University Press.
- Trudgill, Peter. 2008. "Colonial dialect contact in the history of European languages: On the irrelevance of identity to new-dialect formation." *Language in Society* 37(2): 241–254.

- Turner, James. 2014. *Philology: The Forgotten Origins of the Modern Humanities*. Princeton, NJ: Princeton University Press.
- Upton, Clive, Stewart Sanderson, and John Widdowson. (1987). *Word Maps: A Dialect Atlas of England*. London: Croom Helm.
- Upton, Clive, and John Widdowson. 1996. *An Atlas of English Dialects*. Oxford: Oxford University Press.
- Viereck, Wolfgang, and Heinrich Ramisch. 1991. *The Computer Developed Linguistic Atlas of England* (2 vols.). Tübingen: Niemeyer.
- Wagner, Heinrich. 1958–1964. *Linguistic Atlas and Survey of Irish Dialects* (4 vols.). Dublin: Institute for Advanced Studies.
- Wakelin, Martyn, ed. 1972. *Patterns in the Folk Speech of the British Isles*. London: Athlone.
- Wakelin, Martyn. 1977. *English Dialects: An Introduction*, 2nd ed. London: Athlone.
- Wakelin, Martyn. 1987. "The treatment of dialect in English dictionaries." In *Studies in Lexicography*, edited by Robert Burchfield, 156–177. Oxford: Clarendon.
- Wales, Katie. 2010. "Northern English in Writing." In *Varieties of English in Writing: The Written Word as Linguistic Evidence*, edited by Raymond Hickey, 61–80. Amsterdam: Benjamins.
- Walker, James. 2013. "Variation analysis." In *Research Methods in Linguistics*, edited by Robert Podesva, and Devyani Sharma, 440–459. Cambridge: Cambridge University Press.
- Weinreich, Uriel. 1954. "Is a structural dialectology possible?" *Word* 10: 388–400.
- Wells, John. 1982. *Accents of English* (3 vols.). Cambridge: Cambridge University Press.
- Wolfram, Walt, Carolyn Adger, and Donna Christian. 1999. *Dialects in Schools and Communities*. Mahwah, NJ: Erlbaum.
- Wolfram, Walt, and Ralph Fasold. 1974. *The Study of Social Dialects in American English*. Englewood Cliffs, NJ: Prentice-Hall.
- Wolfram, Walt, and Natalie Schilling-Estes. 2004. "Remnant dialects in the Coastal United States." In *Legacies of Colonial English: Studies in Transported Dialects*, edited by Raymond Hickey, 172–202. Cambridge: Cambridge University Press.
- Wolfram, Walt, and Natalie Schilling-Estes. 2006. *American English: Dialects and Variation*, 2nd ed. Oxford: Blackwell.
- Wolfram, Walt, and Ben Ward. 2005. *American Voices: How Dialects Differ from Coast to Coast*. Oxford: Blackwell.
- Woodard, Roger. 2008. "Greek dialects." In *The Ancient Languages of Europe*, edited by Roger Woodard, 50–72. Cambridge: Cambridge University Press.
- Wright, Joseph, ed. 1898–1905. *English Dialect Dictionary* (5 vols.). Oxford: Frowde.
- Wright, Joseph, ed. 1905. *English Dialect Grammar*. Oxford: Frowde.

2 The Dialect Dictionary

JACQUES VAN KEYMEULEN

2.1 Introduction

This chapter discusses the making of dialect dictionaries. We dwell on user-oriented meta-lexicographical considerations, and on the macrostructural and microstructural options that ensue. Special attention is devoted to fieldwork procedures for unwritten language varieties. We address the basic questions of fieldwork: what, where, who, how, and how much?

Dialect will be defined here as a geographically determined language variety (local or regional), which is essentially only orally transmitted and is relatively isolated from the roofing standard variety (if any).¹ It is clear that this type of variety is quickly disappearing in most modern societies, under the pressure of standard language forms. We do not dwell on the question of whether a given language variety is a *dialect* or a *language*. *Language* as against *dialect* (of that language) is, in the first place, a sociological notion rather than a linguistic one, and has to do with the level of codification of pronunciation, orthography, and grammar, usage in official and formal situations, usage in writing, and acceptance by a community of speakers (see Haugen 1966). Some dialects, however, have reached the status of “regional language” (e.g., Low German in Germany, or Limburgian in the Netherlands) under the *European Charter for Regional and Minority Languages*, which was adopted in 1992 under the auspices of the Council of Europe. These enjoy some of the characteristics of “official” languages.²

In defining *dialect* as a geographically determined language variety, for practical reasons we follow the continental European tradition, although the content of the term is widened—especially in the Anglo-American literature—to include *any* variety of a given language, even its standard form.³ For the same reasons, we do not include the lexicography of socially determined lexical varieties such as sociolects (see Green, *in press*), genderlects, ethnolects, technolects, or other specialized vocabularies, although we think that many of the observations below, especially those pertaining to fieldwork, may be useful for them as well. Nor do we include the lexicography of so called “pluricentric languages,” that is, “languages with several interacting centres, each providing a national variety with at least some of its own (codified) norms” (Kloss 1978, 66–67; see also Clyne 1992). Laureys (1997) coined the useful term *natiolect* for “a national standard of a language that is spoken in more than one state.” Natiolects (e.g., Irish English, Austrian German, or Belgian Dutch) do not only exist in Europe, but are very typical forms of geographically defined varieties within world languages such as English, French, Spanish, or Portuguese. Attitudes toward natiolectal

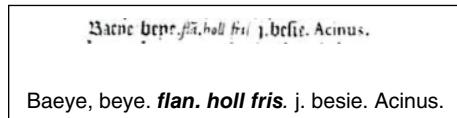


Figure 2.1 Headword *baeye, beye* “berry” (Kiliaan 1599).

vocabulary, and hence its treatment, vary greatly. In some cases, separate dictionaries account for this type of geographical variation (e.g., the *Dictionary of South African English on Historical Principles* [DSAE] for South African English, or the *Variantenwörterbuch des Deutschen* of Ammon *et al.* 2004); sometimes labels are used in the comprehensive dictionaries of the standard languages (e.g., *Belgisch Nederlands* or *België* for the Belgian Dutch words in the *Van Dale* dictionary of standard Dutch). A recent example of a balanced approach to pluricentricity is the *Prisma Handwoordenboek* (Hofman 2014) for Dutch. Thanks to a thorough analysis of digital text corpora, the dictionary is able to accurately label the words which are *natiolectally* restricted on a continuum between *only* “Netherlandic Dutch” and *only* “Belgian Dutch.” In any case, considering natiolectal differentiation as *dialectal* nowadays is increasingly frowned upon as being politically incorrect, since natiolectisms are used by the educated part of a language community in formal oral and written settings. Natiolectal vocabulary is part of a co-standard, not of a sub-standard.

We will also focus on dialect lexicography proper, and not on the way in which dialect words are included (and labeled) in standard-language lexicography. Dialectal or regional words may be incorporated in the latter context for a large number of reasons (e.g., to make literature in substandard language varieties more comprehensible). A very early example of interest in the geographical dimension of a vernacular is the multilingual *Nomenclator Omnium Rerum* (1567) of Hadrianus Junius, who added regional labels to the Dutch words. This tradition—which was to be typical of Dutch lexicography thereafter—was first brought to a pinnacle in Kiliaan’s comprehensive dictionary.⁴ In his Dutch-Latin *Etymologicum Teutonicae Linguae* (1599), he not only translated the headwords into a number of European languages, alongside the definition in Latin, but in a number of cases he also added labels such as *fland.* (Flemish), *holl.* (Hollandic), *fri.* (Frisian), and the like, in order to clarify the geographical origins of his data (Figure 2.1).

In what follows, we will firstly discuss a number of meta-lexicographical considerations concerning dialect lexicography. We then turn to the way the overall wordlist of a dictionary is selected, which leads to its macrostructure. Following this, we devote our attention to the microstructure, that is, the information that is given about the headwords, and to data collection, which in our case boils down to a discussion of lexicographical fieldwork for unwritten language varieties. The final section of the chapter contains some remarks on new technologies and desiderata for the future. Examples are taken from different languages, but with a bias toward Dutch, and especially southern Dutch dialect lexicography.

2.2 The User’s Perspective: Meta-Lexicographical Considerations (Why, and for Whom)

Dialect dictionaries came into being from the eighteenth century onward, in the wake of emerging scientific interest in language history. Moulin (2010, 593) refers to the interesting lexicographic programme of Leibnitz (1697), which aimed at the examination of all German words, the regional ones included. Large-scale dialect lexicography emerged all over Europe—and especially in the German language area—from the end of the nineteenth century onward, together with modern dialectology. A philological scientific paradigm

prevailed, reinforced by Romanticism, with its stress on naturalness, historical preoccupation, and nationalism. Some dialect dictionaries aimed to enrich the comprehensive national dictionaries, whereas others tried to account for both the language and the culture of mainly rural communities. The attitude toward regionalisms, however, was not always positive. In the wake of the rise of standard language ideologies, some dialect word collections came into existence as lists of "errors." Rézeau (1990, 1471), for instance, mentions Desgrouais (1766), stating: "La source du mal, le patois" ["the source of evil, the dialect," my translation] in "Les gasconismes corrigés." Words labelled "Zuidnl." (southern Dutch) in older issues of the standard Dutch dictionary *Van Dale* were put on lists by the Flemings themselves, in order to warn readers against them. Attitudes (and tolerance) toward geographical and other types of language differentiation may differ from culture to culture (see Rézeau (1990) and Hausmann (1990, 1501) on negative attitudes toward Belgian French words, and Görlich (1990) on changing attitudes toward varieties of English).

Dialect dictionaries may be written for societal as well as for scientific reasons. Dialect speakers may wish to have a reference work of their local or regional language variety for practical or for symbolic motives. A local or regional dictionary is obviously handy when a word is not understood by a non-dialect speaker. Local dialect dictionaries also have a codifying effect (or even purpose) because they usually use a dialect spelling for headwords, thus rendering the dialect "writeable" (see also Bernal and Aymerich, in press). The mere existence of a dictionary is very often considered proof of the respectability of the language variety it describes. Another popular motivation is the safeguarding of the dialect vocabulary for a supposedly dialectless future. With that in mind, it is striking that many amateur dictionaries featuring dialectal headwords do not contain some kind of bilingual (standard > dialect) index, which would indeed disclose the dialect vocabulary to the non-dialect speaker of the future. As things stand now, many dictionaries can only be used by good dialect speakers, because these users are the only ones who can decipher the (sometimes home-made) dialect spelling of the headwords. In any case, without an index by which the dictionary is reversed it is impossible to learn dialect words, because a question like "What is the dialect word for 'butterfly?'" cannot be answered without a standard > dialect index. Many dictionaries also want to describe the (disappearing) culture of the dialect-speaking community together with the vocabulary going with it, which sometimes entails an extensive encyclopedic component in the microstructure of the dictionary.

Scientific motivations, such as safeguarding endangered language varieties for historical or ethnological reasons, coincide to some extent with popular ones. Monotopical dialect dictionaries do not in principle differ fundamentally from dictionaries of standard languages. Since every local dialect can be considered a language system of its own, the same macrostructural and microstructural options may be taken as for any standard language dictionary. The specificity of scientific dialect lexicography, however, resides mainly in the geographical dimension of the dialectal type of lexical variety. The *WVD*, for instance, considers as its minimal goal the recording of reliable (preferably orally collected) lexical information for every lexical area within the zone of investigation; the maximal goal is to obtain lexemes for every locality, in order to be able to draw word maps. The collection of exclusively orally transmitted vocabulary (and the lexico-geographical patterns therein) is complementary to the dictionaries of both the historical and present-day periods, which are mostly based on written text corpora.

Collecting dialectisms or regionalisms in order to warn against them in favor of the standard language is nowadays considered old-fashioned. Scientific dialect lexicography is normally strictly descriptive. Since the target users are generally situated in a hypothesized dialectless future, the metalanguage (i.e., the language used to define and discuss dialectal forms) is normally the standard language. The dictionary should also be structured in such a way that both words and meanings are interpretable for non-dialect speakers in the future.

In this sense, dialect lexicography is a kind of bilingual dialect/standard-language lexicography. In sum, a dialect dictionary has to be based on the assumed needs of the (future) user, and should ideally answer two main questions: (a) “What does word X mean?” and—perhaps more importantly for the future—(b) “How is concept/meaning X expressed?” Alongside the last question, a user may be interested in other microstructural elements (see below).

It goes without saying that the front matter (introduction) of a dictionary should be explicit about all meta-lexicographical options, and, as has already been pointed out, the back matter should incorporate a standard >dialect “bilingual” index.

2.3 Macrostructural Considerations

Dialect lexicography differs from standard-language lexicography in three main ways: (1) it is essentially based on fieldwork, (2) an oral language variety has to be rendered in writing, (3) the geographical dimension of this type of language variation has to be accounted for. Moreover, the lexicographer has to decide between a contrastive or a confrontative approach (the terms are from Wiegand (1977)) with regard to the standard language. In the first case, the dialect dictionary contains only the words—or meanings or other deviating elements—that do not occur in the standard variety. Such a dictionary was formerly known as an *idioticon*. In the latter case, the dialect dictionary is meant to account for the totality of the vocabulary, including all lexical and semantic similarities with the standard language.

Any macrostructural option (e.g., options concerning the inclusion of specialized vocabulary, collocations) that is possible for a standard-language dictionary is also conceivable for a dialect dictionary. Although the focus of many dialect dictionaries was, and is, on the oldest layer of the orally transmitted language, it is possible to bring in social parameters other than age alone. The *Dictionary of American Regional English* (DARE) compiled by Cassidy *et al.* (1985–2013) is a good example of what is possible in this respect. DARE takes account of age, gender, and educational, occupational, and ethnic groups.

2.4 Onomasiological or Semasiological Arrangement

Semasiological dictionaries present and explain the meanings of a given word or phrase, in contrast to onomasiological dictionaries, in which words or phrases are presented as expressions of semantically linked concepts (meanings, ideas, notions, word families, and similar relationships).⁵

Dialect dictionaries may have a local or a regional geographical scope, but even “local” dictionaries, especially those for urban dialects, very often account for geographical differentiation, along with social variation of different types. The titles of regional dictionaries may sometimes mislead the user, however. The words included in the *Westvlaams Idioticon* (West Flemish Dictionary; De Bo 1873), for instance, are not necessarily general in the West Flemish dialect, but are recorded *somewhere* in the Belgian province of West Flanders. Sometimes it is also unclear whether the title of a dictionary is to be interpreted in a purely geographical sense, or refers to a linguistic entity proper. Many words in the *Woordenboek van de Vlaamse Dialecten* (WVD, *Dictionary of the Flemish Dialects*), for instance, are not Flemish but Brabantian, since the area of investigation in the east is based on the administrative boundary of the province of East Flanders. Brabantian dialects are spoken in the eastern fringe of that province.

The smaller the geographical scope of a dictionary, the more a semasiological (alphabetical) arrangement becomes evident. If the geographical scope is widened, onomasiology

comes into play. Moulin (2010, 598 ff.) mentions, for example, the *Wörterbuch der obersächsischen Mundarten*, in which an alphabetical arrangement is disrupted by “nests” of cognates (see below). Many regional dictionaries try to account for the geographical component of the vocabulary by resorting to the mesostructure (i.e., the cross-referencing system) of the dictionary.

Grasmücke wie schriftdt. Synonyma: *Fliegenfänger, Fliegenschnapper, Fliegenstecker, Gartengrasmücke, Grasetotschke, Grasshopper, Grasschwappe, Heimchen, Heupferdchen, Hipplich, Hoppe-käppel, Klapperpferdel, Klostermönch, Mönchsgrasmücke, Müller-chen, Schwarzblättel, Spotvogel, Sproachmeesterle, Zaungrasmücke, Zirpvogel; Groasmicke H 69s. Karte Libelle*

The entry *Grasmücke* “warbler” taken from the *Schlesisches Wörterbuch* (SW), with a “nest” of words denoting the same concept (formerly) used in Silesia.

Sometimes word maps unburden the dictionary of an excessively cumbersome reference system.⁶ Alternatively, dictionaries may simply take the shape of an onomasiologically arranged word atlas for a large region, as is the case for the three regional dictionaries covering the three southern Dutch dialect groups (WVD for Flemish,⁷ WBD for Brabantic, and WLD for Limburgian dialects), which are—notwithstanding their titles—in fact not dictionaries but geographically oriented inventories of word usage. They are thematically assigned, each fascicle containing a row of concepts with lists of heteronyms assigned to every concept (see Van Keymeulen 2003). The word collection thus presented offers in the first place material for dialect-geographical research, but detailed semantic information (e.g., the polysemic structure of the meaning) is hard to render in such types of publication.

2.5 Onomasiological Arrangements

An onomasiological arrangement is “inspired by the idea that the reality around us can be roughly divided into a system of concepts [...] That system can be, for instance, logical, philosophical or pragmatic” (Van Sterkenburg 2003, 127–8).⁸ Below, we discuss as an example the overall onomasiological taxonomy of the aforementioned southern Dutch dialect dictionaries. The taxonomy is heavily indebted to the tradition that started with Hallig and von Wartburg (1952). Their *Begriffssystem* (“conceptual system”) was adapted for dialect lexicography (see van Keymeulen 2003) (Figure 2.2). Generally speaking, man is placed at the center of things, and reality is assigned to him in ever broadening circles:

An onomasiological taxonomy of “reality” in our case is a classification of concepts, not of words (or “things,” for that matter), and is thus in principle language-independent. It may be used and adapted for new cultural contexts—and for new dictionaries. The question of whether “chicken” is a bird or a piece of food (and should hence be placed in either Section 4.2 or Section 2.3 in Figure 2.2) can be solved by considering the taxonomy as a dynamic structure instead of a static one. The same concept, indeed, may occur in different classes, depending on the viewpoint one takes. Sometimes the same concept is lexicalized differently, depending on the section it belongs to: in some Flemish dialects a potato is called *erpel* when grown as a plant, but *patat* when cooked and eaten.

Any onomasiological taxonomy should be explicit about its specific purpose. In the case of the regional dictionaries of southern Dutch, the purpose of the arrangement is to account for “the concrete coherence of things in daily life” (Weijnen and Van Bakel 1967, 40; my translation) as experienced by the dialect-speaking community. The main framework of the conceptual classification in Figure 2.2 has, in the author’s opinion, universal value, and

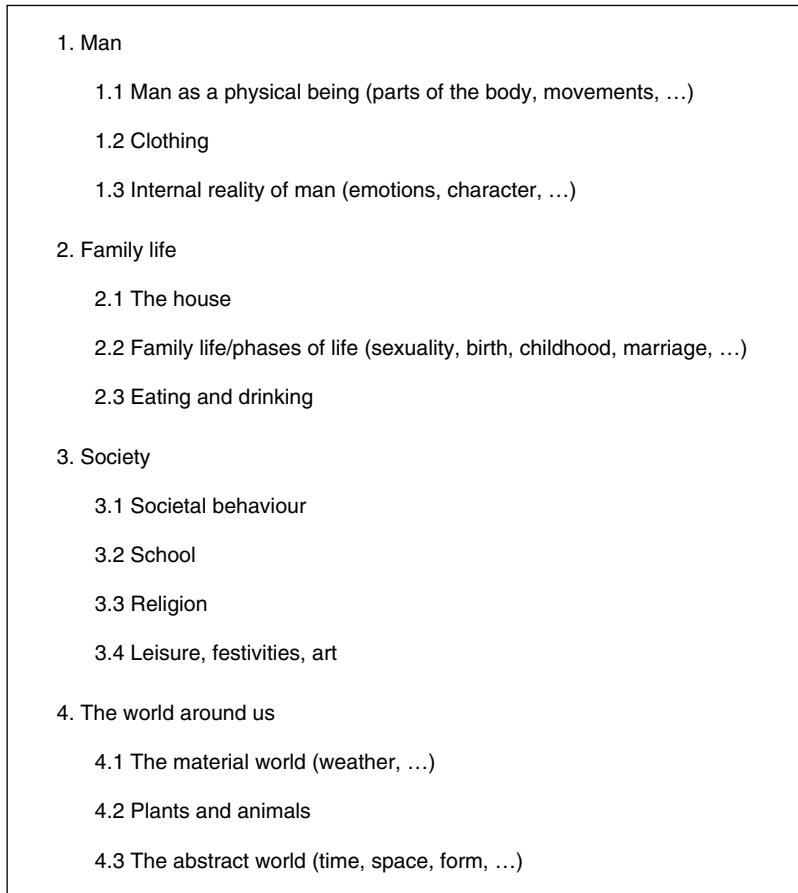


Figure 2.2 Onomasiological classification.

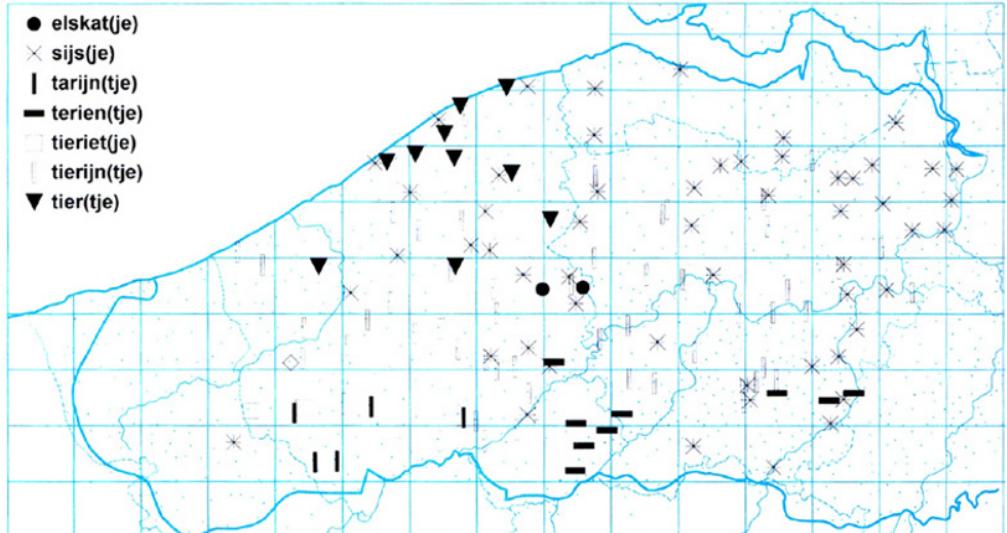
adaptations with regard to specific cultural contexts are only needed at its lower levels. The functionality and frequency of the concept in the everyday life of the average dialect speaker are used as guiding principles in the assignment of a concept to a specific section. The concepts relating to sexuality, for instance, are therefore placed in Section 2.2 “Family life” and not in Section 1.1 “Man as a physical being.” A tomato is a plant, but is more frequently thought of as being eaten than grown; hence, “tomato” is in the first place assigned to Section 2.3 “Eating and drinking.”

Onomasiological dictionaries of the type mentioned above should be supplemented with alphabetical indices. The new technologies available offer many solutions for mesostructural needs of all kinds. For an example, see Figure 2.3.

2.6 Semasiological Arrangement

Most dialect dictionaries are semasiologically (i.e., alphabetically) arranged; the standard-language lexicography is—often implicitly—used as an example. In that way, they can efficiently render microstructural information, especially information pertaining to semantics. Dialect dictionaries covering more than one local dialect have to find a solution to account

SIJSJE : OUD MATERIAAL



SIJSJE

Het sijsje (*Carduelis spinus*) is een kleine, geel en groen gekleurde vink met een zwarte kruin. Het houdt zich het liefst op in elzen en berken, waar het op zoek gaat naar zaden en insecten. Zie afb. 28.



Afb. 28. Sijsjes hangen soms ondersteboven in elzentakken.

WVD 66M (1995), 230 ; WVD 66 (1995), 128 ; N 9 (1961), 6 ; Materiaal Houwen

(1965-1990), 79 ; Materiaal Menschaert (1991), 100 ; Vandecasteele (1978), 191 ; Wielewaal (1952), 333 ; DC 6 (toegift) (1938), 23 ; ZND 6 (1924), 58 ; Willems (1886), 16. Volk en Taal 3 (1890), 103 ; 7 (1895), 102 ; Annalen Land van Waas 73 (1970), 115 ; Heemkring Zele (1989), 103. Zangvogels (1944), 312 ; Avifauna (1975), 17 ; Dialect OZV (1982), 180 ; Nederlandse Vogelnamen (1995), 255.

bosmees : Kruibeke.

elskat(je) : Wingene, Kruishoutem.

♦ Schuiferskapelle, Wingene, Ruijselede.

elsmee : Wdb : Loquela : elsmeese, Destelbergen.

elssij : Vrasene.

elstierientje [*elstrientje*] : ♦ Anzegem.

elstierijntje [*elstrientje*] : Tielt.

♦ Zwevezele. Wdb : De Bo : elstrijtje.

elsvink : Wdb : De Bo.

elzentierijntje [*elzentrientje*] : Wakken.

♦ Ruddervoorde.

elzepikker [ook : *elzepukker*] : Wdb : De Bo.

grastierijntje [*gerstrientje*] : ♦ Tielt.

for diatopical lexical differentiation (see above). Sometimes, though rarely, onomasiological re-arrangements (e.g., word fields) are annexed to alphabetical dictionaries. Moulin (2010, 598) notes for the major wide-area German dictionaries the same “skeletal structure”: namely, a “word-geographic principle” and “the organizational unit of a headword with grammatical information, definitions, illustrative examples, distributional information, and cross-referencing systems.”

2.7 Microstructural Considerations

Microstructural elements of all kinds, such as descriptions of meaning, parts of speech, example sentences, and so on, are best rendered when words, rather than concepts, are taken as the point of departure in the lexicographic description. Since in principle a dialect does not differ from a standard language, the same microstructural options may be taken as those chosen for any standard language dictionary (those concerning definitions, the inclusion of phonological or grammatical information, collocations, example sentences, etymology or labels of any kind). The metalanguage of the dictionary is usually the standard language, although there are exceptions to this rule. A dictionary author indeed may wish to demonstrate the value of a “substandard” variety by rendering both headwords and metalanguage in the very same variety.

The specificity of the microstructure of dialect lexicography resides mainly in the form of the headwords and in the treatment of pronunciation and the geographical diffusion of the vocabulary. In order to form dialectal headwords, a formerly unwritten language variety has to be taken down in writing. Some authors invent orthographies, usually based on the spelling of the standard language, with some adaptations and additions (usually diacritic symbols) in order to try and render the dialect pronunciation in the form of the headword itself, instead of adding it in the microstructure.

râschittink zn. v.; mv. -gen: (boog-schutter) rouwschieting ter ere van over- leden lid. [FFP 22].

rastieël zn. o.: ruif voor paarden. [rasteel; Ofr. rastelier].
z. hoeë.

Example of homemade dialect spellings for headwords, in an attempt to render dialect pronunciation. The “Dutchified” (i.e., normalized toward Dutch orthography) headwords would be **rouwschieting** and **rasteel** respectively (Pletinckx 2003).

Dictionaries with a wide geographical scope have to abstract away from local dialects in order to summarize across a sometimes very large amount of phonological variance in the form of the headword itself. The standardization of the entry form sometimes takes the shape of “framing” the spelling of the dialect headword as if the word belonged to the standard language.

Dialect lexicography should ideally pay detailed attention to the geographical component of all microstructural elements, especially on the lexical, phonological, and semantic levels, if the sources permit. In many cases, the geographical component of the microstructure leaves much to be desired, since sources or preparatory studies are simply lacking (see also Niebaum 1989–1991, 663). In the absence of geographical labels, the user should be aware that words and microstructural elements (with definitions of meaning included) may not apply across the whole of the area covered by the dictionary.

Only a tiny number of microstructural elements, for example the labels for the parts of speech, can be added by the lexicographer him- or herself. Example sentences are usually invented by the author, instead of being taken from sources, in order to illustrate the word meaning; the best example sentences also illustrate dialectal morphosyntactic particularities. Since many lexicographers want to reveal the culture of the dialect speaker, the encyclopedic component of the microstructure is often very extensive (e.g., technical descriptions, illustrations, etc.).

2.8 Data Collection by Fieldwork

2.8.1 Introduction

Dialect lexicography may be based on oral or written sources. Some “regional languages” of different types do enjoy a—sometimes very old—written tradition, the quantity and quality of which varies from language to language (e.g., Low German, Scots, Occitan,⁹ Jamaican English, Papiamento). Dialect texts can be used as sources for dictionaries, but the lexicographer has to be aware of the fact that the “dialecticity” level of the texts may be unclear. The language may vary between dialects with standard language interference and standard language with a dialect flavor.¹⁰ Transcripts of dialectal audio material could in principle serve as a source as well, but it would be extremely time-consuming to establish a large enough corpus for a comprehensive dictionary. Sometimes dictionaries take older dictionaries as their basic material, or already existing wordlists are revised and enlarged. We do not discuss this “plagiarism in alphabetical order”—as it has been jocularly called—here.

In what follows, we focus on synchronic fieldwork for a purely oral language variety.¹¹ We try to propose answers to the five main questions concerning lexicographical data collection: what? where? who? how? and how much?

2.8.2 What?

For a purely oral language variety, both the words and the microstructural elements have to be collected by fieldwork, depending on the selections based on the meta-lexicographical options. The lexicographic focus is normally on words and meanings; dialect lexicography is also typically preoccupied with pronunciation. For some microstructural elements (e.g., morphology), lexicographers very often try to rely on already existing research, if there is any.

2.8.3 Where?

The geographical scope of a dictionary should be well defined. Many dictionaries have too ambitious a title in respect of geography. In a regional dictionary, geographical representativeness should be an important aim, that is, every lexeme/meaning within the area of investigation should be present. This can be guaranteed by establishing a premeditated geographical network of measuring points, which implies that a pilot study designed to determine the lexico-geographical pattern of the area has already been carried out. It is a scientific requirement to be as explicit as possible with regard to the localization of all data. A hybrid dictionary-cum-atlas ideally has one good data point for every locality. This ideal is seldom met, however.

2.8.4 Who?

Informants should be chosen in accordance with the macrostructural options. Lexical knowledge, as it happens, is distributed unevenly in the language community. In traditional dialectology, “NORMs” (Non-mobile Older Rural Males) were usually chosen as informants

(Chambers & Trudgill, 1998, p. 29), since the main aim was to collect the lexicon of the oldest dialect layer. The selection of informants is not always an easy job. The lexicographer may have to call in the help of an intermediary who takes an interest in the project and is in a position to introduce the researcher to possible informants. Schoolteachers and clergymen/-women are very suitable as intermediaries because they are normally highly respected and know a lot of people.

Every informant should meet both objective and subjective criteria. The objective criteria have of course to do with the sociological parameters with regard to the macrostructural options. In any case, the relevant biographical information for every informant should be carefully inquired after and noted down. The subjective criteria can be summarized as follows: interest in the dictionary project, intelligence, introspective capacity in matters of language and talkativeness. Cooperation with the researcher always implies some degree of education on the part of the informant: s/he has to understand the purpose of the project and, in the case of inquiries by correspondence, has to be able to write and read. Men are considered more suited as informants for substandard language varieties than are women; Chambers and Trudgill (1998, p. 61) call the observation that women on average use a greater frequency of higher-status variants than men “the most strikingly consistent finding of all to emerge from sociolinguistic studies in the industrialised western world.”

Informants should always be tested. They evidently want appreciation for their work (e.g., getting informed about the results of the project). Normally, informants are volunteers; if they are paid they tend to get too cooperative, and may deliver nonce answers.

2.8.5 How?

Data collection methods vary according to the language component investigated (phonology, morphology, syntax, lexicon), the relative validity of the elicited data, and the feasibility of data collection being carried out systematically. All methods aim to get as close to “spontaneous speech” as possible. Some methods, however, use experimental settings and produce highly structured data (for eliciting methods in dialectology in general, see also Seiler, 2010). In what follows, we briefly comment on the most important collection methods in lexicography.

2.8.6 (Self-)Observation

It is of course an important advantage if the lexicographer is describing his or her own dialect, although a sound distrust of one’s own knowledge is advisable; self-observation should be checked with good informants. This is especially the case for regional dictionaries, since it is impossible to have good intuitions about the totality of the geographical differentiation within the area of investigation. The lexicographer’s introspection is used to deliver or assess data (see Bergenholz & Mugdan, 1990, p. 1613).

Many lexicographers are attentive listeners, and carry with them notebooks in order to note down words and sentences they overhear. This may yield useful information (e.g., lively example sentences), provided the overheard individual was indeed a representative speaker of the dialect. Many amateur dictionaries are based on material collected in this way, sometimes over many years. In many cases, dictionaries are based on the observations (and notes) of a whole group of volunteering dialect speakers.

Corpora of recordings of so-called “free speech,” whereby a good dialect speaker is invited to talk about something s/he is interested in, may render highly spontaneous language material, which is especially suitable for phonological or syntactic research. For lexicographical purposes, though, this method has only a limited value, as the resulting

corpora are too small (see e.g., Freiburg English Dialect Corpus (FRED), Kortmann & Wagner, 2005). The major disadvantage of observation is that, for obvious reasons, negative evidence is impossible: words that by chance do not occur in the corpus of course cannot be observed. Transcriptions may come in handy for example sentences, or for citations of the closed word classes (which are relatively highly frequent and hard to elicit via questionnaires).

Transcribing dialect recordings is very time-consuming. As time goes by, the transcription of existing collections (e.g., the recordings of the Dutch dialects at the Meertens Institute in Amsterdam and at Ghent University)¹² becomes ever more urgent. Dialect recordings are best transcribed by a native speaker of the dialect, who understands not only the traditional dialect itself, but also the subject matter of the conversation. Transcribers who combine both skills are becoming rare.

2.9 Purposive Systematic Fieldwork

Nearly all major dialect dictionaries are based on purposive systematic fieldwork aimed at investigating the lexical knowledge of informants, using questionnaires administered orally (direct method) or by correspondence (indirect method). Purposive fieldwork is the best method of inquiring into passive dialect knowledge; in dialect loss situations, many words only survive in the memories of members of the oldest generation and have disappeared from active usage. The warning of Seiler (2010, 514) should always be taken to heart: "It must be kept in mind that elicitation procedures create artefacts (task effects, repetition effects, order effects) arising from the unnatural situation the informant is exposed to." It is in fact these artefacts that are the source of information. Methods may differ according to the type of vocabulary investigated; some elements of the lexicon are more easily collected by explicit questioning than others (open versus closed word classes, say, or words with concrete versus abstract meanings), because the introspective capacity of a language user differs according to the different lexical types. The indirect method is considered better for taboo words. Purposive fieldwork normally is carried out thematically, although the ensuing dictionary is alphabetical. Below, we give a brief overview of fieldwork methods used in dialect lexicography. Fieldwork methodologies are also discussed in handbooks for anthropology or sociolinguistics.

2.10 Oral Investigation (Direct Method)

The advantages of oral investigations are evident. The fieldworker controls the interview situation: clarification, feedback, taking notes, and recording are possible. Good phonetic data can only be gathered by this method. The most important disadvantage is the time-consuming nature of the method; it is, however, nearly always used at the initial stages of a project, when the area of investigation is being explored and questionnaires tested. The so-called "observer's paradox" (Labov 1972, 209) (i.e., "spontaneous" speech is influenced by the very presence of the researcher) can to an extent be overcome by putting the researcher in a minority position: interviewing a (small) group of informants, who know each other and are used to speaking the sought-after language variety among themselves, can prove very fruitful, especially because the informants can correct each other. The "intersubjectivity" of the group is thus used to yield valid data that is as close to "objective" as possible. Of course, the fieldworker has to take care to control the group dynamics (e.g., preventing certain informants from becoming too dominant, or inviting more reticent informants—often women—to participate).

The creation of both a social and a scientific "common ground" between interviewer and interviewee(s) is very important. They should—implicitly or explicitly—agree on the aims of

the interview and the definitions of central notions (e.g., what “dialect” is). The fieldworker should also have a sound knowledge of the preconceptions, motivations, and language attitudes of the informant, in order to understand that person’s reactions to questions.¹³ In the case of cooperation on a regular basis, an informant can ideally be turned into a language consultant through on-the-job training. The contact language may cause problems, however; sometimes intermediaries have to be asked to conduct the interview in the presence of the researcher.

A purposive systematic oral investigation is normally structured via a questionnaire (see below), which serves as a steering instrument for a conversation about a specific theme. The task of the researcher is to analyze this meta-lexicographical discourse after it has been collected. Van Keymeulen (1986) investigated some kinds of explanation strategies used by informants: translations into standard language, examples/situations of usage, pointing to things, antonyms/synonyms, naïve analytical definitions, and so forth. In the exploratory stages of the research, an informant should be given enough room to comment or to make associations.

2.11 Investigation by Correspondence (Indirect Method)

Although the direct method is always superior, investigation by postal or electronic correspondence is well suited to (large-scale) lexicographical research. It renders good, and comparable, data. Informants have time to think and the observer’s paradox is circumvented. Feedback is, however, not possible. A major problem concerns how to interpret the notation of dialect words by unskilled informants, who use the standard spelling system or a home-made orthography. The interpretation of these data entails a good knowledge of the dialect phonology on the investigator’s part.

2.12 The Questionnaire

Questionnaires are used in both the direct and the indirect method. In the former case, they structure a meta-lexicographical discourse between fieldworker and informant; in the latter, the questionnaire needs to be designed with the utmost care, since no immediate feedback is possible. Questionnaires are normally thematically organized, and for the most part the metalanguage is the standard variety, which may cause problems for a dialect-speaking informant. Some projects have made extensive use of intermediaries who assisted the sometimes illiterate informants in filling out questionnaires.

The preparation of a lexicographical questionnaire in the case of a relatively unknown dialect has the following phases: (1) exploration of the extra-linguistic reality (ethnological investigation); (2) establishing an inventory of concepts; (3) selection of lexically relevant concepts (e.g., the folk taxonomies in biology); and (4) establishing the way word meanings are stored in the mind of the informant, so as to construct good elicitation techniques. This preparation is normally done orally. Luckily, most lexicographical projects do not begin from scratch, although in some cases (e.g., specialized lexica of traditional crafts) it is a dialect dictionary which is the first description, not only of the words, but also of the craft itself.

2.13 The Structure of the Questionnaire

The structure of the questionnaire should demonstrate the empathetic capacities of its compiler, who has to try to formulate questions and tasks in such a way as to lead the informant to a good answer. The informant’s biographical data (and that of the intermediary, if any)

should be carefully asked for in order to be able to evaluate his/her answers afterward; the most important social parameters are age, gender, occupation, level of education, where the informant was born and bred, and the language (variety) of parents and partner. A questionnaire should bear the date of its composition and ask for the date of filling-in, especially when the age rather than the birth year of the respondent is asked for. In some long-lasting dictionary projects, the same questionnaire is distributed over many years to new informants, and it has to be unambiguously clear how old the informant was when the questionnaire was filled in. Age, indeed, is the most relevant sociological parameter in dialect loss.

A thematic arrangement is advisable because it stimulates introspection and the memory of the respondent. In an introduction (which is, alas, often not read) the researcher has to communicate his or her instructions. It is important to point out that questions can be skipped over in order to avoid nonce answers. The motivation of the informant is strengthened by starting with easy questions. Questions that are likely to raise the strongest emotions (e.g., the words for “greed”) should come before questions about more neutral lexemes (e.g., the words for “economical”). A questionnaire should not be too long; about 150 questions is a good average—such that it can be completed in approximately one hour. Questions should not be too long or too complex, in order to avoid reactions to only a part of the question or task. Control questions are also advisable: one may ask for the same word or meaning a second time, but in a different way.

2.14 Question Types (+ Examples)

Constructing good questions and tasks is a lexicographical art. The indirect method, in particular, can be considered an “experimental setting” that renders comparable data. Very often it is the large scale of the inquiry that validates the data, by means of mass (geographical) comparison (Seiler 2010, 524), although it is possible that all informants are misled in the same way by a badly-constructed question. In what follows, we summarize the main question types used in lexicography, with an example and a short comment accompanying each.

- *Encyclopedic questions* (e.g., “What kind of pears are grown here?”)

Encyclopedic questions deal with the extra-linguistic world; they are typical of the ethnological, preparatory phase of lexicographical research. Lexicographic questions in questionnaires, however, may be masked as encyclopedic ones, as in “How do you feel when everything seems to spin round?” (answer: *dizzy*). It is a useful technique to use as a way of avoiding a complex onomasiological question (see below).

- *Onomasiological questions* (e.g., “What is the word for the small animal covered with brown and white spines?” (*hedgehog*)

An onomasiological question describes a concept (not a meaning) and asks for a lexical expression that can be used to refer to the concept. The technique is widely used for the open word classes. The description of the concept may take different shapes: pictures of prototypical examples (for *concreta*) or analytical definitions (of the *genus proximum + differentia specifica*-type).¹⁴ Complex definitions should be avoided, and replaced by other methods, such as framing a sentence (describing a situation) in which a word should be filled in. Analytical definitions may invoke a “quiz reaction” leading to the corresponding standard word. Suggesting answers, for example, in order to investigate the geographical distribution of a word, is not without its dangers, but it may be used to stimulate the subject’s introspection concerning passive knowledge of disappearing words. Suggestions can be presented in an open or a closed series; in the former, the respondent may add yet another word, in the latter, s/he has to select an answer.

An efficient onomasiological way of eliciting words using the direct method is so-called “picture telling”: the informant is invited to comment on what can be seen in a picture, which may be prepared in such a way as to elicit the sought-after lexical, morphological (etc.) data (see Seiler 2010, 517 ff). Images are in general more likely to prompt good answers than is (complex) verbal definition.

- *Semasiological questions* (e.g., “What does the word *evet* mean?” (*evet* is “newt” in some dialects of southwestern England))

A semasiological question presents a word and a “content” response is asked for (conceptual meaning, connotation, register, usage, etc.). The answer often takes the shape of a translation into the standard language or an attempt at an analytical definition. Sometimes the informant is invited to produce collocations (e.g., expressions, proverbs) or example sentences containing the given word. The semasiological type of questioning requires more thinking and noting down on the part of the informant, which sometimes results in excessively short or ill-conceived answers. Semasiological questions presuppose a word collection of which the semantic characteristics or the geographical distribution are investigated.

A special kind of a semasiologically oriented method is the “acceptability question”: as a closed yes/no-question it inquires about the mere existence of a given word (one with a particular meaning), or it asks for an acceptability evaluation on a Likert scale.

- *Completion tasks* (e.g., “When everything seems to spin round, you feel _____ ?”) In a completion task, a context (sentence) is framed in which a lexical expression should be used. It renders good results if the sentence is unambiguous and leads the informant unequivocally to the sought-after lexeme.
- *Translation tasks* (e.g., *newt* > ?) Translation tasks are not often used in lexicography because the stimulus in the standard language is often just copied or just rendered in dialect phonology.
- *Cyclicity and feedback* Lexicographic fieldwork carried out in a relatively unknown language area should take advantage of the possibility of using cyclical questioning procedures. The answers to onomasiological questions can be used as input to semasiological ones, and vice versa, which may provide new words and meanings. Oral and written methods may control each other.

2.15 How Much?

The question of the saturation of a word collection is difficult to answer. When have enough data been collected? When is the word collection representative of the dialect vocabulary we want to describe? There is, however, always a point at which the continuation of fieldwork (more questions, more informants, and more localities) is felt to be disproportionate to the emergence of new data. With regard to the geographical representativity of the word collection of a regional dictionary, the minimal aim is to record every lexeme/meaning at least once. The maximal aim is to have good data for every locality, in order to be able to draw word maps. The most important advantage of a thematic fieldwork procedure in fact resides in its potential to evaluate the completeness of the word collection.

2.16 New Technologies and Desiderata for the Future

The digital revolution of course did not leave dialect lexicography unaffected. The main difference from digital standard-language lexicography is that dialect lexicography has to account for lexical geography, combining databases and cartographic tools. Digital

dictionaries may combine dictionary, lexical database, sound samples,¹⁵ word maps, etymological information, and encyclopedic information into one gigantic hypertext enterprise (+ search engines).¹⁶ A digital dialect dictionary is free of space constraints and makes myriads of data available to linguists and ethnologists. It also serves as the basis for lexical *dialectometrics* (see Goebel 2010).

An important desideratum for the future is the combination of the existing dialect dictionaries for a given language (and linking them to digital standard language dictionaries). Moulin (2010, 601ff.) describes the University of Trier's *Digitaler Verbund von Dialektwörterbüchern DWV* [Digital Association of Dialect Dictionaries] (<http://www.dvw.uni-trier.de>), where the existing wide-area dictionaries for the German dialects are being integrated in a large dictionary network. At the University of Ghent a *Woordenbank van de Nederlandse Dialecten (WND)* is under construction, in which existing old and modern "amateur" dialect dictionaries are being digitized and converted to the same format (see Van Keymeulen and De Tier 2013).¹⁷ With regard to what might be possible, Moulin (2010, 608) rightfully remarks, "The fundamental preconditions for such complex information systems are the quality and interoperability of the data, ensured by the development and implementation of international standards, plus a willingness to cooperate within the academic community."

Fieldwork nowadays may be done by crowdsourcing, aimed at the co-creation by lexicographer and native speakers of a "living" digital wiki-dictionary. Dictionary projects of "substandard" language varieties are becoming very popular on the internet. This kind of citizen science is very valuable, provided it is monitored well by a skilled lexicographer.

NOTES

1 See also Friebertshauser (1986) and Moulin (2010).

2 The charter explicitly excludes dialects from recognition as regional languages; see Part I (General provisions), Article 1: "regional or minority languages" means languages that are: (i) traditionally used within a given territory of a State by nationals of that State who form a group numerically smaller than the rest of the State's population; and (ii) different from the official language(s) of that State; it does not include either dialects of the official language(s) of the State or the languages of migrants."

3 Dialectology, sociolinguistics, human geography, sociology, and anthropology indeed are being welded together into a highly integrated new science, hence the need for the recent handbook on *Language and Space* (see Auer and Schmidt (2010, vii), who say "The theoretical and methodological reorientation of research into the interplay of language and space is in full swing").

4 Regional labels are already present in Kiliaan's *Dictionarium Teutonico-Latinum* (1588). For Kiliaan, the Brabantic dialect was the "standard" for Dutch. See also Claes (1979, 1991).

5 The definitions are inspired by Hartmann and James (1998).

6 The *Schlesisches Wörterbuch* (SW) of Mitzka (1914) is a good example of the combination of an alphabetical dictionary and word maps.

7 Flemish is—contrary to international belief—not a language but a colloquial term for Belgian Dutch, a "natiolect" (i.e., a national standard) of the Dutch language, in the same way as Austrian German is a natiolect of German. In traditional dialectology, the term *Flemish* has a much more restricted sense (a southwestern Dutch dialect group spoken in French Flanders in France, the two western provinces of Dutch-speaking Belgium, and Zealand Flanders in the Netherlands). The information for *Flemish* given on the Ethnologue website as of July 2014 is regrettably inaccurate.

8 For onomasiological lexicography in general, see Reichmann (1990), Hartmann (2005) and Vossen (in press).

9 The Occitan author Frédéric Mistral received the Nobel Prize for Literature in 1904, together with the Spanish playwright José Echegaray.

- 10 The Flemish author Stijn Streuvels sometimes included invented dialect words in his regional novels, which indeed have found their way into dictionaries as “phantonyms” (i.e., non-existent words).
- 11 The boundary between lexicographical and anthropological fieldwork is sometimes very thin (see, among others, Thieberger 2011; Chelliah and de Reuse 2011).
- 12 See Heikens and Van der Schaaf (1970) and Taeldeman (1970).
- 13 Many dialect speakers are heavily influenced by the prevailing standard language ideology and consider their dialect to be “bad language,” which should not be used in the presence of strangers.
- 14 In the classical definition formula, the superordinate word + the characteristic features which distinguish the target word from the generic term that it is considered a specific instance of, for example, *fir*: a tree (= *genus proximum*) with evergreen needles (= *differentia specifica*) (after Hartmann and James 1998).
- 15 To my knowledge, the first major dialect dictionary which after digitization was supplemented by sound samples for all example sentences was the *Woordenboek der Zeeuwse Dialecten* (WZD, *Dictionary of the Zealand Dialects*). The CD-ROM for WZD was issued in 1999.
- 16 For an early experiment (in 2000) with a dialect lexicographic hypertext, combining words, database, map, sound, etymology, and encyclopedic information see <<http://users.ugent.be/~jvkeymeu/cyberlemmata/>>—SPIN “spider.”
- 17 In 2014, a similar project started at the Meertens Instituut in Amsterdam.

REFERENCES

- Ammon, Ulrich, Hans Bickel, and Jakob Ebner, eds. 2004. *Variantenwörterbuch des Deutschen: Die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol*. Berlin: de Gruyter.
- Auer, Peter, and Jürgen Schmidt, eds. 2010. *Language and Space: An International Handbook of Linguistic Variation, Vol. 1: Theories and Methods* (Handbücher zur Sprach- und Kommunikationswissenschaft (HSK), Vol. 30.1.). Berlin: de Gruyter.
- Bergenholtz, Henning, and Joachim Mugdan. 1990. “Formen und Probleme der Datenerhebung II: Gegenwartbezogene synchronische Wörterbücher.” In *Handbücher zur Sprach- und Kommunikationswissenschaft (HSK)*, Vol. 5.3: *Wörterbücher. Dictionaries. Dictionnaires. Ein Internationales Handbuch zur Lexicographie. International Encyclopedia of Lexicography. Encyclopédie internationale de la lexicographie*, edited by Franz Hausmann, Oskar Reichmann, Herbert Wiegand, and Ladislav Zgusta, 1611–1625. Berlin: de Gruyter.
- Bernal, Elisenda, and Judit Aymerich. In press. “Dictionaries and national identity.” In *International Handbook of Modern Lexis and Lexicography*, edited by Patrick Hanks, and Gilles-Maurice de Schryver. Berlin: de Gruyter.
- Cassidy, Frederic, and Joan Hall. 1985–2013. *Dictionary of American Regional English* (6 vols.). Cambridge, MA: Belknap Press of Harvard University Press.
- Chambers, Jack, and Peter Trudgill. 1998. *Dialectology*. Cambridge: Cambridge University Press.
- Chelliah, Shobhana, and Willem de Reuse. 2011. *Handbook of Descriptive Linguistic Fieldwork*. Dordrecht: Springer.
- Claes, Frans. 1979. “Dialectwoorden bij Kiliaan.” In *Handelingen van de Koninklijke Zuid-Nederlandse Maatschappij voor Taal- en Letterkunde en Geschiedenis*, 33: 35–52.
- Claes, Frans. 1991. “Dialectlexicografie bij Kiliaan.” In *Taal en Tongval, Themanummer 4: (Dialectlexicografie)*: 155–163.
- Clyne, Michael, ed. 1992. *Pluricentric Languages: Differing Norms in Different Nations*. Berlin: de Gruyter.
- De Bo, Leonard. 1873. *Westvlaams Idioticon*. Bruges: Gailliard. [2nd enlarged edition by Joseph Samyn. 1892. Gent: Siffer].
- den Boon, Ton, and Dirk Geeraerts. 2005. *Van Dale Groot Woordenboek van de Nederlandse Taal*. Utrecht: Van Dale Uitgevers.
- Desgrouais, Mr. 1766. *Les Gasconismes Corrigés: Ouvrage Utile à Toutes les Personnes qui Veulent Parler et Ecrire Correctement et Principalement aux Jeunes Gens Dont*

- l'Education n'est Point Encore Formée.* Toulouse: Robert.
- DSAE = Penny Silva, P., Dore, W., Mantzel, D., Muller, C. and Wright, M. (eds.). 1996. *A Dictionary of South African English on Historical Principles*. Oxford University Press.
- Frieberthäuser, Hans, ed. 1986. *Lexikographie der Dialekte: Beiträge zur Geschichte, Theorie und Praxis*. Tübingen: Reihe Germanistische Linguistik 59.
- Goebl, Hans. 2010. "Dialectometry and quantitative mapping." In *Language and Space: An International Handbook of Linguistic Variation, Vol. 2: Language Mapping, Part 1*, edited by Alfred Lameli, Roland Kehrein, and Stefan Rabanus, 433–57 (text), 2201–12 (maps). Berlin: de Gruyter.
- Görlach, Manfred. 1990. "The Dictionary of Transplanted Varieties of Languages: English." In *Handbücher zur Sprach- und Kommunikationswissenschaft* (HSK), Vol. 5.3: *Wörterbücher. Dictionaries. Dictionnaires. Ein Internationales Handbuch zur Lexicographie. International Encyclopedia of Lexicography. Encyclopédie internationale de lexicographie*, edited by Franz Hausmann, Oskar Reichmann, Herbert Wiegand, and Ladislav Zgusta, 1475–1499. Berlin: de Gruyter.
- Green, Jonathon. In press. "Slang lexicography." In *International Handbook of Modern Lexis and Lexicography*, edited by Patrick Hanks, and Gilles-Maurice de Schryver. Berlin: de Gruyter.
- Hallig, Rudolf, and Walther von Wartburg. 1952 [1963]. *Begriffssystem als Grundlage für die Lexikographie: Versuch eines Ordnungsschemas*. Berlin: Deutsche Akademie der Wissenschaften.
- Hartmann, Reinhard. 2005. "Onomasiological dictionaries in 20th-Century Europe." In *Lexicographica: International Annual for Lexicography*, 21: 6–19.
- Hartmann, Reinhard, and Gregory James. 1998. *Dictionary of Lexicography*. New York: Routledge.
- Haugen, Einar. 1966. "Dialect, language, nation." In *American Anthropologist*, 68: 922–935.
- Hausmann, Franz. 1990. "Les dictionnaires du français hors de France." In *Handbücher zur Sprach- und Kommunikationswissenschaft* (HSK), Vol. 5.3: *Wörterbücher. Dictionaries. Dictionnaires. Ein Internationales Handbuch zur Lexicographie. International Encyclopedia of Lexicography. Encyclopédie internationale de lexicographie*, edited by Franz Hausmann, Oskar Reichmann, Herbert Wiegand, and Ladislav Zgusta, 1500–1505. Berlin: de Gruyter.
- Heikens, Henk, and Reimer van der Schaaf. 1970. "Dialecten op de band." In *Taal en Tongval*, 22: 29–58.
- Hofman, Martha. 2009 [2014]. *Prisma Handwoordenboek Nederlands*. Houten: Uitgeverij Unieboek/Het Spectrum.
- Junius, Hadrianus. 1567. *Nomenclator, Omnium Rerum Propria Nomina Variis Linguis Explicata Indicans*. Antwerp: Christopher Plantin.
- Kiliaan, Cornelius. 1599. *Etymologicum Teutonicae Linguae sive Dictionarium Teutonico-Latinum*. [facsimile reprint by Familia et Patria, Handzame 1974].
- Kloss, Heinz. 1978. *Die Entwicklung Neuer Germanischer Kultursprachen seit 1800*. Düsseldorf: Schwann.
- Kortmann, Bernd, and Susanne Wagner. 2005. "The Freiburg English Dialect Project and Corpus (FRED)." In *A Comparative Grammar of British English Dialects: Agreement, Gender, Relative Clauses*, edited by Bernd Kortmann, Tanja Hermann, Lukas Pietsch, and Susanne Wagner, 1–20. Berlin: de Gruyter.
- Labov, William. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.
- Laureys, Godelieve. 1997. "Den svenska-nederländska/nederländsk-svenska ordboken. En bidirectionell ordbok i ett dubbelt binationellt perspektiv." In *Nordiska Studier i Lexikografi*, 4: 249–255.
- Leibnitz, Gottfried. 1697. *Unvorgreifliche Gedanken, Betreffend die Ausübung und Verbesserung der Deutschen Sprache: Zwei Aufsätze*, edited by Uwe Pörksen [1983], 5–46. Stuttgart: Reclam.
- Moulin, Claudine. 2010. "Dialect dictionaries – traditional and modern." In *Language and Space: An International Handbook of Linguistic Variation, Vol. 1: Theories and Methods* (Handbücher zur Sprach- und Kommunikationswissenschaft (HSK), Vol. 30.1.), edited by Peter Auer, and Jürgen Schmidt, 592–612. Berlin: de Gruyter.
- Niebaum, Hermann. 1989–1991. "Diatopische Markierungen im allgemeinen einsprachigen Wörterbuch." In *Handbücher zur Sprach- und Kommunikationswissenschaft* (HSK), Vol. 5.1: *Wörterbücher. Dictionaries. Dictionnaires. Ein Internationales Handbuch zur Lexicographie. International Encyclopedia of Lexicography. Encyclopédie internationale de lexicographie*, edited by Franz Hausmann, Oskar Reichmann, Herbert Wiegand, and Ladislav Zgusta, 662–668. Berlin: de Gruyter.

- Pletinckx, Lode. 2003. *Woordenboek van het Asses: Bijdrage tot de Studie van de West-Brabantse Streektaal*. Asse: Koninklijke Heemkring Ascania.
- Reichmann, Oskar. 1990. "Das onomasiologische Wörterbuch: Ein Überblick." In *Handbücher zur Sprach- und Kommunikationswissenschaft (HSK)*, Vol. 5.2: *Wörterbücher. Dictionaries. Dictionnaires. Ein Internationales Handbuch zur Lexicographie. International Encyclopedia of Lexicography. Encyclopédie internationale de lexicographie*, edited by Franz Hausmann, Oskar Reichmann, Herbert Wiegand, and Ladislav Zgusta, 1057–1067. Berlin: de Gruyter.
- Rézeau, Pierre. 1990. "Le dictionnaire dialectal: L'exemple français." In *Handbücher zur Sprach- und Kommunikationswissenschaft (HSK)*, Vol. 5.2: *Wörterbücher. Dictionaries. Dictionnaires. Ein Internationales Handbuch zur Lexicographie. International Encyclopedia of Lexicography. Encyclopédie internationale de lexicographie*, edited by Franz Hausmann, Oskar Reichmann, Herbert Wiegand, and Ladislav Zgusta, 1467–1475. Berlin: de Gruyter.
- Seiler, Guido. 2010. "Investigating language in space: Questionnaire and interview." In *Language and Space: An International Handbook of Linguistic Variation*, Vol. 1: *Theories and Methods* (Handbücher zur Sprach- und Kommunikationswissenschaft (HSK), Vol. 30.1.), edited by Peter Auer, and Jürgen Schmidt, 512–527. Berlin: de Gruyter.
- SW = Mitzka, Walther. 1914. *Schlesisches Wörterbuch*. Berlin: de Gruyter.
- Taeldeman, Johan. 1970. *Zuidnederlandse Dialecten op de Band*. Gentse Bijdragen 15.
- Thieberger, Nicholas. 2011. *The Oxford Handbook of Linguistic Fieldwork*. Oxford: Oxford University Press.
- Van Dale = den Boon, Ton and Geeraerts, Dirk. 2005. *Van Dale Groot woordenboek van de Nederlandse Taal*. Van Dale Uitgevers.
- Van Keymeulen, Jacques. 1986. "Het Woordenboek van de Vlaamse Dialekten: Kognitieve antropologie in de praktijk?" In *Vruchten van z'n Akker: Hulde-Album Prof. Dr. V.F. Vanacker*, edited by Magda Devos, and Johan Taeldeman, 445–66. Gent: Seminarie voor Nederlandse Taalkunde en Vlaamse Dialectologie.
- Van Keymeulen, Jacques. 2003. "Compiling a dictionary of an unwritten language: A non-corpus-based approach." In *Lexikos*, 13: 183–205.
- Van Keymeulen, Jacques, and Veronique De Tier. 2013. "De woordenbank van de Nederlandse dialecten". In *Electronic Lexicography in the 21st Century: Thinking Outside the Paper (Proceedings of the eLex 2013 conference, 17–19 October 2013, Tallinn, Estonia)*, edited by Iztok Kosem, Jelena Kallas, Polona Gantar, Simon Krek, Margit Langemets, and Maria Tuulik, 261–79. Ljubljana: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut.
- Van Sterkenbrug, Piet. 2003. "Onomasiological specifications and a concise history of onomasiological dictionaries." In *A Practical Guide to Lexicography*, edited by Piet Van Sterkenbrug, 127–143. Amsterdam: Benjamins.
- Van Sterkenbrug, Piet, ed. 2003. *A Practical Guide to Lexicography*. Amsterdam: Benjamins.
- Vossen, Piek. In press. "Onomasiological lexicography." In *International Handbook of Modern Lexis and Lexicography*, edited by Patrick Hanks, and Gilles-Maurice de Schryver. Berlin: de Gruyter.
- Weijnen, Antonius, and Jan Van Bakel. 1967. *Voorlopige Inleiding op het Woordenboek van de Brabantse Dialecten*. Assen: Van Gorcum.
- Wiegand, Herbert. 1977. "Nachdenken über Wörterbücher: Aktuelle Probleme." In *Nachdenken über Wörterbücher*, edited by Günther Drosdowski, Helmut Henne, and Herbert Wiegand, 51–102. Mannheim: Bibliographisches Institut – Dudenverlag.
- WBD = Weijnen, Antonius, et al. 1967–2005. *Woordenboek van de Brabantse Dialecten*. Assen: Van Gorcum.
- WLD = Weijnen, Antonius, Jan Goossens, et al. 1983–2008. *Woordenboek van de Limburgse Dialecten*. Assen: Van Gorcum.
- WVD = Devos, M., et al. 1979–. *Woordenboek van de Vlaamse Dialecten*. Gent: Michiels.
- WZD = Ghijssen, Hendrika. 1964. *Woordenboek der Zeeuwse Dialecten*. The Hague: Van Goor. [Fraanje, Kees, et al. 2003. *Supplement Woordenboek der Zeeuwse Dialecten*. Krabbendijke: Van Velzen].

3 Linguistic Atlases

WILLIAM A. KRETZSCHMAR, JR.

3.1 Introduction

The term *linguistic atlas* means different things to different people. It must have something to do with language, and something to do with maps, but there the agreement ends. The history of spatial analysis of language information begins with the Neogrammarians, but the display of language forms on maps immediately became complicated in theoretical terms. Since the late nineteenth century, a great many innovations have altered our expectations about what language features one might expect to find on atlas maps, and of course the creation of maps has been revolutionized by computer methods in what are generally called Geographic Information Systems (GIS). Theory and practice are inextricably intertwined in all of these developments.

3.2 Terminology and Method

Anyone who googles the term “linguistic atlas” will find the essentially the same definition in a number of lexical resources. The *American Heritage Dictionary* says that a linguistic atlas is “A set of maps recording the geographic distribution of variations in speech. Also called dialect atlas.” This definition buries what “variations in speech” might be, and raises the troublesome issue that such variation occurs in the form of a “dialect.” While the word “linguistic” only suggests some general quality of language and applies it to geography by means of a collection of maps, the term “dialect” assumes that there are real entities called dialects, which should be located in the map collection. “Word geography,” a different relevant term, indicates the study of the areal distribution of particular words without any claim about the reality of dialects, and identifies the locus of variation in speech.

Lee Pederson has offered perhaps the best description of the method for the creation of a linguistic atlas as a product of scientific study in language variation (1995, 36–37):

... a logically ordered and systematic approach that begins with common sense, proceeds through deductive cycles, and concludes in enumeration. It conducts research in a geographic context, but its research concerns a few words of a language, not the language itself and its universe of discourse. [...] American dialectologists, for example, concentrate on sorting and counting components – American English synonyms, morphs, and phones. They are not concerned with the identification of new linguistic classes, semantic,

grammatical, and phonological sets established according to the scientific method. [...] In word geography, [deduction] concerns the engagement of target forms. It takes them first as contrastive lexical sets, and then carries the work forward through segmentation of self-evident morphemes, phonemes, allophones, and distinctive features,¹ according to the needs of a descriptive problem. Taken this way, deductive word geography studies only classes and components of phonological words as they characterize speakers classified according to geographical place and analyzed according to social factors.

Pederson goes on to make an explicit list of the steps in the process of making a linguistic atlas, saying that dialectologists

1. begin with an idea ('Let us compose a linguistic atlas' of a given territory)
2. establish a grid across that territory
3. establish a network of communities within that grid
4. organize a questionnaire of selected items (lexical, grammatical, and phonological issues)²
5. select informants within that network of communities
6. interview informants according to the questionnaire
7. record informant responses (lexical, grammatical, and phonological components) in a finely-graded phonetic notation (Pederson 1995, 37-38).

Pederson's description therefore makes the method "thoroughly deductive, from general to specific, from class to component" (1995, 38). We need to modify his description for linguistic atlases outside the French, Italian, and American traditions only in the last two steps, as some dialectologists asked informants to return written responses to the questionnaire (as did, for example, Georg Wenker in Germany), or, more recently, recorded responses using acoustic phonetic means and not via manual phonetic transcription.

Pederson, however, warns us that the result of the process is open to misinterpretation (1995, 39):

Such analytic word geography ends its work at this point in a taxonomy of observed socio-linguistic facts. But the research invariably implies more than that because planners, editors, and their critics fail to characterize the work at hand. For that reason, a reader expects an identification of dialect areas and a description of dialects within those geographic divisions in a concordance of social and linguistic facts projected across space and through time.

The expectation that a linguistic atlas will reveal dialects, then, exceeds what the method and the process will reveal. A linguistic atlas is a scientific study of language variation that catalogues different word forms representing variations in lexicon, grammar, and pronunciation, especially as these differences can be plotted on maps. The interpretation of these findings with respect to "dialects" is a separate step, one that is generally not a part of a linguistic atlas.

The reason that planners, editors, critics, and readers all expect to find dialects has become clear in recent years, as human speech has been shown to constitute a complex adaptive system. Kretzschmar (2009) demonstrated what linguistic atlas findings really show when one does not jump to the conclusion that dialects will be found there, whereas Kretzschmar (2015) extends the same principles to many areas of linguistics. All of the results from surveys of the lexicon, grammar, and pronunciation in the linguistic atlas tradition show nonlinear frequency profiles for every question, and these nonlinear profiles are arranged in a "scale-free" network. So, for example, Figure 3.1 shows the nonlinear frequency profile for responses to the thunderstorm item in the *Linguistic Atlas of the Middle and South Atlantic States* (LAMSAS).

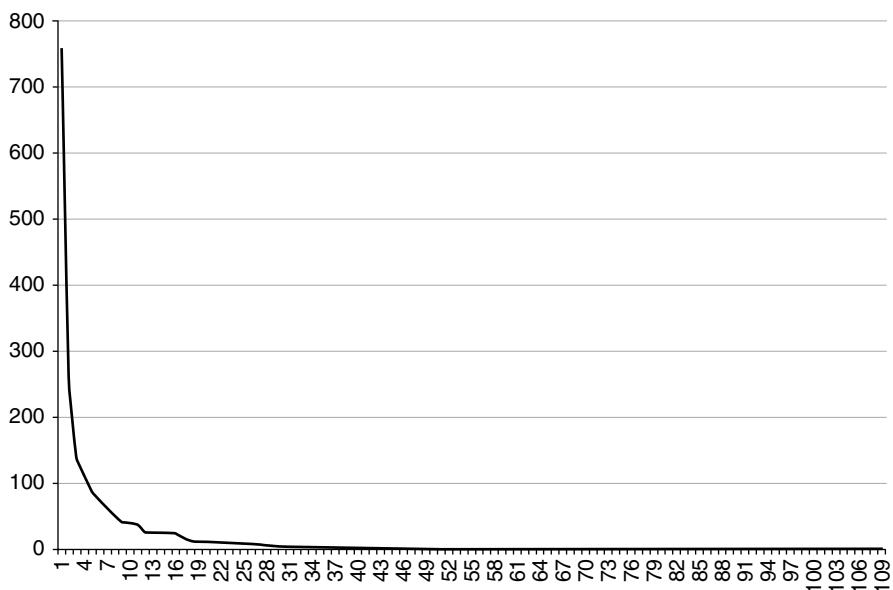


Figure 3.1 LAMSAS responses for the *thunderstorm* item. Word frequency (counts) is shown on the y axis, whereas the x axis represents the number of individual word forms ($N = 109$).

A total of 109 different names for a thunderstorm were reported by the 1,162 people in the survey. The top five names were *thunderstorm* (758 tokens), *thundershower* (247), *storm* (136), *thundercloud* (110), and *electric storm* (87). Most of the names only occur once or twice in the data, but a few are very common, yielding what many would call a fractal distributional pattern with a peak and a long tail. We observe the same nonlinear distribution in every subset of the data as well as in the overall set, which follows the scale-free property of complex systems.

These profiles arise out of the interaction of all people using language. “Nonlinear” simply means that a few ways to say or write something occur very commonly, and a great many ways to say or write the same thing occur only rarely. So, in America, most people say and write *you* for the second-person plural pronoun, but there are many variant usages that occur far less often, like *you guys*, *y'all*, *youse*, *you'uns*, and many others. The “scale-free network” just means that we can expect a nonlinear frequency profile in whatever subset of the entire survey we look at, whether a regional subset or a social subset or a situational subset, and that the very common and uncommon variants may be in a different order in different subsets. For instance, in the American South, *y'all* is a very common form for the second-person plural, with *you* being less common; in Pittsburgh, *you'uns* is very common in many neighborhoods, whereas *you* is still the most common in others; and so it goes on for every subset of speakers we care to inspect. The process operating in the complex system of speech just explains better what we already knew: we tend to talk like the people nearby us, either physically near or socially near, or both, and we tend to use the same linguistic tools that others do when we are writing or saying the same kind of thing. This scaling property helps us to see that we choose to grant the status of a dialect to the speech of groups that we care about for reasons other than speech. Just as languages differ from dialects because, as per Max Weinreich's well-known aphorism, their speakers have armies and navies, complex systems tells us that American Southerners and Pittsburghers have their own dialects not because their language is special in some way, but because the South is special and Pittsburgh

is special in our perception of America, and we then look for linguistic differences to match our perception. Southerners and Pittsburghers remain members of the larger-scale American dialect, sharing most of the variant features in the nonlinear distribution even as they have some of their own high-frequency variants.

Readers (and planners, editors, and critics), therefore, expect a linguistic atlas to provide the evidence for dialects that they already have believed to exist, based on the perceptual advantage that sharp frequency differences provide. As Dennis Preston and others have shown (see Kretzschmar 2009, Ch. 7), readers do not agree about the populations who might speak a dialect, and there is only low-level agreement about where commonly named dialects might exist. That is, in the US and the UK, we share common names for dialect regions like “Northern” or “Southern” or “Midlands” (regions of very long standing in the UK), but we agree only approximately where the borders between them might be. The complex system of speech offers an effectively infinite number of possible groups of speakers who might speak a dialect, and suggests that every speaker will use more than one variety, so there is no way to identify some “correct” set of dialect areas. When people, in Pederson’s words, “fail to characterize the work at hand” (1995, 39), we now understand this observation to mean that people’s perceptions that some correct set of dialect areas *should* exist is not matched by scale-free speech production. The operation of the complex system of speech confirms Pederson’s view that, in practical terms, a linguistic atlas should present word-level survey data accompanied by geographical and social information that allows its readers to consider groups of speakers of interest to them.

3.3 History of Linguistic Atlases

A great many linguistic atlases have been produced in the last century. Joachim Grzega (2009) lists no fewer than 181 projects: 35 German, 28 French, 23 English, and 95 in thirty-three other languages. The linguistic atlas is certainly a popular form for organizing information about language variation.

The first attempt to create a linguistic atlas is generally attributed to Georg Wenker. Dollinger (2015, 13–16) reports two earlier German contenders, Stadler and especially Schmeller, but agrees that Wenker’s work toward his dissertation in the 1870s was the first to use a definite questionnaire (42 sentences to be translated from Standard German into the local variety of German) across a region. Wenker expanded his region from his original focus on Düsseldorf to all of Westphalia, and eventually his method was applied to the entire German-speaking area of Europe. The complete *Deutscher Sprachatlas* eventually completed its publication in 1956 (Wrede, Martin, and Mitzka 1927–1956). As Dollinger reports, Wenker wrote to school inspectors and village teachers, and asked them to find a “suitable” teacher to translate the sentences. There was no established system for these translations and no way to know about the linguistic history of the teachers who executed the translations, and yet the wealth of data elicited provided valuable information about variation in German. Work based on Wenker’s data has continued until the present time, notably in the *Digitaler Wenker-Atlas* (DiWA; www.diwa.info/titel.aspx).

Even from its early days, Wenker’s idea that his survey would reveal neat dialect regions predicted by the Neogrammarian theory of exceptionless change was not supported. Walther Mitzka, the third director of the project initiated by Wenker, described the early results of his project:

As regards the philological controversy about the exceptionlessness of sound laws, the evidence of dialect around Düsseldorf should have brought proof for the adherents of this teaching. The results supported the contrary view. [Wenker] reported in 1885 before the

Giessen conference of philologists that, 'I lived in the fair and calming conviction that these [linguistic] features must completely or nearly completely go together. That assumption turned out soon enough to be utterly mistaken: the boundaries of the contemplated features stubbornly took their own way and often crossed each other.' (Mitzka 1943, 9; my translation).³

Such problems did not stop those who preferred to follow the Neogrammarian program. As I have written elsewhere (Kretzschmar 2002), Hans Kurath, founder of the American Linguistic Atlas Project, disputed the account of Adolph Bach (1950) about Wenker's evidence. In response to Bach's statement "The Linguistic Atlas of Germany was born from the position of the exceptionlessness of sound laws. It has not confirmed this postulate of the Neogrammarians" (my translation),⁴ Kurath has simply written "Wrong," underlined, in the margin of the book in the atlas library. Kurath just says "no!" with a double underline and exclamation point, to Bach's summary statement that "The Linguistic Atlas shows in contrast that, fundamentally, each separate word and each separate word form has its own range, its own borders within the speech area" (my translation).⁵ Kurath was a defender of the Neogrammarian position on sound laws, even though he was confronted in his own atlas with data that required serious accommodation.

In contemporary linguistics, the most prominent defender of the Neogrammarian position has been William Labov (e.g., 1994). The *Atlas of North American English: Phonetics, Phonology and Sound Change* (Labov, Boberg, and Ash 2006) uses the data elicited from a nationwide panel of speakers, and including some in Canada, to defend the dialect regions or sound changes (five shifts, including the Northern Cities Shift and Southern Shift) that Labov has identified in continuing work since the 1970s. Maps typically employ plotted points and isoglosses where data were elicited (see www.atlas.mouton-content.com for examples).

Contemporary with the Neogrammarians was Gaston Paris in France, who saw language variation as a linguistic continuum: "a vast tapestry of which the varied colors flow out from locations [i.e., towns] in imperceptible gradations" (cited in Jaberg 1936, 32; my translation).⁶ Hans Goebel has written extensively (e.g., 1982, 1990, 2003) on the debate between the French and German camps (to use the national designations as a convenient shorthand for two basic approaches to variation), which he has labeled "typophobia" (for Gaston Paris and fellow dialectology advocate Paul Meyer) and "typophilia" (for the Italian linguist Graziadio Ascoli), respectively. The successor of Gaston Paris in the French school, Jules Gilliéron, spent his time with individual words and their histories (1918, 1921, 1922; Gilliéron and Mongin 1905).

The purpose of French-style variation studies is to connect them with particular geographical and social circumstances. As early as 1887, Abbé Rousselot had proposed the importance of the "social condition" of speakers included in a linguistic survey (cited in Pop 1960, 43). Two early, seminal studies considered the speech of towns in their social contexts: Gauchat's famous turn-of-the-century study of the town of Charmey, in the Suisse Romande, reported significant findings about age and gender (Gauchat 1905, cf. Pop 1960, 187–196); so did Gilliéron's less well-known study of Vionnaz, a town in the same region, which was published 25 years earlier, in 1880 (cf. Pop 1960, 178–182). Gaston Paris advocated the creation of similar monographs, modeled after Gilliéron's study, for every town in France. The goal of this initiative would be to isolate characteristic linguistic features that were not shared with neighboring places, to determine the relation between the "patois" and the standard language, and to consider circumstances that conditioned the entry of new words (Pop 1960, 48). Patois is the "local speech," the set of habits peculiar to a place, but does not have formal boundaries, even though it is identified with a geographic location. No claim is made that any patois is a systematic subvariety of a language, like the Neogrammarian view of

dialect, since the patois is characterized merely by the list of features that differentiate it from the speech of other locations and from the standard language. The atlas survey, then, helps to identify the words that might be recognized as components of a patois, and the different words that were variants of questionnaire items might also be compared in different locations.

The French method of gathering data used a fieldworker instead of the postal questionnaire. Only one fieldworker, Edmond Edmont, was used for the *Atlas Linguistique de France* (ALF; Gilliéron and Edmont 1902–1910). He reached 639 French communities by bicycle. The questionnaire Edmont used covered about 2,000 words and phrases, and informants' responses were written down in fine phonetic transcription. The advantage of employing a fieldworker over using a postal questionnaire lies in the greater consistency of data collection and the larger extent of the questionnaire, whereas the disadvantage is that far fewer locations could be covered. Control over the quality of the data has been a decisive factor, however, since the fieldworker method has been used in nearly all major projects since Gilliéron (the exception is Chambers' Golden Horseshoe project in Canada; see Chambers 1998). Phonetic transcription, most commonly using the International Phonetic Alphabet, allows for consistent representation of speech sounds, as well as word forms and grammar, and has been supplemented in recent decades by acoustic phonetic measurements of recorded interviews.

Gilliéron, in the view of his student Karl Jaberg, was possessed of "a certain mathematical spirit" that made Gilliéron concentrate on the words themselves as the basis for linguistic science, rather than as components of a patois (1936, 28). Jaberg himself believed that words are not subject to the same kind of processing, notably statistical processing, as the data of the natural sciences. He was an advocate of the *Wörter und Sachen* ("words and things") school, a branch of cultural anthropology, which preferred the ethnographic approach. The German dialogue concerning theory with the French appears clearly in the work of Jaberg. He claimed that "If we could establish a Linguistic Atlas of France from the eleventh century, we would doubtless see dialect boundaries draw themselves with a precision that would refute Gaston Paris" (my translation; 1936, 32).⁷ Jaberg looks backward to a time when there were 'real' dialects, at the same time as focussing on words and things. For Jaberg, as for Kurath, the French and German approaches coexisted.

Immediate successors of ALF included Jaberg and Jud, who created an atlas of Italian areas of Switzerland (Jaberg and Jud 1928–1940). The French approach has come to characterize the contemporary strand of American dialectology as expressed above by Pederson, whose emphasis on the linguistic continuum matches the findings of complex systems research better than the common perception of dialects. The American Linguistic Atlas effort, promoted and begun in 1929 by Hans Kurath, followed, with modifications, the lead of the French method in adopting a long questionnaire (about 800 questions) designed to sample everyday speech. For each large region covered (a pilot project in New England, and then other wide areas chosen for convenience, not on the basis of suspicions about dialects), communities were chosen with regard to culture, settlement, and demographics so as to include historically important places and cultural groups within an even spread of area and population. Within each of the communities, two speakers were normally selected as representative of the community because of life-long residence there: one was a member of the oldest living generation, with little education or compensating experience, and the other was younger and better educated, with a less insular outlook. In 20% of the communities a member of a local elite—people with the best education and best access to high culture—was interviewed. Questioning styles avoided "how do you say..." questions in favor of less-direct approaches, in order to reduce the formality of the interview situation (see Pederson et al. 1972). Interviewers noted down responses in Kurath's modification of IPA phonetics (see Kurath 1939, 1939–1943), indicated any special circumstances in which the responses

were given, captured informants' comments, and made a detailed biographical sketch for each informant. Kurath's interview plan for the American atlas effort thus replicated the French style of elicitation, whereas Kurath's own analyses tended toward the Neogrammarian interest in finding dialects. Kurath's analytical method worked from the distribution of single variants to the creation of regional boundaries (e.g., Kurath 1949, Kurath and McDavid 1961). The *Survey of English Dialects* (SED) began later than the American atlas, but in some ways remained more like the ALF in that fieldworkers interviewed more than one person in a community without differentiating them as individuals (Orton 1962–1971). SED analyses have included both single-item studies and delineation of dialect regions (e.g., Orton and Wright 1974; Orton, Sanderson, and Widdowson 1978; Upton, Sanderson, and Widdowson 1987).

The history of linguistic atlases thus exists in tension between the German approach that expects to find dialect areas, and the French approach that focuses on separate words. The practice of using fieldworkers has become nearly universal owing to the improved control that it permits over the data elicited. While the interest of the German atlas makers has been mainly geographical, the French method from an early date also considered social considerations, and the application of the French fieldwork style in America created the possibility that individual speakers could be considered as representatives of social categories as well as geographical locations.

3.4 Data and Maps

Linguistic atlases generate huge volumes of paper records: originally the responses to Wenker's set of questions, then field book pages from the ALF and other atlases using fieldworkers. In addition to the responses themselves, atlas makers must keep records of the speakers who responded, that is, metadata. In what follows, this chapter will focus mostly on the American linguistic atlas effort, but the principles explained apply to many atlases.

The original method to make data available from the American linguistic atlas was presentation on maps. In the *Linguistic Atlas of New England* (Kurath 1939–1943), actual transcriptions were hand-lettered on maps (see Figure 3.2 for an example). On this map, *bureau*, *dresser*, *chiffonier*, and other words are represented, in full phonetic detail, along with rivers and political boundaries, as if the words were part of the landscape. The implication is that there is some direct connection between a linguistic feature and a geographic place, between words and the land. Another early option for this type of map used shadings and some symbols instead of lettering, as in Figure 3.3, by Gilliéron (1918), which illustrates French words for "bee" (note that while the French-speaking parts of countries neighboring France are represented, neither Brittany nor Corsica are indicated on the map, as other languages were spoken in those areas). The American pioneers in the field, Hans Kurath and Raven McDavid, soon came to prefer the creation of list manuscripts to the entry of the data directly onto maps. They made it easier for Kurath and McDavid to plot symbols on maps instead of the data themselves. These lists of all of the answers from all of the informants were never published except as a few fascicles of the *Linguistic Atlas of the Middle and South Atlantic States* (LAMSAS; McDavid and O'Cain 1980). Publication in list form, however, was completed for the SED (Orton 1962–1971).

In the computer age, lists of responses and metadata have become databases. Kretzschmar and Schneider (1996, but originally prepared in the late 1980s) provides a detailed account of the transformation of the historical LAMSAS into a modern digital resource. Table 3.1 shows a part of the database that stores metadata for the LAMSAS project, including identifiers for the speaker, a range of social characteristics, and longitude/latitude for geocoding.

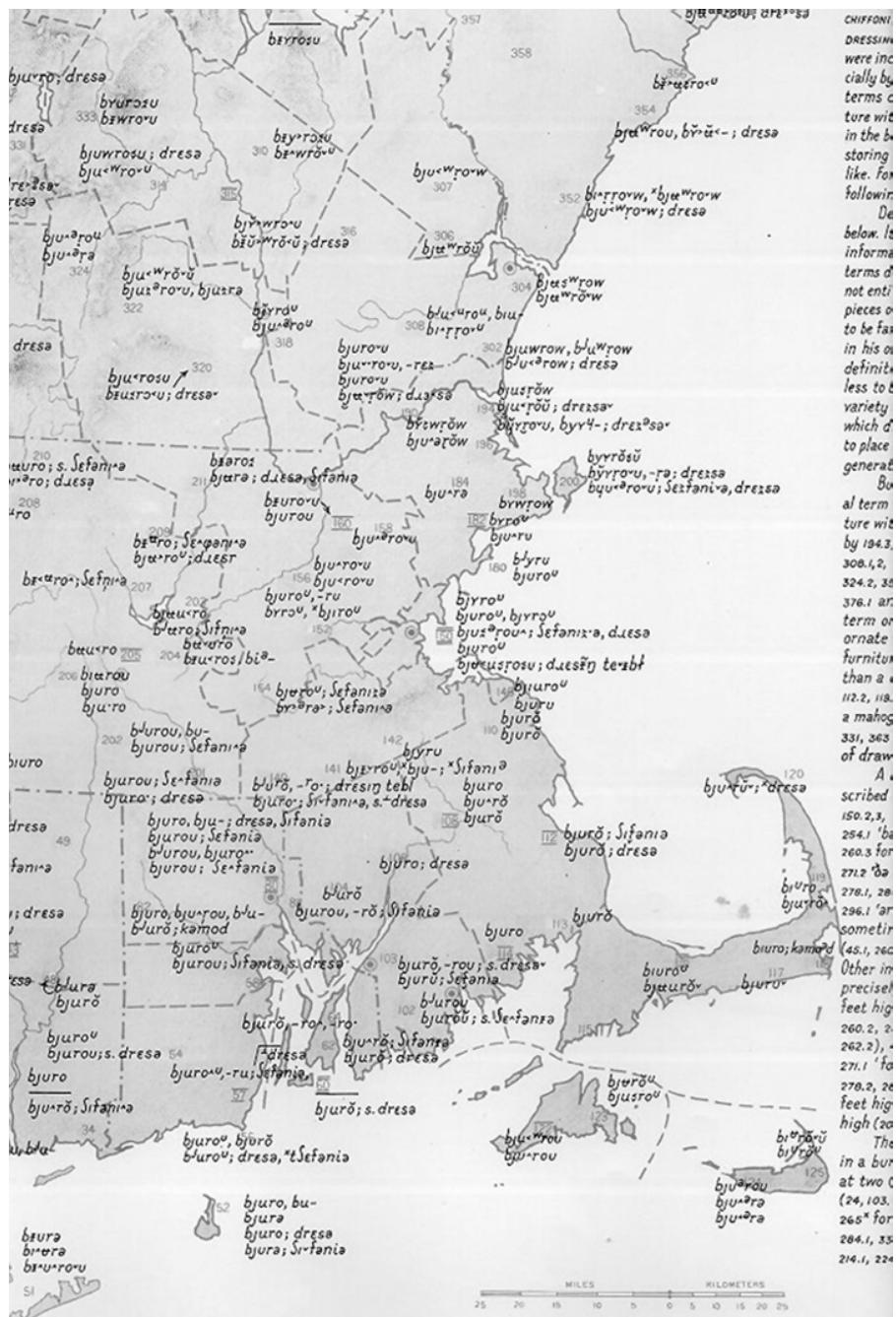


Figure 3.2 Partial map of *bureau*, *dresser*, *chiffonier* (and others) from the Linguistic Atlas of New England (Kurath 1939–1943).

Table 3.2 shows the responses of the same speakers listed in Table 3.1 to the dragonfly question. The “key” column for “informid” links the metadata to the responses in a relational database, so that users can browse and search the responses according to the social characteristics of locations in the metadata. For the American atlas, all of the information is freely

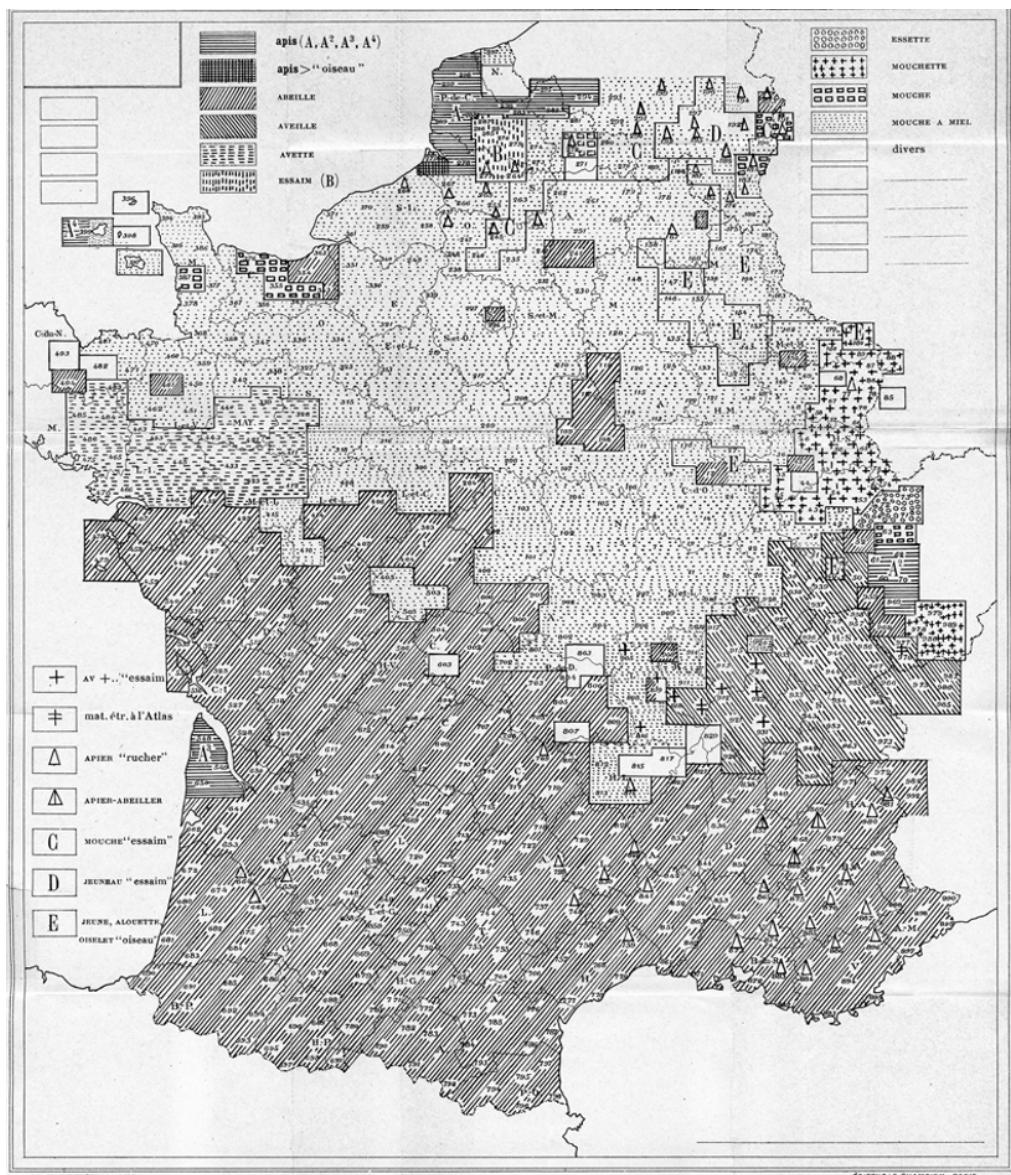


Figure 3.3 Gilliéron's (1918) map of "bee" variants in France and neighboring countries.

downloadable from the Linguistic Atlas Project (LAP) website (www.lap.uga.edu), to the extent that it has been digitized. Figure 3.4 shows a map of the *snake doctor* responses to the dragonfly question, as produced automatically on the LAP website between 1996 and 2015. The map shows a square at each of the communities where a response to this question was elicited, and “x” where no response was elicited. The squares are filled where the target response was obtained from at least one speaker, and open where only other responses were elicited. The legend indicates how many individuals (of 1,162) used the form in question across the LAMSAS survey. This style of map differs from those in Figures 3.2 and 3.3 by showing only one variant at a time, not all of the data, and it offers a comprehensive picture of the available data and the condition of each community

Table 3.1 Partial LAMSAS speaker database.

<i>Informid</i>	<i>Oldnumbe</i>	<i>Aux</i>	<i>Fw</i>	<i>Ws</i>	<i>Year</i>	<i>Inftype</i>	<i>Exp_s</i>	<i>Cult</i>	<i>Sex</i>	<i>Age</i>	<i>Educ</i>	<i>Occup</i>	<i>Race</i>	<i>Commtype</i>	<i>Community</i>	<i>Statelong</i>	<i>Lat</i>	
NY1	50	N	L	E	1933	I	B	N	M	53	3	F	W	R	Suffolk Co.	NY	72.30389	41.13889
NY2A	51.1	N	L	E	1933	I	A	N	M	63	0	F	W	R	Suffolk Co.	NY	72.185	40.965
NY2B	51.2	N	L	E	1933	I	A	N	M	85	3	F	W	R	Suffolk Co.	NY	72.27861	40.92528
NY2C	51.3	Y	L	E	1933	II	A	N	F	80	2	K	W	U	Suffolk Co.	NY	72.30139	40.93778
NY3A	401	N	L	M	1941	I	A	N	M	71	4	F	W	R	Suffolk Co.	NY	72.91583	40.77917
NY3B	403	N	L	M	1941	I	A	N	M	70	3	F	W	R	Suffolk Co.	NY	73.42611	40.86806
NY3C	402	N	L	M	1941	I	B	N	M	50	4	F	W	R	Suffolk Co.	NY	73.00195	40.86861
NY3D	404	N	L	M	1941	I	B	N	M	48	3	F	W	R	Suffolk Co.	NY	73.30139	40.9125

Table 3.2 Partial database of dragonfly responses for LAMSAS speakers.

<i>Informid</i>	<i>Item</i>	<i>Doubtflag</i>
NY1	snake doctor	N
NY2A	snake doctor	N
NY2B	snake doctor	N
NY2C	snake doctor	N
NY3A	darning needle	N
NY3A	darning needle	N
NY3B	darning needle	N
NY3B	dragonfly	N
NY3B	dragonfly	N
NY3B	darning needle	N
NY3C	darning needle	N
NY3C	darning needle	N
NY3D	darning needle	N
NY3D	darning needle	N

with respect to the survey. It is easier to read than the earlier maps, and also easier to make since the map can be prepared online in just a moment of real time, whereas earlier maps of all the data required many hours to create. Because all of the data are available for download from the LAP website, many other mapping styles can be tried by individual users on their own computers (see, e.g., Kretzschmar 2013a).

Graphical computer displays all depend on the idea of overlays: that is, of superimposing different layers of graphical information in order to produce a composite picture. This is true, for example, of the standard PC Windows display, which superimposes user-selectable icons on a user-selectable background, and then opens additional windows on top of the main window. GIS use the layer principle to establish a base map and then to add user-selectable layers, each of which contains some particular information (see Kretzschmar 2013b). Figure 3.4 is actually composed of several layers: the base map, the locations of communities, the symbols plotted at the locations, and the fill color assigned to the relevant squares. Only the last two layers change with each map created online. The symbol and color layers access information from the response databases, and link it to the location information (“geocoding”) for each community. GIS can of course be used for many purposes, not just in linguistic atlases.

GIS also provide tools for what is known as “technical geography,” which refers to the use of spatial statistics. The use of statistics in linguistic atlases has greatly accelerated in recent decades, after its beginnings in the “mathematical spirit” of Gilliéron. Work with linguistic atlas data has generated a number of new statistical procedures for language variation studies, among them spatial autocorrelation and density estimation (see Kretzschmar 1996). The use of lists and databases for storage and presentation of data enabled much easier counting of word forms and other response types, and the use of various statistics has capitalized on the property of differential frequency of different linguistic variants. Most recently, linguistic atlas data has made it possible to understand speech as a complex system (Kretzschmar 2009, 2015), which shows that the underlying pattern of such differential frequencies is “fractal,” nonlinear and scale-free (i.e., self-similar at different levels of analysis). Linguistic atlases began and remain in the forefront of what is now known as Big Data, where quantitative methods are required to make sense of the variation.

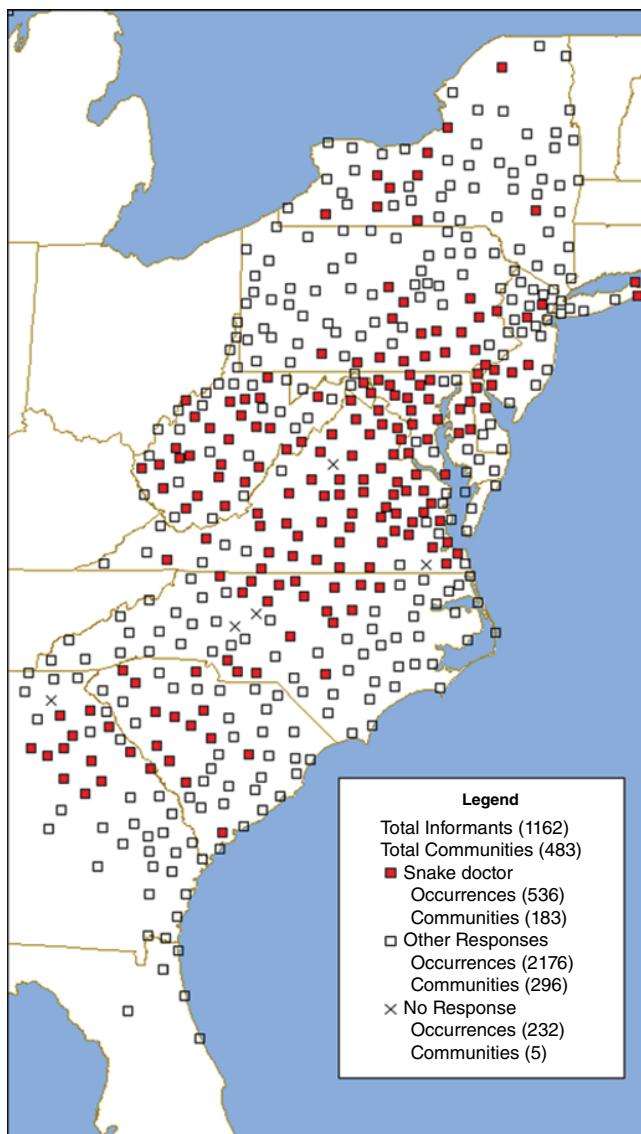


Figure 3.4 Map of *snake doctor* responses from LAMSAS, LAP website.

Besides lists of responses and metadata, linguistic atlases now also provide audio and image data. Field interviews began to be recorded in the 1960s, when good portable tape recorders began to become available. The use of recording has allowed separation of the role of conducting the interview and transcription of its content, so that now many fieldworkers can be used in a project with central, controlled transcription to maintain consistency of the data. Beginning with the *Linguistic Atlas of the Gulf States* (LAGS; Pederson 1986–1992), the tape recording became the primary record of the interview, and recordings were archived instead of the earlier practice of reuse of audio tapes. Many of these audio recordings have now been digitized (the LAGS recordings account for 5,500 hours of the 8,000 hours of interviews on the LAP website), and many newer recordings made for linguistic atlases are born digital using new digital recorders. Because it has proven to be uneconomical to keyboard

the huge volume of paper records from linguistic atlases, the field book pages and editorial materials can now be scanned for permanent archives and for putting online for users to see. It is possible, going forward, that these digital images can be transcribed (for audio recordings) and keyboarded (for images) by means of crowdsourcing. Linguistic atlases have played an important role in the public corpus movement, which aims to put linguistic data of many kinds online to make it more available to users (see Kretzschmar *et al.* 2006).

3.5 Linguistic Atlases and Linguistic Theory

The idea for linguistic atlases was born in the nineteenth century among the Neogrammarians, at the very beginning of modern linguistics as a field of study. Linguistic atlases have not remained the same over time, as this chapter has shown in some detail, and yet they are all still focused on the collection of information from many speakers over a wide geographical area. The most influential linguist at the turn of the last century, Ferdinand de Saussure, preferred to found the science of linguistics on linguistic structure (Saussure 1916). He emphasized the difficulty of working with real speech as a justification for a more abstract view of *langue*, which became an object, a social fact, in the same way that at the same period Émile Durkheim talked about human society in the founding of sociology (Durkheim 1895). Saussure, however, did discuss linguistic atlases as an alternative to his linguistics of language structure, acknowledging a different possibility for thinking about language.

Before World War II, linguistic atlases were popular, and at the same time American structuralism as represented by the work of, for instance, Leonard Bloomfield, grew to describe many of the central ideas of modern linguistics (Bloomfield 1933). After the war, however, structuralism became difficult as linguists consulted more than just one speaker in order to describe the structure of a language. In the 1950s, Noam Chomsky famously swept away the need for surveys of speakers in favor of a highly rational approach to the description of linguistic structure. As a result, linguistic atlases became less and less popular over time, and especially came in for criticism by many sociolinguists who preferred the new generative model (but, interestingly, not by Labov himself, who followed the generative model but also defended the utility of American linguistic atlases). The theoretical achievement of linguistic atlas surveys in America came through Kurath's *Word Geography* (1949) and *Pronunciation of English in the Atlantic States* (PEAS; Kurath and McDavid 1961), which were followed considerably later by Kurath's theoretical treatments of urban speech and areal linguistics (1970, 1972). The SED (Orton *et al.* 1962–1971) was a more conservative survey that did not take advantage of Kurath's innovations. By the 1970s it was too late for Kurath to confront sociolinguistics with his more traditional Neogrammarian viewpoint. Raven McDavid, Kurath's successor as editor-in-chief of the American atlas effort, had done prodigious amounts of fieldwork and had contributed to linguistics and language variation studies in numerous ways, but under his direction the American atlases continued to lose favor in the face of Chomsky and Labov. His main effort, and a successful one, was to secure the evidence of the American atlases, whether by archival arrangements for original field materials or by publication in microform. Lee Pederson contributed to the advancement of linguistic atlas methods in LAGS, published important theoretical statements (as noted above), and began the conversion of the linguistic atlas to computers, but he, too, was unable to make much impression on sociolinguistics or linguistics more generally. The same can be said for the SED in Britain, which Clive Upton continued to nurture but which fell out of the mainstream of linguistics. The ALF continued in France, and work continued on the German Sprachatlas in Marburg. These atlases were also bypassed by the mainstream. The main point at the end of the twentieth century was that a few scholars had resisted the generative impulse sufficiently to preserve the evidence collected by linguistic atlases, and to continue to collect additional evidence in small ways.

The debate between generativists and the linguistic atlases appears clearly in the review by Samuel Keyser (1963) of Kurath and McDavid's PEAS (1961). He attacked the phonemicization practices of PEAS and used rule ordering in a generative subsystem to account for the diphthongs in *five*, *twice*, *down*, and *out* for informants from Charleston, SC, New Bern, NC, and Winchester, VA. He proposed two rules for generating the observed pronunciations from an underlying form (1963, 310), and suggested that rule ordering could account for different dialects. With this procedure, Keyser predefined a group of speakers, from one segment of the Atlantic States region as represented by his sample informants; he then predicted the inventories for two diphthongs in three subregions on the basis of system. This is manifestly not what Kurath and McDavid were trying to do when they used linguistic atlas data to try to find the boundaries of large Neogrammarian dialect regions. To the extent that the review is not simply a disagreement about phonemicization theory, Keyser did not address dialect in the same sense as Kurath and McDavid took it, and the review argues at cross purposes. Whether or not Keyser's generative subsystem actually accounts for the diphthong inventories of the informants he selected (see Davis 1983, 138–139), their arguments do not actually address each other. The generative model just talks about something different from what linguistic atlases are about. Thus, generativists who dismiss linguistic atlases and their data as "performance-based" are correct in terms of their own model, but the argument is not effective in any larger sense about how we should study speech. It is simply true that linguistic atlases are not generative.

The best reply to such arguments is the new discovery that speech is a complex system. Technological advances such as sound recording and computers, along with advances in reliable survey research methods, have made it possible to do now what Saussure did not find feasible a hundred years ago: to study what people actually say and write, *parole*. Linguistic atlases have kept up with technological change, for all the time that they were unpopular at the end of the twentieth century, and the Big Data perspective in them has been crucial to the recognition of the operation of the complex system of speech. The title of the foundational study of speech as a complex system, *The Linguistics of Speech* (Kretzschmar 2009), directly addresses the Saussurean dichotomy between *langue* and *parole*, between linguistic structure and linguistic behavior. We now know that linguistic behavior is prior to any sort of structure: our impression of grammars, as well as of dialects, derives from our perception of the nonlinear frequency profiles that always emerge from the complex system. Grammar is not generative but is instead an observational artifact, something that derives at some remove from the language behavior in the continual interactions of speakers in every location and in every situation of use (see Kretzschmar 2015, Ch. 4). The fact that linguistic atlases have survived all this time has, at length, turned the tables on Saussure's decision to prefer linguistic structure. While it will take more time to come for the field of linguistics to take advantage of our new understanding of complex systems, and thereby to find the right balance between structure and behavior, the systematic survey research of linguistic atlases will provide a wonderful source of data through which this new balance can be achieved.

NOTES

1 As explained in Kretzschmar (2009, 53–55), Saussure argued that "concrete entities," linguistic features of different kinds, are what we choose to isolate from the stream of speech. They are not naturally given. Pederson here is suggesting that units like morphemes, phonemes, allophones, and distinctive features are self-evident in that we choose to isolate them and recognize them for themselves.

- 2 These selected items are cues to elicit what the informants will say. So, if you were trying to elicit words for what to call a *thunderstorm*, the item might be "If there is thunder and lightning with it, you are having a ____." The wording of questions was not specified in Pederson's atlas work, so that a fieldworker might write down the response to an item when it occurred in conversation, or as part of a more general question like, for *thunderstorm*, "Tell me about the weather around here."
- 3 'Im philologischen Streit um die Ausnahmslösigkeit der Lautgesetze sollten diese Zeugnisse der Mundarten rings um Düsseldorf den geographischen Beweis für die Anhänger dieser Lehre erbringen. Das Ergebnis führte zur gegenteiligen Meinung. 1885 berichtet er [Wenker] vor der Philologenversammlung in Gießen: "Ich lebte in der schönen und beruhigenden Überzeugung, diese Charakteristika müßten ganz oder nahezu ganz zusammengehen. Jene Voraussetzung erwies sich bald genug als eine durchaus irrite, die Grenzen der vermeintlichen Characteristika liefen eigensinnig ihre eigenen Wege und kreuzten sich oft genug".'
- 4 'Der Sprachatlas des Dt. Reiches ist geboren aus dem Kampf um die Ausnahmslösigkeit der Lautgesetze (s. §20). Er hat dieses Postulat der Junggrammatiker nicht bestätigt.'
- 5 'Der Sprachatlas zeigt im Gegensatz zu dieser Annahme, daß im Grunde jedes einzelne Wort und jede einzelne Wortform ihre eigenen Geltungsbereiche, ihre eigenen Grenzen im Sprachraum besitzen.'
- 6 'Une vaste tapisserie dont les couleurs variées se fondent sur tous les points en nuances insensiblement dégradées'.
- 7 "Si nous pouvions établir un Atlas linguistique de la France au XIe siècle, nous verrions sans doute les limites dialectales se dessiner avec une précision qui démentirait G. Paris."

REFERENCES

- American Heritage Dictionary of the English Language*, 5th ed. 2011. Boston: Houghton Mifflin Harcourt.
- Bach, Adolf. 1950. *Deutsche Mundartforschung*, 2nd ed. Heidelberg: Winter.
- Bloomfield, Leonard. 1933. *Language*. New York: Holt.
- Chambers, Jack. 1998. "Inferring dialect from a postal questionnaire." *Journal of English Linguistics*, 26: 222–246.
- Davis, Lawrence. 1983. *English Dialectology: An Introduction*. University, AL: University of Alabama Press.
- Dollinger, Stefan. 2015. *The Written Questionnaire in Social Dialectology: History, Theory, Practice*. Amsterdam: Benjamins.
- Durkheim, Émile. 1895. *The Rules of Sociological Method*. Paris: Alcan.
- Gilliéron, Jules. 1918. *Généalogie des Mots qui Désignent l'Abeille*. Paris: Champion.
- Gilliéron, Jules. 1921. *Pathologie et Thérapeutique Verbales*. Paris: Champion.
- Gilliéron, Jules. 1922. *Les Étymologies des Étymologistes et Celles du Peuple*. Paris: Champion.
- Gilliéron, Jules, and J. Mongin. 1905. *Scier dans la Gaule Romance*. Paris: Champion.
- Gilliéron, Jules, and Edmond Edmont. 1902–1910. *Atlas Linguistique de la France* (9 vols.). Paris: Champion.
- Goebl, Hans. 1982. *Dialektometrie*. Vienna: Austrian Academy of Science.
- Goebl, Hans. 1990. "Methodische und Wissenschaftsgeschichtliche Bemerkungen zum Diskussionkomplex 'Unita Ladina'." *Ladinia*, 14: 219–257.
- Goebl, Hans. 2003. "Graziadio Isaia Ascoli, Carlo Battisti e il ladino." In *I Linguaggi e la Storia*, edited by Antonio Trampus and Ulrike Kindl, 273–298. Bologna: Il Mulino.
- Grzega, Joachim. 2009. *Sources for Onomasiological Studies*. <http://www1.ku-eichstaett.de/SLF/EngluVglSW/OnOn-4.pdf>.
- Jaberg, Karl. 1936. *Aspects Géographique du Langage*. Paris: Droz.
- Jaberg, Karl, and Jakob Jud. 1928–1940. *Sprach- und Sachatlas Italiens und der Südschweiz* (8 vols.). Bern: Zofingen.
- Keyser, Samuel. 1963. "Review of Kurath and McDavid 1961." *Language*, 39: 303–316.
- Kretzschmar, William. 1996. "Quantitative areal analysis of dialect features." *Language Variation and Change*, 8: 13–39.
- Kretzschmar, William. 2002. "Dialectology and the history of the English language." In *Studies in the History of English: A Millennial Perspective*, edited by Donka Minkova, and Robert Stockwell, 79–108. Berlin: de Gruyter.

- Kretzschmar, William. 2009. *The Linguistics of Speech*. Cambridge: Cambridge University Press.
- Kretzschmar, William. 2013a. "Computer mapping of language data." In *Research Methods in Language Variation and Change*, edited by Manfred Krug, and Julia Schlüter, 53–68. Cambridge: Cambridge University Press.
- Kretzschmar, William. 2013b. "GIS for language and literary study." In *Literary Studies in a Digital Age: An Evolving Anthology*, edited by Ray Siemens, and Ken Price. New York: MLA. [online] <http://dlsanthology.dev.mlacommmons.org/>
- Kretzschmar, William. 2015. *Language and Complex Systems*. Cambridge: Cambridge University Press.
- Kretzschmar, William, Jean Anderson, Joan Beal, Karen Corrigan, Lisa Lena Opas-Hänninen, and Bartłomiej Plichta. 2006. "Collaboration on corpora for regional and social analysis." *Journal of English Linguistics*, 34: 172–205.
- Kretzschmar, William, and Edgar Schneider. 1996. *Introduction to Quantitative Analysis of Linguistic Survey Data: An Atlas by the Numbers*. Thousand Oaks, CA: Sage.
- Kurath, Hans. 1939. *Handbook of the Linguistic Geography of New England*, 2nd ed. Providence, RI: Brown University, for ACLS.
- Kurath, Hans. 1939–1943. *Linguistic Atlas of New England* (3 vols.). Providence, RI: Brown University, for ACLS.
- Kurath, Hans. 1949. *A Word Geography of the Eastern United States*. Ann Arbor, MI: University of Michigan Press.
- Kurath, Hans. 1970. *The Investigation of Urban Speech*. Publication of the American Dialect Society, 49: 1–40.
- Kurath, Hans. 1972. *Studies in Area Linguistics*. Bloomington, IN: University of Indiana Press.
- Kurath, Hans, and Raven McDavid. 1961. *The Pronunciation of English in the Atlantic States*. Ann Arbor, MI: University of Michigan Press.
- Labov, William. 1994. *Principles of Linguistic Change, Vol. I: Internal Factors*. Oxford: Blackwell.
- Labov, William, Charles Boberg, and Sherry Ash. 2006. *Atlas of North American English: Phonetics, Phonology and Sound Change*. Berlin: de Gruyter.
- McDavid, Raven, and Raymond O'Cain. 1980. *Linguistic Atlas of the Middle and South Atlantic States* (Fasc. 1–2). Chicago: University of Chicago Press.
- Orton, Harold, et al. 1962–1971. *Survey of English Dialects*. Intro. and 4 vols. in 3 pts. Leeds: Arnold.
- Orton, Harold, Stewart Sanderson, and John Widdowson. 1978. *The Linguistic Atlas of England*. London: Croom Helm.
- Orton, Harold, and Nathalia Wright. 1974. *A Word Geography of England*. London: Seminar.
- Pederson, Lee. 1986–1992. *Linguistic Atlas of the Gulf States* (7 vols.). Athens, GA: University of Georgia Press.
- Pederson, Lee. 1995. "Elements of word geography." *Journal of English Linguistics*, 23: 33–46.
- Pederson, Lee, Raven McDavid, Charles Foster, and Charles Billiard. 1972. *A Manual for Dialect Research in the Southern States*, 2nd ed. Tuscaloosa, AL: University of Alabama Press.
- Pop, Sever. 1960. *La Dialectologie*. Louvain: by the author.
- Saussure, Ferdinand de. 1916. *Cours de Linguistique Générale* (edited by Charles Bally, Albert Sechehaye, and Albert Riedlinger). Paris: Payot.
- Upton, Clive, Stewart Sanderson, and John Widdowson. 1987. *Word Maps*. London: Croom Helm.
- Wrede, Ferdinand, Bernhard Martin and Walther Mitzka, eds. 1927–1956. *Deutscher Sprachatlas*. Marburg: Elwert.

4 Structural Dialectology

MATTHEW J. GORDON

4.1 Introduction

In the 1950s, several scholars, led by Uriel Weinreich, advocated a new approach to dialect study they called “Structural Dialectology.” Researchers in this paradigm introduced into dialectology elements of linguistic analysis borrowed from structuralist theory. At the level of phonology, they compared regional dialects in terms of the dialects’ phonemic systems rather than just on the basis of phonetic forms, as was common in traditional dialectology. A number of the methods and ideas developed in structural dialectology shaped the field in general, and its legacy is still felt in current research such as the *Atlas of North American English*.

Examining dialectological research done in the early twentieth century, one is struck first by the extraordinary level of detail. In linguistic atlases of the period, dialect variation is recorded in long lists of lexical alternatives and in phonetic transcriptions bursting with dia-critical fine-tuning, and the geographical distribution of the variant forms is represented in exquisite maps. The dedication it took to produce this work is certainly admirable, but it is hard to know what to do with the results. It is interesting to read of regional alternatives to *teeter-totter* and *seesaw*, and to see that some people pronounce *ewe* as [jou], whereas others say [jʌʊ], [hiu], [ɛu], and so on. But drawing larger lessons from the copious examples presents a challenge.

Critical examinations of the linguistic atlas tradition raise questions of representativeness. Whose speech is studied? Which components of languages are investigated? How are the data collected? Dialectologists of this early period primarily studied the speech of older rural people by eliciting select words and transcribing their pronunciations in phonetic detail (Chambers and Trudgill 1998). In the middle of the last century, this style of dialectology came under fire, with challenges issued from both inside and outside the field (see Petty 1980, 101–116). Sociologists, anthropologists, and eventually, sociolinguists called for expanding the types of speakers included in dialectological studies. Critics also decried the limitations of the methodology for exploring only a narrow slice of what distinguishes dialects by emphasizing lexical evidence, and for collecting data in only one stylistic context in which subjects were always aware that their usage was under investigation.

This chapter examines a related mid-twentieth-century effort to bring much-needed reform to dialectology and to bridge the gap between dialect study and mainstream linguistic theory. At the time, structuralism occupied that mainstream, and the call to improve communication across the disciplines was issued under the structural dialectology banner.

Key issues that emerged from research carried out according to this new approach are reviewed below, but the stage for this discussion is first set with some background on structuralism and its perceived conflict with dialectology.

4.2 Structuralism

The structuralist framework was developed toward the close of the nineteenth century and played a prominent role in linguistics throughout much of the twentieth century. In outlining the birth of linguistic structuralism, Seuren (1998, 144–145) underscores the work of the Polish scholar Jan Baudouin de Courtenay, who pioneered the concept of the phoneme and whose writings influenced Prague School linguists such as Nikolai Trubetzkoy and Roman Jakobson. The figure most often credited with establishing a new structuralist approach to language is the Swiss linguist Ferdinand de Saussure, who developed a number of ideas that became guiding principles of the field. These include the distinction between *langue* (language) and *parole* (speech), which contrasts the language system existing in the minds of native speakers (*langue*) from actual productions of speech (*parole*).

The view of language as a system lies at the center of structuralism. In the famous description by Meillet, who studied with de Saussure, language is “un système où tout se tient” (“a system where everything holds together,” as the phrase is commonly translated; see further, Koerner 1996). Language structure derives from a complex array of interlocking pieces across multiple layers. Sounds exist in a web of relationships at the level of phonology, morphemes at the level of morphology, and so on. Phonemic contrast is one such relationship in phonology. For example, aspirated [p^h] and plain [p] function differently in Hindi and English: in Hindi they represent distinct phonemes, whereas in English they are connected as allophones of a single phoneme.

The structural linguist therefore treats language “as a unique and closed system whose members are defined by opposition to each other and by their functions with respect to each other, not by anything outside of the system” (Weinreich 1954, 388). One of the hallmarks of this approach is the insistence on describing each language in its own terms, based solely on the observed patterns. In the United States, the rise of structural linguistics was closely connected to the study of “American Indian” (i.e., Native American) languages, as reflected in the careers of the two early pillars of American structuralism, Edward Sapir and Leonard Bloomfield (see Hymes and Fought 1981). Languages like these cannot be accurately analyzed in terms of a pre-established set of categories informed by the traditional study of Indo-European languages. Instead, the analyst must approach each language objectively, deploying a methodology of “discovery procedures” to reveal linguistic structures (Harris 1951).

The description proceeds systematically, working first at the level of phonology before moving on to morphology and grammar. Many of these techniques will be familiar to current students of linguistics. Thus, sounds are investigated by laying out their distributions to establish patterns of contrast or complementary distribution. The linguist may also appeal to the judgments of native speakers as to whether two sounds are the same or different. Similarly, morphemes are identified by carefully comparing utterances that differ minimally in meaning.

The structuralist view of language stands at odds with dialectology on several fundamental points. Most significant is the assumption of the “closed system,” as noted by Weinreich (1954, 388). Naturally, structuralists recognize that languages vary in terms of dialects as well as speaking styles. But the researcher sets out to describe a particular *variety* of a language. This method reflects the view of language as a system of structural relations. If one dialect differs structurally from another, it represents a distinct system that would need to be

described on its own terms. This approach tends to treat dialect variation as an obstacle to the analysis, a source of noise to be filtered out from the data (see also Gordon 2013, 30–33). Clearly, then, structuralism stands in strong contrast with dialectology, where documenting differences drives the research agenda.

This disciplinary contrast is heightened by the structuralists' insistence that valid evidence comes only from observations of the language system itself. Thus, the argument that a given language has, say, an inflectional category of tense or a phonemic stress contrast rests on the comparison of forms within that language, rather than on any preconceptions of how languages (generally, or in some areal or genetic group) operate. Dialectologists, while sharing some of this perspective—especially with regard to empirical data—do not shy away from looking beyond language structure for explanations.

The structuralist framework dominated linguistic thinking in both Europe and North America throughout much of the twentieth century. The approach that began in linguistics spread to the study of literature and anthropology. Structuralism's reign in its home discipline ended in 1957, when the "Chomskyan revolution" ushered in the era of Generative Grammar. At least, so goes the narrative told by supporters of the new orthodoxy (e.g., Newmeyer 1986). In truth, many structuralist ideas continue to influence linguistic research today.

4.3 Dialectology as a Non-Structural Linguistics

Many people find dialect differences fascinating, and there is a long history of amateur-led dialect study. In England, for example, people with an interest in folklife and local history have been collecting samples of rural dialect, often in the form of glossaries or regional vocabulary, since the seventeenth century (Petyt 1980; Shorrocks 2000). The focus of these early amateur dialectologists on lexis reflected a broader antiquarian enthusiasm, as the dialect words were taken as archaisms. Grammar and accent were of less interest, since they were typically viewed as mere corruptions of standard English (Shorrocks 2000, 87). When linguistics arose as a scholarly discipline in the nineteenth century, dialect study became more systematic. Still, some aspects of the earlier tradition of amateurism endured.

The emphasis on rural speech certainly persisted among professional dialectologists, and interest in lexical variation also continued to drive dialect study for much of the twentieth century. To be sure, pronunciation and grammar were not excluded from the agenda, but they typically played second fiddle to vocabulary studies. It is also fair to suggest that the historical leanings of early informal dialect study persisted. Dialectologists have traditionally sought to record the retention of older words and pronunciations. Thus, in most cases their sampling procedures favor older subjects, whose speech can provide a window into an earlier time period.

As suggested above, the dialectologist's interest in historical trends conflicts with the structuralist's commitment to system-internal explanations. Following de Saussure, structural linguists tend to paint a sharp line between synchronic and diachronic matters, and their procedures for describing the workings of a given language draw only on synchronic evidence. Indeed, the conflict in scholarly practices looms even larger when we remember that the history that dialectologists engage with is not just linguistic. Discussions of such non-linguistic topics as settlement patterns and topography abound in dialectological research (see Kretzschmar 1998, 174–175).

The lexical focus of much dialectological research was certainly out of step with structural linguistics. Documenting the word stock of a language was incidental to the endeavor of describing the language system in structural terms. Simply put, words were not regarded as structural elements. As Kurath (1945, 207) described this view, "The vocabulary or lexicon of

a language does not constitute a linguistic system but refers to the ‘practical world’ in which the speakers of a given language move and have their being.”

Even when dialectologists collected material that might fall within the structuralist’s purview, their data-handling practices failed to promote cross-disciplinary conversations. In phonology, for example, dialectologists documented pronunciation by transcribing targeted words uttered by the interviewee. The objective was to record in great phonetic detail how the relevant words were spoken. The ideal fieldworker would operate like a transcribing machine, accurately recording each utterance in precise phonetic notation (Moulton 1972, 215). The method sought to document speech with a minimum of interpretation from the researchers. In fact, dialectologists’ commitment to “a full and fair presentation of their evidence” (Kretzschmar 1998, 167) runs very deep, and has guided the dissemination of results since the field’s earliest days. The best-known product of dialectological research, the linguistic atlas, contains map after map displaying the variants of some feature recorded from speakers across a certain territory (see Kretzschmar, this volume). Critics have dismissed dialectology as having nothing to contribute to linguistic theory. As Trudgill (1999, 2) characterizes this critique, “The accusation has been one of ‘butterfly collecting’—that dialectologists are engaged in collecting data for the sake of collecting data.” But those who dismiss dialectology in this way fail to recognize the priorities that drive this line of research. Dialectology is “a data-oriented discipline,” as McDavid (1980, 280) puts it, adding:

However fine the theoretical extrapolations one may wish to make, the dialectologist’s first duty is to present the data in such a way that any reader can replicate the conclusions – or failing to replicate them, can show where the original statement went astray.

In this conflict over how the results of dialectological research were presented we see the seeds of the movement toward “structural dialectology.” Researchers working within this paradigm recognized the value of the meticulously collected data that underpin traditional dialectology, arguing that a more meaningful description of dialect variation results when that empirical foundation takes into account broader structural principles.

4.4 Weinreich’s Watershed Essay

While some scholars had been advocating structural approaches to dialect study since the 1930s, the catalyst behind the explosion of this work in the second half of the twentieth century was a 1954 essay by Uriel Weinreich, in which he asked “Is a structural dialectology possible?” It should come as little surprise that Weinreich answers his question in the affirmative. The issues he highlighted set the research agenda in this area for decades thereafter.

Weinreich sketches the problem of the apparent incompatibility between dialectology and structural linguistics in terms that are now familiar. He critiques the structuralist assumption of a homogeneous system as a fiction, noting that even an individual idiolect shows variation along a stylistic dimension. In comments that foreshadow the argument developed more fully in Weinreich, Labov, and Herzog (1968), he pleads for linguists to abandon the hypothesis of “absolute uniformity” in favor of exploring descriptions that can represent the inherent variability of language (1954, 389–390). Naturally, any rapprochement requires changes to dialectological practice as well. Weinreich challenges the tradition of ignoring structural relations, whereby “existing dialectology usually compares elements belonging to different systems without sufficiently stressing their intimate membership within those systems” (1954, 391). The clearest illustration of that non-structural approach is found in the habit of recording pronunciations in purely phonetic terms, without noting the *phonemic* status of the sounds.

To facilitate the proposed structural approach to dialectology, Weinreich introduces the notion of “diasystem,” which represents a higher-level construction of the linguistic system that is composed of the various systems found in a language’s dialects. More practically, the diasystem offers a formalism for representing points of structural agreement and divergence between dialects. Consider the following example from Yiddish, which I have simplified from Weinreich’s description (1954, 394):

$$\begin{matrix} // & \underline{1/i:\sim i/} & \approx e & \approx \underline{1/a:\sim a/} & \approx o & \approx u & // \\ 1,2 & 2 i & & 2 a & & & \end{matrix}$$

Double slashes mark the representation of the diasystem for the varieties denoted by the subscript numerals. Here, 1 stands for the Polish dialect of Yiddish, and 2 for the Lithuanian variety. Single slashes and tildes show phonemic contrasts within a dialect, whereas double tildes represent contrasts at the level of the diasystem (i.e., across all the varieties). This formula thus compares the seven-vowel system of the Polish dialect with the five-vowel system of Lithuanian Yiddish. These varieties agree in distinguishing /e/, /o/, and /u/, but differ by virtue of the length contrasts of /i:~i/ and /a:~a/ in Polish Yiddish where Lithuanian has only /i/ and /a/, respectively.

The diasystem notation provides a convenient means of summarizing dialect differences in structural terms. As Weinreich notes (1954, 394), however, it serves best for characterizing differences of phonemic inventory, mapping the phonemic contrasts of one variety onto another. But dialects also differ in terms of which words contain which phonemes. For example, both the Polish and Lithuanian dialects of Yiddish distinguish /o/ and /u/, but some words that have /o/ in one variety have /u/ in the other. Weinreich acknowledges that such distributional differences cannot readily be discerned from a diasystemic representation like the one shown above.

The construction of a diasystem—indeed, the ability to compare dialects from a structural perspective at all—depends on having some means of identifying the systems to be compared. Here we face an age-old dialectological challenge: how to determine boundaries between dialects, where the variation generally falls along a continuum. How many speech differences between two locations must we identify to count them as representing different dialects? As Weinreich observes (1954, 396), the problem becomes especially acute for researchers working within a structural dialectology, because the approach requires the breaking down of the speech continuum into discrete varieties for analytical purposes. It had become commonplace for dialectologists to supplement their linguistic evidence with “extra-structural criteria” such as geographical features in designating dialect boundaries. To Weinreich, this style of “external dialectology” has much to offer, and he recommends it as a useful supplement to research guided by structural principles. He imagines a scenario in which a careful structural analysis might highlight differences of historical or theoretical interest that could be overlooked by traditional approaches. In such an instance, however, a satisfying explanation of why those differences fall where they do might draw on patterns of communication or other “external” factors (1954, 398–399).

In sum, Weinreich sought to expand dialectology by incorporating elements of structuralism. Analyzing dialects in structural terms, he maintained, would not only make dialectologists’ work more relevant to a wider community of scholars, but would also produce stronger accounts of the linguistic phenomena they study. He stated very clearly, however, that structural dialectology does not supplant existing approaches; rather, “[i]ts results promise to be most fruitful if it is combined with ‘external’ dialectology without its own conceptual framework being abandoned” (1954, 400).

4.5 Describing Dialects Structurally

Structural dialectology shares with traditional approaches the goal of understanding similarities and differences across varieties of a language. Where the structural dialectologist diverges from other paths is in how dialects are compared. Moulton (1968) described the contrast as it plays out in phonological studies:

... traditional dialectology asks only a single question, for example: What does the vowel of a given word sound like at each of the many points under investigation? This is a useful sort of question to ask, and it has led to many valuable insights into the geographical dimension of human language. Structural dialectology also tries to obtain exactly this same information; but it then goes on to ask a second and more revealing question, namely: What position does this vowel occupy within the total vowel system at each of the many points under investigation? (Moulton 1968, 453)

The reasoning behind this “second and more revealing question” lies in the diversity of ways that dialects may differ even at the phonological level. Documenting phonetic forms, as Moulton acknowledges, is a crucial step in a dialectological investigation, but does not tell the whole story, and may in fact lead researchers astray.

Weinreich (1954, 391–392) presents a constructed example illustrating the answers that emerge from asking different questions. He imagines four speakers representing four locations. Each says the word “man,” with the following results:

- A. [man]
- B. [man]
- C. [mɒn]
- D. [mɒn]

These data represent the responses to the question posed by Moulton’s traditional dialectologist. This phonetic picture suggests a binary division that groups A with B and C with D.

A structural analysis might result in a very different picture. Suppose, as Weinreich posits, that the dialect spoken by A has a phonemic length distinction such that the [a] used here represents a short /ă/ phoneme, which contrasts with long /a:/-. Dialect B has no such distinction: [a] represents the phoneme /a/. In C’s dialect, the [ɒ] is a positional allophone (of a phoneme /a/) that appears between nasal consonants. The [ɒ] in D’s dialect is a conditioned allophone, but of /o/ rather than /a/. In this way, posing Moulton’s second question adds a phonemic layer to the dialect picture:

- A. [man] =/män/
- B. [man] =/man/
- C. [mɒn] =/man/
- D. [mɒn] =/mon/

From the perspective of the vowel system, it seems now that B and C have a stronger affinity that separates them from both A and D.

Weinreich’s simple example demonstrates the light that structural analysis can shed on “raw” dialectological data, and illustrates some of the ways that dialects may differ phonologically. He acknowledges the influence of earlier work by the Prague School linguist Nikolai Trubetzkoy, who had outlined a typology of dialect differences in a 1931 paper (see Petyt 1980 for a summary in English). Trubetzkoy lays out three major ways in which dialects vary from a phonemic perspective (see also Wells 1982, 73ff.). Under the heading

“phonological differences,” Trubetzkoy lists cases in which one dialect’s inventory contains phonemes that are missing from another dialect’s inventory. Weinreich’s Yiddish example, in which the short /i/ ~ long /i:/ contrast found in Polish dialects corresponds to the single phoneme /i/ in Lithuanian dialects, illustrates such an inventory difference.

Trubetzkoy labels as “phonetic differences” those instances in which dialects share a phoneme but differ in how it is phonetically realized. The difference between B and C in Weinreich’s constructed example falls into this category, as the /a/ appears as [a] in one dialect but as [ɒ] in the other.

The third of Trubetzkoy’s major types involves a difference of “etymological distribution.” Here, two dialects again have the same phonemic inventory, but differ in terms of the lexical incidence of those phonemes. Weinreich alludes to a situation of this kind in noting that both the Polish and the Lithuanian Yiddish dialects distinguish /o/ from /u/ while also pointing out that some cognate words have /o/ in one variety and /u/ in the other.

Trubetzkoy’s typology of dialect differences clearly pertains only to the realm of phonology, and this is the area in which structural dialectology has its most direct applications. Certainly, the literature on structural dialectology is dominated by discussions of phonological research. Still, the general principles of structuralism, especially the exploration of linguistic elements as members of a system, can be applied to other dimensions of dialect variation.

Given the general lack of attention paid to the lexicon by structural linguists (see above), it comes as little surprise that vocabulary has not figured prominently in structural dialectological work. Nevertheless, Weinreich (1954, 399) offers an example from Yiddish that at least points toward a structural treatment of lexical variation. The word *shtul* means “chair” in some dialects, but more specifically “easychair” in others, and thus plays different roles in the semantic systems of the different dialects. In areas where *shtul* means “chair” it stands in contrast with words designating finer distinctions such as *fotél* “easychair” and *benkl* “little bench,” whereas in other areas *shtul* marks the more specific meaning and *benkl* serves as the broader term.

Pilch (1972) discusses similar examples of lexical difference, and suggests applications of structural dialectology to grammatical variation. In the area of morphology, for example, he notes (1972, 176) that the second-person pronoun *you* functions differently in standard English, where it marks both singular and plural, from how it does in the many dialects in which it stands as singular in opposition to plural forms such as *youse* or *y’all*. In a more syntactic vein, he considers cases of “constructional syncretism,” where a difference between two structures becomes neutralized. In some dialects of German, according to Pilch’s example (1972, 178), the adjective *kalt* “cold” functions differently in an impersonal construction such as *Ihr ist kalt* “she is feeling cold” from how it functions in *Sie ist kalt* “she is cold-hearted.” This distinction is collapsed in other varieties, with both meanings being conveyed by *Sie ist kalt*. Such examples hint at how structural dialectology might be extended beyond phonology.

4.6 From Phones to Phonemes

A key plank in the platform for a structural dialectology is the description of pronunciations in phonemic terms. The dialectologist should produce an account not only of the sounds that occur in targeted words but also of what phonemes those sounds represent. Approaching dialects on these structural terms allows one to distinguish surface-level observations, such as the phonetic contrast of [man] versus [mɒn] in Weinreich’s example, from the presumably more consequential patterns at the phonemic level. Still, the methods of traditional dialectology produce data that are not always easily interpreted in structural terms, and the

process of moving from the recorded sounds (phones) to phonemes can be less than straightforward.

The way of documenting pronunciation differences in the linguistic atlas tradition is a useful and, some would argue, even necessary step in establishing dialect boundaries. As noted above, having fieldworkers record pronunciations in such fine-grained phonetic detail was a strategy designed to avoid potential bias introduced by interpreting what was heard in terms of preconceived categories. In any case, some dialectologists argued, this practice does not preclude later analysis through a structuralist (or any other theoretical) lens. Supporters of the traditional approach claimed, for example, that the phonemic system of a given dialect remained recoverable amid the detailed phonetic records produced in a linguistic atlas project (e.g., Orton 1962, 28, quoted by Rydland 1972, 309).

Rydland (1972) takes on this claim with an examination of data from two locations represented in the *Survey of English Dialects* (SED). True to expectations, the amount of phonetic detail recorded for the SED is breathtaking. For the dialect of the village of Andreas on the Isle of Man, Rydland reports that the SED materials list 122 differences of vowel quality, including 47 monophthongs and 75 diphthongs, many of which also appear with up to three different designations of vowel length. Based on their phonetic similarity, Rydland organizes the raw data into 30 categories for which he attempts to establish the phonemic status. Drawing on standard structuralist procedures such as the presence of contrastive pairs, he confidently identifies 17 vowel phonemes. But several of the phonetically defined categories remain difficult to interpret in phonemic terms. The SED materials provide minimal pairs (e.g., [sou] "so" versus [sou] "sew"), which support the conclusion that the two sounds represent distinct phonemes. However, Rydland remains doubtful, noting that "the examples may not be properly minimal pairs: it is possible that they merely exemplify free variation between phonetically similar sounds" (1972, 316).

Ultimately, the uncertainty stems from the limitations of the data. The SED records some sounds in only a few words, and records those words only a few times. Moreover, when a speaker uses a particular sound in a given word, the fieldworker does not normally ask whether that word could also be pronounced with a different vowel. Such lines of questioning were a standard component of the discovery procedures followed by structural linguists (e.g., Harris 1951), and would have potentially eliminated some of the ambiguity in the SED data, had they been adopted. Rydland acknowledges this point. He concludes that even though the "SED does not provide enough material for a complete phonemic analysis of these dialects," it does allow one to draw "fairly close approximations to the actual phonemic systems." He also recommends supplementing the existing material by carrying out "additional, phonemically orientated field-work" targeting phonologically ambiguous points (1972, 324).

Kurath and McDavid (1961) weigh in on the suggestion that phonemic systems can be inferred from the phonetic records of traditional dialectology in *The Pronunciation of English in the Atlantic States* (PEAS). This landmark study draws on the records of the *Linguistic Atlas of the Eastern United States*, a project directed by Kurath for which the fieldwork was carried out in the 1930s and 1940s. While these atlas data "were recorded as heard by trained observers without reference to a general scheme of phonemicization or to the phonemic structure of the separate idiolects" (Kurath and McDavid 1961, 1), the material has been reexamined through a structuralist lens, and PEAS describes the documented variation in phonemic terms. In their main data chapters, Kurath and McDavid walk through each vowel phoneme, describing the regional variants or "diaphones," as well as any positional allophones. They also explore phonological variation in terms of lexical incidence, describing regional patterns related to the appearance of particular phonemes in particular words (e.g., /ɛ/, /ɪ/, or /i/ in *deaf*).

Juxtaposed with Rydland's (1972) study, PEAS certainly presents a tidier picture of how linguistic atlas data can be configured within a structuralist framework. Actually, this

comparison is unfair, since Kurath and McDavid do not actually work out their phonemic analysis in the book but rather present it, with little argumentation, as a reasonable account of the facts. Each phoneme has several phonic types (diaphones and allophones) associated with it, but the details of how those types were assigned to particular phonemes do not appear in *PEAS*.

We can gain a sense of how *PEAS* operates as a work of structural dialectology via a simple example. Eastern dialects of American English vary in their treatment of /ɛ/, /e/, and /æ/ before /r/, with some areas preserving a three-way distinction that is collapsed into two, or even into a single vowel. Thus, for some speakers *merry*, *Mary*, and *marry* are distinct, while others distinguish *marry* from the homophonous *merry* and *Mary*, and others make no distinction between the three forms at all. Prior to *PEAS*, a researcher interested in this variation might consult the atlas materials where the pronunciations of relevant words are recorded. This phonetic evidence, however, might be ambiguous in phonemic terms. As Pilch (1972, 167) noted, even if one finds words like *cherry* and *married* transcribed with different vowels,

The difference of transcription may mean contrast, but it may not. The fieldworker may have responded with slightly different transcriptions to two forms with the same vowel heard on different occasions. Or the informant may, within the normal field of allophonic dispersion, have used two slightly different variants minutely recorded by the fieldworker.

In *PEAS* we find maps representing the variation heard in particular words (just as in earlier work such as the *Linguistic Atlas of New England*), but the results are coded phonemically. Thus, for *Mary* (Map 50) Kurath and McDavid distinguish several phonetic realizations of /e/, as well as a separate category for locations where the phoneme /ɛ/ appears in this word. Moreover, their discussion explicitly notes where words of the *Mary* class rhyme with those of the *merry* and *marry* classes (1961, 124–125). As to how such determinations were made, we can infer a process that surveys the phonetic range of vowels in each word class and carves out phonemes on that basis. Keyser (1963, 304) suggests that Kurath and McDavid operated under a condition of “strong biuniqueness,” by which he means “the constraint that allophones of a given phoneme must be phonetically distinct from allophones of other phonemes.” Having made this theoretical assumption, the assignment of phones to phonemes seems relatively straightforward. Indeed, as Kurath and McDavid confidently observe, “if the diaphones of the /e/ of *paper*, the /ɛ/ of *head*, and the /æ/ of *ashes* are known, one can determine fairly safely which of the three phonemes a speaker has in *stairs* or *Mary*” (1961, 101).

4.7 Diasystemic Analysis

For the dialectologist, the point of determining the phonological system of a given variety is largely to provide a means of comparing that system with others. For reasons noted earlier, the fundamental orientation of dialectology toward tracking features across varieties (i.e., linguistic systems) runs against the grain of structuralism, which emphasizes relationships *within* a given system. Weinreich’s (1954) introduction of the diasystem represents an attempt to reconcile structuralist and dialectological thinking by constructing a higher-level system to account for partial similarities across related dialects. The diasystem relates the structural relationships within one system to those within others, indicating, for example, how phonemic contrasts in dialect A align with those in dialect B. The concept of the diasystem and the formalism associated with it (see above) generated substantial discussion and debate throughout the heyday of structural dialectology.

As the diasystem concept was being developed in dialectology, the intellectual climate in American linguistics was being shaped by interest in an “overall pattern” approach to English phonology. Developed by several authors over a period of a decade or so, this approach culminated in *An Outline of English Structure* (Trager and Smith 1951). The work presents an “overall pattern” in the sense that its system is meant to cover all varieties of English. It shares with Weinreich’s diasystem, then, the goal of constructing a system at a higher level than the individual idiolect or dialect. Still, Trager and Smith are less concerned with the details of how one dialect’s structure aligns with others than they are with producing a system flexible enough to describe any dialect. Thus, the system they propose is meant to embrace all possible systems of English. In the area of phonology, for example, they posit an inventory of nine simple vowels, each of which may function as a syllabic nucleus alone or in combination with one of three “semivowels,” /y w h/ (/y/ = IPA /j/). No speaker of English makes use of all 36 of these vocalic possibilities, as each dialect draws on a subset of them for its inventory.

Trager and Smith’s analysis drew criticism from a variety of scholarly directions (see, e.g., Kurath 1957, and works cited by Stockwell 1959). Many of the objections from dialectologists centered on the highly abstract treatment of vowels, and particularly on the “binary” interpretation of phonetic diphthongs as composed of two phonemes (e.g., the vowel of *out* as /a/+/w/). Kurath and McDavid (1961, 4) argued that such an approach resulted in counter-intuitive results when applied to actual speech patterns recorded in dialect studies. Moulton (1968, 457) joined in this challenge, arguing that dialect data from Swiss German revealed similar problems for a binary treatment of diphthongs. These critiques illustrate how evidence from structural dialectology can inform questions of interest to phonological and linguistic theory in general.

Despite certain shared goals, the overall pattern approach of Trager and Smith (1951) differs fundamentally from the diasystem construct in structural dialectology. In Pulgram’s (1964, 76) apt characterization, a diasystem “analyses dialects singly and then recombines them, for a good purpose, in some kind of formula,” whereas the overall system “merely adds them together.” Still, even the diasystem construct has faced a mixed reception among dialectologists. Cochrane (1959) elaborated the core concept by drawing a distinction between “diaphonemic” variation, whereby dialects differ in terms of their inventories of phonemes, and “diaphonic” variation, in which dialects differ only in how a shared phoneme is realized phonetically. He outlines a range of diaphonemic relations that serves as a typology of how phonemic contrasts may be aligned across dialects, illustrating the process of diasystem construction using a case study of dialects of Australian English.

Other investigators carrying out research within the framework of structural dialectology found the diasystem and its formalism less satisfying. Moulton (1960) takes exception to the way the diasystem concept as developed by Weinreich and Cochrane prioritizes phonemic inventories over lexical incidence. In his study of Welsh dialects, Thomas (1964) reiterates such concerns and identifies similar problems with treating morphological variation in diasystemic terms. The issue is partly notational. A diasystemic formula can easily represent cases of a single phoneme or morpheme in one dialect corresponding to two or more phonemes or morphemes in another dialect, but it is much harder to represent situations where cognates differ in their phonemic assignment or grammatical treatment across dialects. Drawing on Swiss German data, Moulton (1960, 176) notes that two dialects may appear identical when only their phonemic inventories are considered, whereas an examination of lexical correspondences (i.e., which words have which phoneme) shows almost no agreement. When presented in diasystemic notation, either view of the relationship between the dialects is misleading.

While highlighting the limitations of the diasystem construct, Moulton’s critique stems from using diasystems in ways other than Weinreich probably intended. What Moulton

(1960) is engaged in a dialectological study carried out from a structural perspective. The diasystem is a source of frustration because it fails to adequately show the intricacies of how the dialects he analyzes are related. But Weinreich had other purposes in mind in proposing the diasystem concept. He introduces it (1954, 389–390) as a mechanism to benefit “structural linguistic theory,” providing a higher-level system constructed from the “discrete and homogeneous systems” that share partial similarities. In this way Weinreich sought to broaden the focus of structural analysis beyond the single system (i.e., the dialect). The diasystem illustrates what dialectology can bring to the table of structural linguistics. That Moulton and other dialectologists saw little value in it is therefore not surprising. As Pulgram (1964) summarized these conflicting perspectives:

Moulton compares dialects in order to determine their relationship and their place in a larger system... Weinreich starts from the knowledge thus provided, and the structural comparison his diasystem yields is, I believe, something to be read out of it conveniently because of an adroit and meaningful arrangement of the information. In other words, the results of Moulton’s dialectology and comparison are grist for the mill of Weinreich’s. (Pulgram 1964, 80).

4.8 Isoglosses and Dialect Boundaries

The discussion of the diasystem has underscored some of the new insights that structural dialectology brought to theoretical (structural) linguistics in general, but the framework that Weinreich advocated undoubtedly had a greater and more lasting impact on dialectology itself. This impact is evident, as we have seen, in the new parameters that dialectologists came to consider in their analyses, but also in new ways of conceiving of boundaries between dialects.

How many dialects are there in a particular region, nation, and so on? This kind of question reflects a common conception of dialect variation, one that assumes that dialects exist as discrete entities separated by clear boundaries. Much to the frustration of the non-linguist who asks the question, the honest answer to the question is, “It depends on what criteria we use to distinguish dialects.” How different must two varieties appear in order to be considered representative of distinct dialects? Dialectologists may, in fact, challenge the premise and question whether dialects as such even exist. As Ivić (1962) framed the matter:

Upon closer inspection, considering the multitude of isoglosses crossing one another in all possible directions, territorial dialects have often proved to be fictitious entities, established on the basis of arbitrarily chosen features, or on the basis of extralinguistic criteria. This fact led either to the negation of the very existence of territorial dialects or to the opinion that this concept has a relative character, and that the central concept of dialectology should be that of the isogloss. (Ivić 1962, 34)

Ivić clearly directs his criticism at traditional approaches to dialectology, where the relative value of isoglosses in a given territory may be a subjective matter. Moulton (1968, 456) levels similar charges, and even suggests that investigators in this tradition would often give preference to those isoglosses that supported their preconceived vision of dialect areas.

The remedy prescribed by Ivić, Moulton, and other structural dialectologists was naturally to adopt structural criteria for dialect comparison. This approach, they claim, leads to a more objective description of regional variation. As Moulton (1968) optimistically described the process: “we analyze some part of the phonological structure [...], plot the locations of the various systems, and let the geographical structure of speech unfold as it will” (Moulton 1968, 456–457).

The argument for a structural dialectology is premised on the belief that structural differences between dialects are more significant than non-structural differences. Therefore, boundaries drawn on the basis of structural patterns offer not simply an alternative to the traditional picture, but a truer view. Kurath and McDavid certainly endorse this line of thinking when they write that for the purpose of “determining the degree of difference between dialects and in evaluating the relative importance of the boundaries between speech areas” (Kurath and McDavid 1961, 2), phonemic heteroglosses outrank those related to phonetic difference or to lexical incidence.

Researchers have used the results of a structural investigation to challenge dialect divisions based on traditional criteria. Moulton (1960) takes up this cause, showing how some boundaries that separate dialects of Swiss German on the basis of phonetic data disappear and new ones emerge when those varieties are compared instead in terms of phonemic systems. Stankiewicz (1957) carries out a similar structuralist revision of Polish dialects. He notes that the traditional approaches to Slavic dialectology were historically oriented, and in the case of Polish prioritized the reflexes of certain sound changes over other dialect differences, even when those changes resulted in purely phonetic isoglosses. A synchronic approach that contrasts varieties by their phonological inventories produces not only a different view but, Stankiewicz maintains (1957, 52), a more objective one in which the dialect areas are defined more clearly.

Delineating dialects on a structural basis can lead to the appearance of neatly bounded regions, and in this way may run to the opposite extreme of the problem observed above. Whereas a non-structural approach may paint a landscape of intersecting and non-intersecting isoglosses on which the investigator must arbitrarily impose some order so as to arrive at separate dialects, the structural approach draws strict boundaries between what are assumed to be unique systems. Stankiewicz (1957, 48) positions as a central issue for dialectology this “problem of defining the relation between continuity and discreteness.” He advises researchers to attend to both dimensions by describing what unites dialects as well as what distinguishes them. He outlines some ways that structural dialectologists might fruitfully expand their scope of inquiry by looking at phonemic systems in terms of their organizing features, noting, for example, that all Polish dialects are united by their lack of phonemic distinctions of stress and length, a property that also clearly differentiates them from neighboring Slavic languages. On the strength of such examples, Stankiewicz (1957, 56) argues that admitting “a broader frame of phonemic relevance enables us likewise to discern continuity, where on the basis of strict phonemic criteria we would see only discreteness.”

4.9 The Legacy of Structural Dialectology

For decades, linguists have produced work that confirms Weinreich’s (1954) suggestion that a structural dialectology is not only possible but also productive, with benefits for linguistic theory in general as well as for the study of dialect variation. The structural dialectology framework has proven fruitful in investigations focused on particular issues (e.g., Moulton 1960, Stankiewicz 1957, and other works cited throughout this chapter) and on a larger scale as with the *Linguistic Survey of Scotland* (Catford 1957) and *PEAS* (Kurath and McDavid 1961; see also citations in Moulton 1972).

Despite the tremendous interest it generated from the mid-1950s into the 1970s, structural dialectology as an intellectual movement or a sub-discipline of linguistics seems to draw little recognition from scholars today. When the term “structural dialectology” appears in current scholarship, it is typically paired with past-tense verbs and positioned as a historical topic. So, we might ask: what happened to structural dialectology?

One facile story of structural dialectology's demise ties it to the shrinking relevance of the two fields it sought to bridge. Structuralism met its fate as Generative Grammar ascended to prominence in linguistic theory following the Chomskyan revolution that was sparked in the late 1950s. A decade or so later, dialectology, which had long occupied a rather peripheral space in twentieth-century linguistics, was finally put out of business when its work on linguistic variation came to be taken over by sociolinguistics.

A more nuanced account of this history, which I give only hints of here, would observe that many elements of structural linguistics continue to hold sway in linguistic scholarship today, even within the Generativist camp. To give one obvious example, the phoneme still plays a principal role in phonology. Moreover, the structuralist practice of positing the language system as self-contained and uniform survives in Chomsky's focus on 'an ideal speaker-listener, in a completely homogeneous speech community' (1965, 3).

The other half of the story is even more problematic. The relationship between sociolinguistics and dialectology has been explored from several perspectives (see, e.g., Gordon 2013, Kretzschmar 1995, Trudgill 1999). It may suffice here to observe that, while sociolinguistics opened valuable new avenues for investigating linguistic variation, it did not supplant dialectology, as evidenced by the simple fact that researchers continue to label their field dialectology (as this volume testifies). Nevertheless, it would be a mistake to assume from this terminological continuity that dialectology has not evolved from the way it was practiced in the early twentieth century. In this respect, the push for a structural dialectology had a profound effect. Attention to structural patterns has become a regular, even necessary, element of dialect study. If the phrase "structural dialectology" has little currency today, we can chalk that up to the apparent redundancy of the initial adjective.

By way of conclusion, I discuss a recent study that is emblematic of the evolution of a dialectology shaped by structuralism. *The Atlas of North American English* (ANAE) (Labov, Ash, and Boberg 2006) explores regional variation across the continent, with a stated emphasis on phonological patterns. Though known as a pioneering figure in sociolinguistics, Labov positions the study squarely within the tradition of dialectology and identifies its major goal as "the re-establishment of the links between dialect geography and general linguistics" (2006, v). With these strong echoes of the call for rapprochement made by Weinreich (1954), who was Labov's mentor, it comes as no surprise to see the influence of structural dialectology throughout ANAE.

The structuralist orientation of the work is immediately apparent in the symbols chosen to represent sounds. Labov and his colleagues adopt a version of the phonemic notation promoted by structural linguists like Trager and Smith (1951). This kind of system evoked strong negative reactions from dialectologists like Kurath and Moulton (see discussion above), but it serves ANAE as a framework for exploring structural relations because it organizes the inventory of vowel phonemes into subclasses (i.e., short versus ingliding versus upgliding). The value of this classification stems largely from the predictions it makes about the dynamics of the vowel system.

Labov has drawn on this model of the structure of vowel systems extensively in his research on sound change (e.g., 1991, 1994, 2008). He has developed influential accounts about the mechanisms underlying common changes such as mergers and chain shifts. American structural linguists devoted significant attention to the former, which involve the loss of a phonemic contrast, but they took less interest in chain shifts because these changes involve subphonemic variation with no effect on the inventory of phonemes. Labov's thinking about chain shifts and mergers reflects a strain of structuralist theory associated with Weinreich's mentor, André Martinet (e.g., 1955). The approach treats phonological systems in spatial terms. So, for example, when one vowel shifts, we expect a reaction from its neighbors. The moving vowel may set off a chain shift by pushing a vowel out of its path or pulling a vowel into its wake, or the initial movement might trigger a merger as two vowels

come to occupy the same space. Moulton was an early advocate of this analytical approach. For example, in his studies of Swiss German dialects (1960, 1968) he argues that the various vowel systems observed today resulted from alternative strategies to remedy an earlier asymmetry (a “hole” in vowel space).

The ANAE methodology allows for the exploration of these structuralist conceptions of vowel system dynamics on a large geographical scale. The project relies on acoustic measurements of the first and second formants, which serve as correlates of vowel height and frontness and therefore present vowels as objects in a two-dimensional space. This technique makes it possible to investigate the status of the many vowels shifts active in American English today. Because their interview protocols also elicited pronunciations of certain key minimal pairs and asked participants directly about rhymes and homophones, the ANAE researchers could document examples of phonemic merger. This approach recalls techniques developed as discovery procedures for structural analysis (e.g., Harris 1951).

In sum, the ANAE explores questions of how and why vowel systems change, and does so from a perspective focused on structural and spatial relations among vowels. At the same time, its contribution to linguistic theory in general is matched by its significance for dialectology in particular. It stands as an unparalleled resource for information about regional patterns of North American English. It presents a valuable and fresh view of the dialect geography of the continent, and has already sparked a wave of research following up on its insights. For this reason, it serves not only to illustrate the survival of theories and practices developed in structural dialectology, but also to reaffirm the fruitfulness of this approach.

REFERENCES

- Catford, John. 1957. “The Linguistic Survey of Scotland.” *Orbis*, 6: 105–121.
- Chambers, Jack, and Peter Trudgill. 1998. *Dialectology*, 2nd ed. Cambridge: Cambridge University Press.
- Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.
- Cochrane, George. 1959. “The Australian English vowels as a diasystem.” *Word*, 15: 69–88.
- Gordon, Matthew. 2013. *Labov: A Guide for the Perplexed*. London: Bloomsbury.
- Harris, Zellig. 1951. *Structural Linguistics*. Chicago: University of Chicago Press.
- Hymes, Dell, and John Fought. 1981. *American Structuralism*. The Hague: Mouton.
- Ivić, Pavle. 1962. “On the structure of dialectal differentiation.” *Word*, 18: 33–53.
- Keyser, Samuel. 1963. “Review of *The Pronunciation of English in the Atlantic States* by Hans Kurath and Raven I. McDavid.” *Language*, 39: 303–316.
- Koerner, Ernst. 1996. “Notes on the history of the concept of language as a system ‘où tout se tient.’” *Linguistica Atlantica*, 18: 1–20.
- Kretzschmar, William. 1995. “Dialectology and sociolinguistics: Same coin, different currency.” *Language Sciences*, 17: 271–282.
- Kretzschmar, William. 1998. “Analytical procedures and three technical types of dialect.” In *From the Gulf States and Beyond: The Legacy of Lee Pederson and LAGS*, edited by Michael Montgomery, and Thomas Nunnally, 167–185. Tuscaloosa, AL: University of Alabama Press.
- Kurath, Hans. 1945. “Review of *Outline of Linguistic Analysis* by Bernard Bloch and George L. Trager.” *The American Journal of Philology*, 66: 206–210.
- Kurath, Hans. 1957. “The binary interpretation of English vowels: A critique.” *Language*, 33: 111–122.
- Kurath, Hans, and Raven McDavid. 1961. *The Pronunciation of English in the Atlantic States*. Ann Arbor, MI: University of Michigan Press.
- Labov, William. 1991. “The three dialects of English.” In *New Ways of Analyzing Sound Change*, edited by Penelope Eckert, 1–44. New York: Academic Press.
- Labov, William. 1994. *Principles of Linguistic Change, Vol. 1: Internal Factors*. Oxford: Blackwell.
- Labov, William. 2008. “Is a structural dialectology practical? Re-deploying Weinreich’s approach to diasystems.” In *Evidence of Yiddish*

- Documented in European Societies: The Language and Culture Atlas of Ashkenazi Jewry*, edited by Marvin Herzog, Ulrike Kiefer, Robert Neumann, Wolfgang Putschke, and Andrew Sunshine, 217–229. Tübingen: Niemeyer.
- Labov, William, Sharon Ash, and Charles Boberg. 2006. *The Atlas of North American English: Phonetics, Phonology and Sound Change*. Berlin: de Gruyter.
- Martinet, André. 1955. *Économie des Changements Phonétiques*. Bern: Francke.
- McDavid, Raven. 1980. *Varieties of American English: Essays by Raven I. McDavid, Jr., Selected and Introduced by Anwar S. Dil*. Stanford, CA: Stanford University Press.
- Moulton, William. 1960. "The short vowel systems of northern Switzerland: A study in structural dialectology." *Word*, 16: 155–182.
- Moulton, William. 1968. "Structural dialectology." *Language*, 44: 451–466.
- Moulton, William. 1972. "Geographical linguistics." In *Current Trends in Linguistics, Vol. 9: Linguistics in Western Europe*, edited by Thomas Sebeok, 196–122. The Hague: Mouton.
- Newmeyer, Frederick. 1986. *Linguistic Theory in America*, 2nd ed. Orlando, FL: Academic Press.
- Orton, Harold. 1962. *Survey of English Dialects: Introduction*. Leeds: Arnold.
- Petyt, Malcolm. 1980. *The Study of Dialect: An Introduction to Dialectology*. London: Deutsch.
- Pilch, Herbert. 1972. "Structural dialectology." *American Speech*, 47: 165–187.
- Pulgram, Ernst. 1964. "Structural comparison, diasystems, and dialectology." *Linguistics*, 4: 66–82.
- Rydland, Kurt. 1972. "Structural phonology and the Survey of English Dialects: A critical evaluation of the material." *Zeitschrift für Dialektologie und Linguistik*, 39: 309–326.
- Seuren, Pieter. 1998. *Western Linguistics: An Historical Introduction*. Oxford: Blackwell.
- Shorrocks, Graham. 2000. "Purpose, theory and method in English dialectology: Towards a more objective history of the discipline." In *Debating Dialect: Essays on the Philosophy of Dialect Study*, edited by Robert Penhallurick, 84–107. Cardiff: University of Wales Press.
- Stankiewicz, Edward. 1957. "On discreteness and continuity in structural dialectology." *Word*, 13: 44–59.
- Stockwell, Robert. 1959. "Structural dialectology: A proposal." *American Speech*, 34: 258–268.
- Thomas, Alan. 1964. "Some aspects of a structural dialectology." *The Transactions of the Honourable Society of Cymmrodorion*, (1964, Part II): 313–343.
- Trager, George, and Henry Smith. 1951. *An Outline of English Structure*. Norman, OK: Battenberg.
- Trubetzkoy, Nikolai. 1931. "Phonologie und Sprachgeographie." *Travaux du Cercle Linguistique de Prague*, 4: 228–234.
- Trudgill, Peter. 1999. "Dialect contact, dialectology and sociolinguistics." *Cuadernos de Filología Inglesa*, 8: 1–8.
- Weinreich, Uriel. 1954. "Is a structural dialectology possible?" *Word*, 10: 388–400.
- Weinreich, Uriel, William Labov, and Marvin Herzog. 1968. "Empirical foundations for a theory of language change." In *Directions for Historical Linguistics*, edited by Winfred Lehmann and Yakov Malkiel, 95–195. Austin, TX: University of Texas Press.
- Wells, John. 1982. *Accents of English* (3 vols.). Cambridge: Cambridge University Press.

5 Dialectology and Formal Linguistic Theory: The Blind Man and the Lame

FRANS HINSKENS

"The men of experiment are like the ant, they only collect and use; the reasoners resemble spiders, who make cobwebs out of their own substance. But the bee takes the middle course; it gathers its material from the flowers of the garden and field, but transforms and digests it by a power of its own"

(Francis Bacon, *Novum Organum*, 1620/2004, 153)

5.1 The Dialectological Investigation of Dialect Variation, in Brief

Contra the Neogrammarians, who introduced the concept of grammatically blind and lexically exceptionless sound laws—the effects of which can be obscured by analogy or borrowing—late nineteenth-century dialectologists explicitly took the idiosyncrasies of individual lexical items into account (as per Jaberg's axiom “in reality each word has its own particular history” (Jaberg 1908, 6)).¹ This tendency to concentrate on isolated linguistic forms is a manifestation of atomism, that is, the inclination not to embed in the grammar the phenomena under study. Findings are typically not interpreted in the context of existing linguistic theories, which has led to the marginalisation of dialectology in the wider field of linguistics.

A related (disputable) trait of traditional dialectology is its tendency not to embed the phenomena being investigated in the verbal repertoires of the speakers and the speech community under study. The sociodialectological approach to dialect variation does provide this possibility, but with its tendency to treat linguistic variables in isolation rather than properly embedding variation in language structure, the sociolinguistic approach to language variation and change unmistakably inherited some of the features of dialectology. Related to this is the fact that the construction, testing, and revision of theories of language change does not appear to have been given much priority in the sociolinguistic approach to dialect variation, the biggest and probably most important exception to this generalization being the work of William Labov (1994, 2001, 2010).

5.2 On the Formal Theoretical² Investigation of Dialect Variation

Muysken's (2014) "brief history of 20th century linguistics" only mentions "de Saussure: structure, Chomsky: deep structure." Generative theory characteristically focuses on the meta-grammar, that is, the parameters, principles and constraints that are assumed to govern the way grammars work. This hypothetical meta-grammar is claimed to be biologically programmed, and to manifest itself in a necessarily abstract way as the common core of natural languages. The meta-grammar, usually referred to as Universal Grammar (UG), can be seen as a kind of decision tree, whose nodes are called parameters. Every individual language can be uniquely defined as a specific constellation of choices made for the respective parameters. The decision tree representing UG is built in such a way that it reflects an essential trait of natural language, namely modularity, that is, the fact that the different parts of grammar (syntax, phonology, etc.), though interrelated, are internally autonomous to a certain degree. Certain instances of linguistic change can be a side effect of the modular organization of language, which may make it possible for abstract principles to interact.

Not all adherents of formal theory seem to be sufficiently aware of the fact that "heterogeneity and variation are not abnormalities but part of the normal condition of language" (Kiparsky 1988, 370), although there is a growing interest in geographical (inter-systemic) and quantitative (intra-systemic) variation among formal linguistic theorists. There is, however, no standard view of language variation in formal theoretical models. Several generativists have tried to understand the smallest differences between dialects as manifestations of universal principles underlying the organization of language systems. The smallest difference (at the level of language as a system shared by the members of a community; cf. Chomsky's (1995) *E-language*) is thus explained on the basis of the highest common denominator (*I-language*, language as a cognitive commodity). Dialect features are thus sometimes explained as different instantiations of language universals or as instantiations of different language universals. But how does it work concretely?

From the 1980s onward, generative syntax was dominated by Principles and Parameters (P&P) theory, which looked at UG as an invariant system of abstract principles, some of which permit at most a specified degree of variation within a given language. Originally, this notion of variation referred to differences between languages (macro-parametric variation), but the approach came to be applied to cross-dialectal variation (micro-parametric variation). From this line of research, deeper insights into the universal set of parameters were expected, in terms of their form as well as the substantial variation they allow.

Whereas in P&P variation resides in the computational system, in the Minimalist theory of generative syntax variation is located in the lexicon: all relevant parameters are encoded in the feature specification of individual items in the lexicon. Barbiers sketches a specific model to account for syntactic micro-variation, which "avoids the tendency found in much generative work to explain syntactic variation by syntactic principles exclusively" (2013, 24). In Barbiers' model there is a role for cognition, body ("brain, oral tract, etc.") and society ("groups, contact, history, etc."). This is in line with the Minimalist view, according to which variation is not only located in the syntactic module, but also in other linguistic and non-linguistic dimensions. Barbiers' model also in principle has room for frequency of usage and conventionalization, which in Barbiers' perception are relevant to the question of "why certain syntactic variables are sensitive to sociolinguistic specialization while others are not" (2013, 23).

While the concepts of rule and derivation had faded into the background in nonlinear phonological theory, they were entirely abandoned in Optimality Theory (OT). In OT, a set of constraints is assumed which determine the way in which the surface structure is allowed

to deviate from lexically-underlying representations. In principle, all constraints are both universal (although their relative importance is language-specific) and “soft,” as they can all be violated by conflicting constraints. The only generative capacity of the model resides in a function labeled GEN (“generator”), which projects an unlimited set of possible output candidates from a single lexical input form. All output candidates are rated according to their success in complying with the ranked constraints set; the candidate that best satisfies the relevant high-ranked constraints is selected as the optimal one. (It is never perfect, since it will always violate certain relevant lower-ranked constraints.) Constraints come in three kinds. Faithfulness constraints require the phonetic output form to be maximally identical with the underlying form, whereas markedness constraints make the phonetic output conform to prosodic and articulatory requirements. A third type of constraints guards the alignment of prosodic and grammatical structures.

In OT analyses, variation resides in the constraint ranking. Inter-systemic variation is commonly described through subtle differences in constraint ranking (e.g., Herrgen’s 2005 analyses of cross-dialectal variation in word-final [t] deletion in modern southern and southwestern dialects of German). In this “multiple grammars” scenario,³ every variety is viewed as a categorical grammar of its own.⁴ Intra-systemic variation is typically analyzed as partial ordering of constraints. The oldest variant of this type of analysis is based on the concept of floating constraints. This theoretical variant (represented in e.g., Nagy and Reynolds 1997) has meanwhile disappeared from the stage. The second variant is unranked constraints, the situation in which several constraints entirely coincide, producing multiple optimal candidates as output, although not every logically possible ranking emerging from this scenario will necessarily yield another winner (e.g., Anttila 1997, 2002). A closely related way to capture intra-systemic variation is by associating a relevant domain of the language to a different constraint ranking, a concept known as “co-phonology.” The third variant is essentially different from the other two. In the case of continuous ranking (Zubritskaya 1997) and stochastic OT (Boersma and Hayes 2001), the constraint ordering is not discrete but normally distributed, and the tails in the distributions of neighboring constraints overlap, which accounts for the variation (see Figure 5.1). The range of the intervals between neighboring constraints can vary.

Harmonic Grammar (HG) is a family of models in which constraints, which can be language-specific, are assigned a specific weight, rather than an absolute position in the constraint ranking (Smolensky and Legendre 2006; Pater 2008, 2014); lower-ranked constraints can ‘join forces’ to match the violation of higher-ranked constraints. Within HG the same range of theoretical variants apply as in OT: a stochastic HG model is discussed in Boersma and Pater (2016), for example.

A mechanism known as Lexicon Optimization guarantees that input forms do not necessarily need to be identical in different dialects. Nevertheless, unlike Minimalist syntax, formal phonology does not operate on the basis of the assumption that variation resides in the lexicon. In all mainstream theories of phonology, variation is located in the computational system, whereas in OT and related declarative models it lies in the constraint ranking.

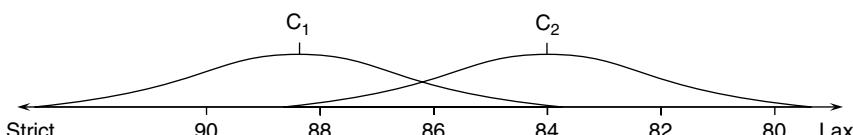


Figure 5.1 Constraint ranking on a continuous scale with stochastic evaluation (adapted from Clark 2005, 213).

Generative studies of language variation are often based on data from existing grammatical descriptions, an approach which is known as the “armchair method.” Alternatively, and even more in line with generative research traditions, the data consist of intuitions concerning the well-formedness of linguistic output, the source of which is typically a single native speaker (commonly the researcher, a methodological decision which is understandable given the fact that syntactic phenomena in particular tend to occur infrequently in spontaneous speech). This is what is referred to as “internal evidence.” In this view, “external” evidence concerns data from actual language use, and diachronic data—as well as data from geographical, social and stylistic variation—are relegated to the domain of E-linguistics. In generative studies of language variation, explanations are preferably I-linguistic in nature, and they are never historical/diachronic.

5.2.1 Weaknesses

Markey (1986) distinguished three “explanatorily adequate basic epistemologies” of language change:

- I. People (as communicative beasts) do things to language systems [...]
- II. Languages *qua* systems do things to people (The Whorf-Sapir Hypotheses).
- III. Languages *qua* systems have a life of their own apart from people.

Markey (1986, 16–17)

The first perspective can be termed the materialistic position, as opposed to position II, which is idealistic in nature. This latter position may make people deliberately change aspects of their language use (e.g., avoiding sexist language use) with the aim and conviction that doing so may cause others to change their ideologies. Position III, finally, can be labeled the biologicistic view. Language variation and change are in part social phenomena. Hence, a type I epistemology is called for. Formal theories largely file under the type III epistemology, and they are claimed to be psychological theories. Generative theory conceives of language as a cognitive object. It operates in a fundamentally different dimension than the (socio-)dialectological approach to language variation, since “the central dogma of sociolinguistics is that the community is prior to the individual. [...] language is seen as an abstract pattern located in the speech community and exterior to the individual” (Labov 2010, 7). The ultimate object of sociolinguistic research is the community grammar; a speech community is defined as a group of people who share the same set of linguistic norms. Apart from well-formedness judgements (as data), norms do not play a role in formal theory, let alone sociolinguistic prestige.

Until recently there were only few theoretical variation studies in which the position of the language variety under scrutiny seemed to matter. In this respect the world seems to be changing, since some researchers who approach language variation from a formal theoretical perspective have come to realize that they abstract away from the fact that dialects “change fast, due to rapid changes in their sociolinguistic context—factors involved include but are not limited to the lack of codification, the erosion of the dialect landscape due to social changes, and the low status of dialect varieties” (Bennis and van Oostendorp, 2013, 676).

The methodological construct of the “ideal speaker/hearer” is diametrically opposed to the concept of “inherent variability,” the insight that variation is a deep property of (living) language. Moreover, most dialect speakers have (at least receptive) knowledge of related varieties at their disposal. As King (2013, 448) points out, “it is often difficult to tease out what sort of knowledge the speaker is drawing on. Is knowledge of the dialect in question entirely separate from knowledge of more prestigious varieties?”

Generative theory is strictly synchronic. That is not incompatible with language variation as such, but it does not align with variation as a synchronic reflection of an ongoing process of language change.

Formalists often call dialectological and sociolinguistic research “descriptive.” In generativist rhetoric this term is often used as an invective, whereby studies that do not address any formal-theoretical questions are dismissed. In Chomsky’s philosophy, explanatory adequacy is determined through “evaluation metrics,” including economy and simplicity, both of which are relevant to language acquisition. In this regard, however, language change is out of the picture, since it is the diachronic issue *par excellence*.

Formal theory is deeply committed to categorical reasoning. “A probabilistic approach,” King (2013, 452) argues, “would go against the basic assumption [...] that grammar and use are modularly distinct.” Gradience is problematic and relegated to phonetics and diachrony. Linguistic systems are conceived as closed systems without, as it were, “frayed hems.” The restrictive operationalization of the concept of dialect variation as cross-dialectal variation is also problematic. Like all linguistic phenomena, cross-dialectal variation is in principle approached as the presence or absence of a discrete feature. The only refinement concerns potential differences in the structural conditioning of the feature at issue in the dialects concerned.

There is also an asymmetry in accessibility: most “theory-driven” studies are only accessible to those who are sufficiently familiar with the theoretical matrix. Given the expiration date of many generative proposals, it is by no means self-evident that acquiring the necessary familiarity is worthwhile. By contrast, most studies of a specific phenomenon in a specific language area that are not primarily theory-driven are accessible to every linguist who is sufficiently familiar with traditional linguistic and phonetic terminologies and classifications (although these are being gnawed at in modern typological research, as in e.g., Haspelmath, 2007). The obvious bridge across this divide is the construction of generally accessible databases of linguistic data; a sustainable approach of this sort enables “computer-aided armchair linguistics” (Fillmore, 1992). More on this recent development, in which many theoreticians actively participate, is given below.

5.3 Why Should They Collaborate?

Like Bacon’s (1561-1626) empiricist “men of experiment,” who collect data without altering them, dialectologists tend to be at best theoretically shortsighted. The generativist “reasoners” weave theoretical webs out of their own material; many of them are methodologically challenged, especially data-wise. Yet together they can make headway. Together they can gather the best material and convert it into something that is superior to the original material.

There are several ways in which the study of (both inter- and intra-systemic) dialect variation can profit from the types of theories that have been developed in formal linguistics. The following considerations can be made with regard to specific dialect features:

- the theory can inform the decisions underlying the selection of dialect features to be studied, although the selection will typically not solely be based on considerations of a strictly linguistic nature;
- in-depth structural analysis, whether or not in a formalized fashion, can be indispensable when deciding whether or not a phenomenon constitutes a case of quantitative variation. Variation in itself is usually linguistically structured, and is necessarily part of the larger structure of the linguistic system. In these respects, formal theory can offer analytical

depth—as Rooryck (2014) puts it, “theoretical linguistics renews the toolbox of descriptivists”—although dialect research usually has more to offer than mere description. With respect to theory as a tool or heuristic device, Labov (1997, 146–147) ponders, “The chief value of formal models, I believe, is to draw the attention of empirical investigators to undetermined relationships and unanswered questions that they may have overlooked. Once such questions have been raised, and clearly formulated, the chief purpose of the model has been achieved. It may then fruitfully be dissolved and replaced by other models, which will reveal new aspects to be investigated. The cumulative character of the enterprise lies not in the models, but in the gradual development of our knowledge through further inference and investigation.”⁵ Linguistic analysis can counteract the “atomistic” approach to dialect features that is typical of dialectology in that its practitioners have had a tendency “to treat linguistic forms in isolation rather than as parts of systems or structures” (Chambers and Trudgill, 1998, 33). Linguistic analysis, indeed, can help to embed the features in the structure of the grammar at large, by specifying the conditions on their distribution, linking the features at hand to related phenomena, and so on.

- sometimes theory can help elucidate the *raison d'être* of a specific dialect feature, that is, of why there is variation at this specific point in grammar (insofar as it was not borrowed from another dialect or language). It can, in other words, help to reveal the “innere Kausalität” (“inner causality”; Moulton 1961) of a phenomenon, and thus help to answer questions regarding the actuation and constraints, the transition, and the embedding of a change in the linguistic structure (Weinreich, Labov, and Herzog 1968; Labov 1972).

With respect to the study of changes in the usage of a dialect feature:

- an explanation of the phenomenon at issue may be the basis for predictions about possible future changes (say, the fate of a given dialect feature in the course of processes of structural dialect loss), provided that the account is grounded in a general theory;
- formal theory can sometimes also provide interpretable insights into the structural relationships between different features (elements, structures, or processes) of a dialect. The nature of these relationships is part of what distinguishes language varieties from each other, and they can play a role in processes of linguistic change. For example, one of the features which set apart the Ripuarian dialects of Dutch (spoken in the extreme southeastern part of the language area) is dorsal fricative deletion (DFD). In lexical morphemes with rhymes consisting of a short vowel followed by a dorsal fricative + /t/, the fricative can be deleted. As a result of compensatory vowel lengthening, non-low vowels develop a schwa offglide. Examples are:

(1) nax(t)	~	na:t (st. Dutch)	nacht	“night”
(2) ze:ət			zegt	“says”

(2) shows that DFD applies to inflected forms, but it does not apply to derived forms, as in:

(3) jəwɪç(t)	~	jəwɪ:ət	gewicht	“weight”
jəwɪçtiç	~	*jəwɪ:ətiç	gewichtig	“weighty”

The deletion of word-final [t] (WFtD) in clusters is one of the most widespread non-standard phenomena in the Dutch language area. In Ripuarian dialects of Dutch, variable

WFtD is fully productive following obstruents, and it can affect every word-final [t], that is, /t/ and /d/ (Hinskens 1992, 244–248). Examples are:

(4a)	rəsəpt	~	rəsep	<i>recept</i>	"prescription"
	rəsəptə	~	*rəsəpə		
	hoft	~	hof	<i>hoofd</i>	"headmaster"
	hofdə	~	*hovə		
	ɛçt	~	ɛç	<i>echt</i>	"real"
	ɛçtə	~	*ɛçə		
(4b)	wirəkt	~	wirək	<i>werkt</i>	"works"
	jewirəkt	~	jewirək	<i>gewerkt</i>	"worked" (past part.)
(4c)	ə le:f kriŋk			<i>een lief kind</i>	"a sweet child"
	ə le:ft	~	ə le:f		"a sweet [one]"

WFtD applies to [t], which is part of a lexical representation (4a), as well as to affixal [t] (4b). (4c) contains a word-final [t] in a grammatical function that is lacking in standard and other dialects of Dutch; here [t] is as pronominal affix marking neuter nouns that are not lexically expressed (Hinskens and Muysken 1986; Hinskens 1992, 180–181). This [t], too, can fall prey to WFtD. (4b, c) show that WFtD is blind to morphological structure, which is one of the properties of post-lexical processes.

In view of the coexistence of DFD and WFtD in Ripuarian dialects of Dutch, we are presented with the question of what happens to Ripuarian variants of items such as *nacht* "night" and *licht* "light," which in principle constitute input for both phenomena. DFD and WFtD are disjunctive: they cannot simultaneously apply on the same word. In words with this structure, DFD and WFtD bleed each other (Koutsoudas, Sanders, and Noll 1974), that is, one destroys the input of the other. Traditionally in Ripuarian dialects, words of this type systematically show DFD; both processes apparently apply in accordance with the "Elsewhere condition" (Kiparsky 1973). This condition says that, whenever a given form obeys the structural description of two different rules, the more specific rule applies. In such cases, the more general rule is blocked, but it does apply elsewhere. Indeed, the structural description of DFD, viz. /Vçt/, forms a proper subset of that of WFtD, viz. [[-son](t/d)]_{phw}. In Ripuarian dialects, the latter process applies exceptionlessly elsewhere. Moreover, as a lexical rule, DFD will apply before post-lexical WFtD.

If processes of dialect leveling proceed gradually in linguistic respects, then this analysis would lead to the prediction that DFD, which is conditioned on several levels, will be given up in favor of WFtD, which is automatic and exceptionless, and therefore, easier to acquire, and hence more resistant. For the dialects at issue this amounts to the prediction that variants such as [na:t], [li:at], "night," "light," will gradually be replaced by variants such as [nax], [liç].

Recordings of both elicited and conversational dialect use were made using a stratified random sample of 27 male speakers of the Ripuarian dialect of Rimburg; the speakers represent three different age groups. For DFD each of these speakers realized, *inter alia*, four words (a) in which in the Ripuarian dialects of Dutch form part of the potentially DFD-susceptible set, (b) in which /Vçt/ is in absolute final position, and (c) in the realization of which no other dialect features play a role, such that linguistic covariation effects are excluded. In the realization of these four words, WFtD is in principle possible. Each individual realization can be characterized as (DFD, WFtD). (1,1), that is, application of both DFD and WFtD is impossible. With respect to the three possible realization types: (1,0) is a realization of the type [na:t], whereas (0,1) is a realization of the type [nax]. (0,0), the realization type in which neither the dorsal fricative nor the final [t] are deleted, did not occur in the recorded data. This part of the database hence consists of 108 observations (4 words × 27 speakers); 4 realizations are lacking. The distribution of the remaining 104 realizations is summarized in Table 5.1.

Table 5.1 Findings: loss of DFD, territorial gain for WFtD.

(DFD, WFtD)	Older	Middle	Younger	Total
(1,0)	35	30	31	96
(0,1)	0	3	5	8
Total	35	33	36	104

There is an apparent-time decrease in the use of DFD variants, a decrease that is completely absorbed by WFtD. According to the outcomes of a significance test ($\chi^2=4.95$; $df=2$; $.05 < p < .10$), the data show tendential support for the hypothesis that DFD undergoes dialect loss but WFtD does not. On the contrary, WFtD gains from DFD's loss of ground.

Linguistic analysis is indispensable when it comes to answering the question of whether, and to what extent, cross-dialectal similarities in processes of language change are motivated either by (universal) structural tendencies or rather by common "external," for example, sociolinguistic, factors. The study of specific dialect features can, in turn, deepen formal theory in several ways. Since all language varieties must conform to universal theories of grammar, the patterning of specific instances of variation can serve as a test for formal theories.⁶ With the exception of language acquisition research, there hardly exists a tradition in linguistics of testing theories on the basis of behavioral data. Perhaps the best-known example in sociolinguistics is Guy's (1991) study, in which he puts his exponential model of t/d-deletion, and through this model some central tenets of Lexical Phonology, to the test. Properties of specific dialect features can disprove, modify or corroborate a proposed analysis. As an example, in his *Generatieve fonologie van het Nederlands* [Generative phonology of Dutch] from 1981, Geert Booij makes specific claims regarding vowel reduction and (re)syllabification. One of Booij's claims is given in (5), as follows:

- (5) 'The application of vowel reduction may result in resyllabification of the word undergoing this rule. In e.g. "pastoor" ["priest", with the stress on the second syllable – FH], the [s] is ambisyllabic, because it is preceded by a short vowel. But after reduction the [s] is preceded by a schwa, as a result of which [s] now belongs to the second syllable only.' (Booij, 1981, 151 – my translation, FH).⁷

The ambisyllabicity part of this claim is supported by the observation that (except for in a few interjections) lax vowels never appear in open syllables in Dutch. However plausible this might be, standard Dutch does not provide any substantive evidence for the second part of Booij's claim, although dialectology can help. A phenomenon distinguishing most Limburg and Ripuarian dialects from other varieties of Dutch is the palatal realization of /s/ and /z/ in absolute syllable-initial position before a consonant. An informal representation of this correspondence rule is given in (6):

(6) standard language		dialect
*[s]C	~	*[ʃ]C
*[z]C	~	*[ʒ]C

The examples in (10) serve to illustrate this.

(7) standard language		dialect	
[s]mal		[ʃ]ma'l	"tight, narrow"
[z]wart		[ʒ]wat	"black"

So in the dialect of Dutch spoken in Rimburg, for instance, one never finds:

- (8a) * bə'sty:əR "board of directors,"

which is identical to the standard variant. Because of (6), the dialect has

- (8b) bə'fty:əR

Similarly:

- (8c) * yə'stiç yə'ftriç "mental home"

However, in coda position following a vowel, an "etymological" /s/ remains [s], for example, in

- (9) * 'me'ʃtəR 'me'stəR "master," "teacher"
 * 'pəʃta 'pasta "paste," "pasta"

In the case of *pastoor*, palatalization is excluded in the variant with the unreduced vowel, thus

- (10a) * pa'ʃtu:əR pa'stu:əR
 pə'ʃtu:əR pə'stu:əR

and, likewise, in the variants of the toponym *Maastricht* (with stress, again, on the second syllable):

- (10b) * ma'ʃtriç ma'striç
 mə'ʃtre'ç mə'stre'ç

The fact that /s/ can be palatalized when the preceding vowel is reduced implies that in these dialects it is no longer part of the first syllable. At the same time, these data imply that in the variants with the full, unreduced vowels the [s] does not belong exclusively to the second syllable. The possibility that in the latter case it is only part of the first syllable is, however, not excluded as far as these dialects are concerned. In this case, dialect data can be adduced in favor of a theoretical claim.

The study of (apparent time or diachronic) changes in the usage of a given dialect feature can adduce evidence *pro* or *contra* specific theoretical proposals. For instance, processes of chain shifting (as are e.g., taking place in North American English and Austrian German)⁸ are problematic for OT, because they involve opacity. Opacity is troublesome for OT, which is basically a "flat," non-derivational model. In OT the various phases in a chain shift, which present themselves as many different sound changes, would not apply serially but in parallel, with the effect that all relevant vowels would be realized identically, viz. as the output of the last change in the chain. Within OT, several proposals have been advanced to solve these problems, but some, for example, "Sympathy" (McCarthy 1999), did not survive and the remaining ones—including "Turbidity," which says that the derivation of a segment is traceable in its representation—(Goldrick 2001) are disputed.

Genuine synergy between dialectology and formal linguistic theory is visible only rarely, but there is at least one such case in the recent history of the study of the trajectory of sound change. With respect to the intensive spread of linguistic change, the so-called "Neogrammarian controversy," that is, the distinction between Neogrammarian sound change (which is phonetically gradual, lexically abrupt, and exceptionless) and lexically

diffuse sound change (which is phonetically abrupt and lexically gradual, and hence not exceptionless; Scheutz 1988, 1608), has been the subject of a particularly fruitful exchange of ideas between Labov (1994) and Kiparsky (1995).

5.4 How Can They Collaborate?

Modern dialectology positions variation along social or social-geographical dimensions, the methodology is usually quantitative, and the phenomena and their conditioning are perceived to be probabilistic. In formal theory, by contrast, the focus is rather on psychology, the method is inductive, and languages are conceived deterministically as rule-governed systems. In view of the “yin and yang” nature of language (Bailey 1982), that is, the fact that language has both neuro-biological and socio-communicative sides, either perspective is legitimate and at the same time reductionistic, and many members of the two camps are aware of this. Referring to “autonomous phonology” as “any approach to phonological investigation that assumes that the object of phonological enquiry can be studied in its own right, relatively independently of the study of factors such as the social context in which the speakers are located,” Carr (2000, 74, 84) claims that “a fully autonomous phonology is unsustainable, since the data to be accounted for cannot be divorced from social context and are inherently variable.” The latter aspects are, to a large degree, matters of convention, and there is a non-obvious “relation between sociophonetic variation and UG as a natural object” (2000, 96). What role, then, does the “natural” play in the “conventional?”

The answer to this question calls for both a theoretically informed approach and more refined analyses of better data than generative theory traditionally builds upon. The types of *data* need to be broadened, even in those cases in which the empirical basis is confined to “internal” evidence. According to Lloret (1997, 201), regular—that is, systematic and recurrent—data “are the usual concern of formal linguistics, but the marginal [non-systematic but recurrent] data also hide relevant facts of the language and thus should be the explicit concern of formal linguistics too.” Lexically diffuse sound change typically results in that kind of data.

Obviously, sharing (rich) data can bring together dialectologists and theoreticians. Ideally, digital fieldwork recordings of relatively natural speech are stored in sound archives that are preferably parts of databases, which are connected with similar databases for related dialects,⁹ enriching all databases with standardized labeling of metadata¹⁰ will enhance research into larger geographical domains.¹¹ Once other, mainly older and comparable, and typically questionnaire-based data have also been made electronically available, cartographical tools, adapted so as to enable the mapping of both older and newer data,¹² will facilitate genuinely diachronic research.

Simultaneously, the types of *analysis* need to be refined. In formal theory the language use of non-linguists has meanwhile been discovered, which has improved the reliability and the generalizability of the findings. In less than two decades, micro-variation has developed into an important object of theoretical research. Now variation in the social dimension and (in the wake of the work of pioneers such as Lightfoot, Kroch, and Kiparsky) in the time dimension will have to find its way into theoretically-informed analyses.

Both for inter- and intra-systemic variation, ongoing processes of dialect leveling have far-reaching consequences regarding the question of the level at which the data can be analyzed. It is now time to abandon the position that a given phenomenon can be treated as a discrete variable, and instead to approach every non-standard phenomenon as a continuous variable. Researchers need to realize that dialect reality has become a fragmented reality; this complicates research, but approached from this angle, the findings will be more valid. And generative theory is in principle compatible with contextualized data, quantitative methods, and probabilistic relations.

Theories and models should not be expected to be able to account for 100% of the variance in the data, although the overall model should of course be as parsimonious as possible. As there are often competing internal tendencies, simple yet highly predictive accounts of dialect variation are probably an illusion. The claims deduced from formal theory should be considered as probabilistic explanations (cf. Kiparsky 1972, 222) or as predictions regarding favoring or disfavoring constraints.

5.5 How Have They Collaborated so Far? And What Has It Achieved?

For years, the Competing Grammars model introduced by Kroch (1989) has been a leading generative approach to (morpho-)syntactic variation. In this model, which is closely related to the “multiple grammars” scenario of early OT approaches to phonological variation (cf. section 2 above), on an abstract level, there is no such thing as intra-systemic variation. All variation results from the availability of two categorical grammars that differ in the setting of one particular parameter.

A selection of thorough studies inspired by formal theory and focused on *syntactic* variation in different languages is presented in Cornips and Corrigan (2005), a volume in which P&P theory is well represented. Yang (2003) demonstrates an elaboration of P&P theory which is explicitly constructed to deal with quantitative variation.

In generative syntactic approaches to micro-variation, the accent has meanwhile shifted to Minimalism (see above). There even exists a specialized international journal that is devoted to the Minimalist-informed study of (micro- and macro-) variation, entitled *Linguistic Variation*, now in its fifteenth edition, and edited by Jeroen van Craenenbroeck.

Van Craenenbroeck (2014) calculates the differences in geographical spread between linguistic variables pertaining to verb cluster orders in the SAND data (see below); for each pair of verb cluster orders, a distance is calculated based on their geographical spread. On the resulting distance matrix, Van Craenenbroeck carries out a Multiple Correspondence Analysis. On the basis of the resulting dimensions, the author tests a set of syntactic micro-parameters, which had been proposed in the theoretical literature on verb clusters. It turns out that there is a clear empirical basis for several of these parameters.

Elaborating on earlier work with Smith (Adger and Smith 2005), Adger (2006) presents a model based on Minimalist syntax and Distributed Morphology in which the interpretation and checking of “uninterpretable” (i.e., purely grammatical) features allow the quantitative modeling of the patterning of instances of morpho-syntactic variation, which fits empirically established distributions astonishingly well. This approach has been refined by Nevins and Parrott (2010), who claim that their proposal is compatible with the original variable rule model. Taking an alternative tack, Bresnan (2002) has proposed OT models in which low-ranked and thus inactive constraints nevertheless play a role in shaping quantitative morpho-syntactic variation.

For the study of *phonological* variation, Guy (2011) is a concise yet well-documented and highly readable historiography, which also summarizes recent work. Kostakis (2010) proposes a model based on “classical” OT in which language change is a matter of constraint demotion to account for intra-systemic variation. To this end, he introduces the concept of “Vestige constraint,” a phantom of a demoted constraint that behaves in an output-output fashion “as a sort of a receptacle for variants uttered in a linguistic community” (2010, 2476). Cardoso (2007) models developmental data in a Stochastic OT framework, whereas Pater (2014) presents a HG analysis of Canadian Raising; his model includes language-specific constraints. The Harmonic-Grammar weight associated with a given constraint is increased for every candidate where it is satisfied. The candidate with the highest sum (“Harmony”)

wins. The impact of HG and stochastic OT seems to be growing. These frameworks serve to model both language acquisition and intra-systemic variation, typically in tandem.

In the meantime, especially in the study of phonological variation, the paradigm debate concerning cognitivist models (Usage-based Phonology, Exemplar Theory, etc.; see below) is becoming more and more important, making the theoretical part of the field richer and deeper.¹³

5.5.1 Net Results and Added Value

In general terms: what has the collaboration achieved so far? And what does the net balance look like?

On the negative side, the marriage between (socio-)dialectology and formal theory does not seem to be an utterly happy one, not because they still do not know each other very well—that may not necessarily stand in the way of conjugal bliss—but because both partners distrust each other and spend little time together.

On the positive side, the past 15 years have seen a sharp growth of theoretically inspired large-scale dialect geography projects. The digitization of linguistic research doubtlessly plays an important role, as does the fact that powerful servers and fast internet connections make the data broadly accessible to other researchers. The digitization consists of a chain of major technical improvements that catalyze the speed and quality of the collection and analysis of large amounts of data.

In theoretical syntax circles, too, it has become “clear that informally gathered intuitions are not always a satisfactory basis for syntactic theorising. It is also clear that experimental methods are sometimes necessary and may provide richer data than informal methods. It is clear too that corpus data can be valuable in various ways. Above all it is clear that questions about data are more important than is sometimes assumed” (Borsley 2005, 1479). This quote is from Borsley’s introduction to a thematic issue of *Lingua* (vol. 115(1), pp. 1475–1666) entitled “Data in Theoretical Linguistics.” In the eight contributions, themes such as gradience, soft constraints, and magnitude estimation are discussed. There are further encouraging signs of this type, such as Penke and Rosenbach’s (2007) collection of papers (*cum* extensive discussion) on types of evidence and argumentation in (morpho-)syntax and, for phonology, the overview of actually used and potentially relevant data types in van Oostendorp (2013). Most of these texts contain pleas for a diversification in the types and sources of data studied for theoretical purposes, which points at a general contemplation among theoreticians of the empirical basis of their work.

In phonological theory the insights have emerged that variation in the sound component of language has a range of possible loci, can accordingly be constrained in different ways, and thus cannot be approached in a standard way (Hinskens 1998). Generally, the degree of awareness and consequently the degree of manipulability of a given phenomenon increases in accordance with the following cline: phonetic implementation < postlexical processes < lexical phonological rules < lexicalized sound change. The productivity of sound changes decreases along the same cline. Differences in the position of a sound change on this cline can be conceived as a type of gradience, and can cause inter-systemic variation beneath the surface (Ramsammy 2015). Similarly, syntactic variation has a wide range of potential “origins” (UG, psychology, physiology, society), as Barbiers (2013) argues. Here, the degree of awareness and thus the degree of manipulability of a phenomenon generally increase in accordance with the following cline: realization of syntactic structures (e.g., doubling of Wh-elements in syntactic dependencies) < morphosyntax (e.g., verbal inflection) < lexicalized syntactic phrases (such as verbs and pronouns).

From the cross-linguistic comparison of findings from sociolinguistic research it appears that variable phenomena which occur in different languages are often influenced in the same way by the same or similar linguistic factors (cf. Tagliamonte 2011). This implies that the

internal conditioning of language variation can give important indications of possible universal constraints. A closely related insight is what Bresnan *et al.* (2001) label “stochastic generalization,” that is, generalizations that are categorical for some language(s) but probabilistic in others. Sometimes the contextual conditioning of a variable phenomenon goes in the direction of a complementary distribution; it is easy to imagine that, say, allophony can be the end result of a gradual diachronic development, in which the preference of a variant for one type of context (e.g., more WFtD before a consonant) and the dispreference for a complementary context type (less WFtD before a vowel) on either side grew into a categorical distribution (no [t/d] before C, always [t/d] before V, an alternation that is comparable in some respects with one of the few aspects of French liaison that are fully understood). Teleologies of this type cannot easily be made visible in a strictly synchronic approach, but can be made so using an apparent-time approach, and more definitely still on the basis of real-time replications. At the same time, stochastic generalizations are another reason to include intra-systemic variation in the grammar; cf. “[i]f the canonical ordering and the obligatory cases are part of the competence grammar, but the quantitative preference is treated as performance, then a larger generalization is lost” (Wasow 2002, 139).

5.6 The Roads Ahead

The last few decades have seen the rapid development and spread of “cognitivist” approaches to language, including Cognitive Grammar (Goldberg 2006) and, for the sound components, Exemplar Theory (ET; Johnson 1997) and the closely related Usage-Based Phonology (Bybee 2001). In these *paradigms*, which are conceived as alternatives to generative theory, lexical items and their properties (regarding form, function and usage, and including all sorts of type and token frequencies) have a pivotal position. Each realization (or “exemplar”) of an item, with all its phonetic, semantic, and extra-linguistic attributes, is supposed to be stored in memory, where it is connected with other items and their many properties. From this huge multi-dimensional memory cloud, grammar emerges from the bottom up. The geographical and social distribution of the tokens (taking the place of variants) are part of the stored extra-linguistic properties.

In the cognitivist paradigm, corpora play a central role, not least as the source of lexical frequencies. For the study of dialect variation, cognitivist approaches have attractive sides. First, variability is assumed to be represented directly in memory in the shape of concrete exemplars, which are assumed also to contain social-indexical information (e.g., Docherty and Foulkes 2000). Second, the model is not based upon deterministic principles, but rather probabilistic ones; as such, it seems to match the nature of most documented instances of language variation. Third, just like adherents of cognitivist approaches to language, many sociolinguists studying language variation reject the analytical distinction between diachrony and synchrony that is applied by adherents of formal theories. Cognitivist approaches can be implemented relatively straightforwardly for the study of dynamic aspects of language such as acquisition and processes of language change.

Barbiers (2013) sketches a Minimalist model for syntactic variation that can in principle accommodate frequency effects. In generative phonology (including OT and HG), models that allow room for frequency effects are also being developed. For example, in connection with lexicalization, which is typically the last phase in the life cycle of a sound change, token frequency can be argued and demonstrated to play a central role (Hinskens 2011; Bermúdez-Otero 2012). With regard to phonological variation, several hybrid models uniting generativist/OT and ET approaches are being developed. Some of these are sketched in Hinskens *et al.* (2014, 13–14). In these domains, three-way traffic between dialectology, formal theory and cognitivist theory may even come into vogue; cf. Nagy (2013, 437–438).

For the study of sound change, Forced Alignment and automatic Vowel Extraction and measurement (FAVE; Labov *et al.* 2013) is very promising, because of the precision of the method

and the enormous time savings, which enable researchers to concentrate on questions of interpretation. Other than phonetics, new *methods of analysis* can also be imported from other disciplines, including from statistics. More advanced multivariate analyses such as path analysis, a technique to estimate both the direct effects of certain variables and the indirect effects of the same variables or others via intervening variables (as applied in Villena Ponsoda 2014), enable more careful modeling of the interplay between internal and extra-linguistic forces in processes of language change. This can add depth to any generative model of language variation.

There is a need for multi-dimensional mapping techniques, which would enable the production of maps on which static as well as dynamic data can be displayed simultaneously in geographical space, in several types of social/cultural space and in time (see Thun 2010). It would be interesting to apply multi-dimensional cartography to data relating to clusters of phenomena or even grammatical (sub-)modules, for examples, a constraint ranking (Sloos and van Oostendorp 2012), instead of specific phenomena. Such cartographical techniques can help to bring to light the gradual internal generalization of some structural phenomenon in the course of its diffusion, which may well be relevant to formal theory.

Last but definitely not least, theoretically enriched dialectological research could focus on sets of *questions*, for example, regarding aspects of the development of diglossic into diaglossic repertoires that is currently taking place all over Europe. To what extent is the development of continua between dialects and standard varieties (which the Marburg-based REDE project focuses on)¹⁴ socially motivated and supported by functional dialect loss, growing command and usage of the standard or near-standard varieties, and the like? How are those changes internally conditioned? Is it on the basis of grammatical similarities, drift tendencies, or markedness? These general questions can be addressed through answers to more specific ones such as "which types of dialect features are generally ousted 'on the way upwards,' and which remain?"

More light can probably be shed on the complex conditioning of language variation and thus on the relative roles of dialectological, formal and possibly cognitivist approaches from the perspective of stability. Which types of variable phenomena remain stable through time and space, and why? Is it for linguistic or extra-linguistic reasons, or both? If both, then how do linguistic and extra-linguistic forces interlock?

Acknowledgments

Many thanks to Sjef Barbiers, Hans Bennis, Edoardo Cavigani, Jeroen van Craenenbroeck, Peter Gilles, Ben Hermans, Pieter Muysken, and Johan Rooryck for their help and advice. They are not responsible for any flaws, nor do they necessarily agree with the views expressed here.

NOTES

1 The idea encapsulated by this phrase is Hugo Schuchardt's, in his famous *Über die Junggrammatiker, gegen die Lautgesetze* (1885), and in a short addendum to this, entitled *Worte als Individuen*; see Spitzer (1921). But there is nothing in Schuchardt's *œuvre* that is as concise and lapidary as Jaberg's famous dictum. Schuchardt's admirer Jules Gilliéron (Jaberg's teacher and an atlas man more than anything else; see Goebel, this volume) does not seem to have published anything that contains the idea, although it has sometimes been attributed to him.

- 2 Or generative. The designation “formal” is to be related to Saussure’s definition of *langue* as “une forme, non une substance” (“a form, not a substance”). With the term “forme,” Saussure referred to the structure of the relations holding between linguistic elements (Siertsema 1980, 196–198).
- 3 Closely related to the Competing Grammars model (Kroch 1989); see section 5 below.
- 4 Barbiers (2010, 127–138) contains a sketch of the developments in the past four decades of the history of the mainstream generative accounts of phonology and (morpho-)syntax, from rule-based to constraint-based approaches, both in general and with respect to language variation. Bennis and van Oostendorp (2013) garnish these general lines with summaries of some main studies for the Dutch language area. Alber (2014) is an attempt to model micro-variation with the aid of recent insights and methods from OT.
- 5 One could add that, whereas in Popper’s view the competition between theories is central, in practice much scientific work amounts to the extension of the domain of a theory or the interpretation of new facts in the light of an existing theory. This is what Kuhn (1962, 114, 156 ff) has labeled “normal science.” Most research in formal linguistic theory is accumulative Kuhnian “normal” science, rather than Popperian inductivistic, in nature.
- 6 The very existence of language variation as such is rather a problem for functionalist theories that look at language primarily as a medium for communication.
- 7 Cf. Kager (1999, 165, 306); Booij (1995, 131).
- 8 In both cases with considerable cross-dialectal differences; cf. Labov (2010) and Moosmüller and Scheutz (2013), respectively.
- 9 Such as *Sprekende Kaart*, lit. “Speaking Map,” at the Meertens Instituut website: http://www.meertens.knaw.nl/projecten/sprekende_kaart/svg/, accessed July 1, 2014.
- 10 For example IMDI, cf. <http://www.mpi.nl/IMDI/>.
- 11 As is already under way in *Mimore* and *Edisyn*; cf. Barbiers and Goeman 2013; Hinskens and van Oostendorp 2013, 73.
- 12 As per the Marburg-based *Digitaler Wenker-Atlas*, DiWA, accessible at <http://www.diwa.info>.
- 13 A content analysis of two international peer-reviewed journals can be found at <http://bit.ly/1MSaaIR> and <http://bit.ly/1VEtXkj>.
- 14 <http://www.regionalsprache.de/>.

REFERENCES

- Adger, David. 2006. “Combinatorial variation.” *Journal of Linguistics*, 42: 503–530.
- Adger, David, and Jennifer Smith. 2005. “Variation and the Minimalist programme.” In *Syntax and Variation: Reconciling the Biological and the Social*, edited by Leonie Cornips, and Karen Corrigan, 149–178. Amsterdam: Benjamins.
- Alber, Birgit. 2014. “Microvariation inside typological space.” Paper presented at Old World Conference in Phonology (OCP) 11, Leiden/Amsterdam, January 24th 2014.
- Anttila, Arto. 1997. “Deriving variation from grammar. In: *Variation, Change and Phonological Theory*, edited by Frans Hinskens, Roeland van Hout and Leo Wetzel. 35–68. Amsterdam, John Benjamins.
- Anttila, Arto. 2002. “Variation and Phonological Theory.” In: *Handbook of Language Variation and Change*, edited by Jack Chambers, Peter Trudgill, and Natalie Schilling-Estes. 206–243. Blackwell, Oxford, U.K., and Malden, Massachusetts.
- Bacon, Francis. 1620 [2004]. *Novum Organum*. In *The Instauratio Magna Part II: Novum Organum and Associated Texts*, edited by Graham Rees, and Maria Wakely, XXX-XXX. Oxford: Oxford University Press.
- Bailey, Charles-James. 1982. *On the Yin and Yang Nature of Language*. Ann Arbor, MI: Karoma.
- Barbiers, Sjef. 2010. “Language and space: Structuralist and generative approaches.” In *Language and Space: Theories and Methods*, edited by Peter Auer, and Jürgen Schmidt, 125–142. Berlin: de Gruyter.
- Barbiers, Sjef. 2013. “Where is syntactic variation?” In *Language Variation – European Perspectives IV*, edited by Peter Auer, Javier Caro Reina, and Göz Kaufmann, 1–26. Amsterdam: Benjamins.
- Barbiers, Sjef, and Ton Goeman. 2013. “Research results from on-line dialect

- databases and dynamic dialect maps." In *Language and Space: Dutch*, edited by Frans Hinskens, and Johan Taeldeman, 646–663. Berlin: de Gruyter.
- Bennis, Hans, and Marc van Oostendorp. 2013. "Grammar and geography or vice versa." In *Language and Space: Dutch*, edited by Frans Hinskens, and Johan Taeldeman, 664–679. Berlin: de Gruyter.
- Bermúdez-Otero, Ricardo. 2012. "The architecture of grammar and the division of labour in exponence." In *The Morphology and Phonology of Exponence*, edited by Jochen Trommer, 8–83. Oxford: Oxford University Press.
- Boersma, Paul, and Bruce Hayes. 2001. "Empirical tests of the gradual learning algorithm." *Linguistic Inquiry*, 32: 45–86.
- Boersma, Paul and Joe Pater. 2016. "Convergence properties of a gradual learning algorithm for Harmonic Grammar." In: *Harmonic Serialism and Harmonic Grammar*, edited by John McCarthy and Joe Pater. 389–434. Sheffield: Equinox.
- Booij, Geert. 1981. *Generatieve Fonologie van het Nederlands*. Utrecht: Spectrum.
- Booij, Geert. 1995. *The Phonology of Dutch*. Oxford: Oxford University Press.
- Borsley, Robert. 2005. "Introduction." In *Data in Theoretical Linguistics (special issue of Lingua, 115(11))*, edited by Robert Borsley, 1475–1480. Amsterdam: Elsevier.
- Bresnan, Joan, Shipra Dingare, and Christopher Manning. 2001. "Soft constraints mirror hard constraints: Voice and person in English and Lummi." In *Proceedings of the LFG 01 Conference, Hong Kong*, edited by Miriam Butt, and Tracy Holloway King. Available at: <http://web.stanford.edu/~bresnan/lfg01.pdf> [accessed 29th November 2015].
- Bresnan, Joan. 2002. "Pidgin genesis and optimality theory." In *Processes of Language Contact: Case Studies from Australia and the Pacific*, edited by Jeff Siegel, 145–173. Montreal: Fides.
- Bybee, Joan. 2001. *Phonology and Language Use*. Cambridge: Cambridge University Press.
- Cardoso, Walcir. 2007. "The variable development of English word-final stops by Brazilian Portuguese speakers: A stochastic optimality theoretic account." *Language Variation and Change*, 19(3): 219–248.
- Carr, Philip. 2000. "Scientific realism, sociophonetic variation, and innate endowments in phonology." In *Phonological Knowledge: Conceptual and Empirical Issues*, edited by Noel Burton-Roberts, Philip Carr, and Gerard Docherty, 67–104. Oxford: Oxford University Press.
- Chambers, Jack, and Peter Trudgill. 1998. *Dialectology*. 2nd ed. Cambridge Textbooks in Linguistics. Cambridge University Press.
- Chomsky, Noam. 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.
- Clark, Brady. 2005. "On stochastic grammar." *Language*, 81(1): 207–217.
- Cornips, Leonie, and Karen Corrigan, eds. 2005. *Syntax and Variation: Reconciling the Biological and the Social*. Amsterdam: Benjamins.
- Docherty, Gerard, and Paul Foulkes. 2000. "Speaker, speech, and knowledge of sounds." In *Phonological Knowledge: Conceptual and Empirical Issues*, edited by Noel Burton-Roberts, Philip Carr, and Gerard Docherty, 105–129. Oxford: Oxford University Press.
- Fillmore, Charles. 1992. "'Corpus linguistics' or 'Computer-aided armchair linguistics'". In *Directions in Corpus Linguistics*, edited by Jan Svartvik, 35–60. Berlin: de Gruyter.
- Goldberg, Adele. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.
- Goldrick, Matthew. 2001. "Turbid output representations and the unity of opacity." *Papers from the Annual Meeting of the North-Eastern Linguistics Society (NELS) 30(1)*: 231–245.
- Guy, Gregory. 1991. "Explanation in variable phonology: An exponential model of morphological constraints." *Language Variation and Change*, 3: 1–22.
- Guy, Gregory. 2011. "Sociolinguistics and formal linguistics." In *The Sage Handbook of Sociolinguistics*, edited by Ruth Wodak, Barbara Johnstone, and Paul Kerswill, 249–264. Los Angeles: Sage.
- Haspelmath, Martin. 2007. "Pre-established categories don't exist: Consequences for language description and typology." *Linguistic Typology*, 11(1): 119–132.
- Herrgen, Joachim. 2005. "Sprachgeographie und Optimalitätstheorie am Beispiel der t-Tilgung in Auslaut-Clustern des Deutschen." *Zeitschrift für Dialektologie und Linguistik*, 72(3): 278–317.
- Hinskens, Frans. 1992. *Dialect Levelling in Limburg: Structural and Sociolinguistic Aspects*. PhD thesis, University of Nijmegen [abridged and revised version published under same title by Niemeyer, Tübingen, 1996].
- Hinskens, Frans. 1998. "Variation studies in dialectology and three types of sound change." *Sociolinguistica*, 12: 155–193.

- Hinskens, Frans. 2011. "Lexicon, phonology and phonetics. Or: Rule-based and usage-based approaches to phonological variation." In *Linguistic Universals and Language Variation*, edited by Peter Siemund, 416–456. Berlin: de Gruyter.
- Hinskens, Frans, Ben Hermans, and Marc van Oostendorp. 2014. "Grammar or lexicon. Or: Grammar and lexicon? Rule-based and usage-based approaches to phonological variation." *Lingua*, 142: 1–26.
- Hinskens, Frans, and Pieter Muyken. 1986. "Formele en functionele benaderingen van dialectale variatie: De flexie van het adjektief in het dialect van Ubach over Worms." In *Syntax en Lexicon; Veertien Artikelen bij Gelegenheid van het Emeritaat van Albert Sassen*, edited by Cor Hoppenbrouwers, Ineke Schuurman, Ron van Zonneveld, and Frans Zwarts, 13–24. Dordrecht: Foris.
- Hinskens, Frans, and Marc van Oostendorp. 2013. "Language and space in Dutch: Wishes for the future." In *Language and Space: Dutch*, edited by Frans Hinskens, and J. Taeldeman, 60–80. Berlin: de Gruyter.
- Jaberg, Karl. 1908. *Sprachgeographie*. Aarau: Sauerländer.
- Johnson, Keith. 1997. "Speech perception without speaker normalization: An exemplar model." In *Talker Variability in Speech Processing*, edited by Keith Johnson, and John Mullennix, 145–165. San Diego, CA: Academic Press.
- Kager, René. 1999. *Optimality Theory*. Cambridge: Cambridge University Press.
- King, Ruth. 2013. "Morphosyntactic variation." In *The Oxford Handbook of Sociolinguistics*, edited by Robert Bayley, Richard Cameron, and Ceil Lucas, 445–463. Oxford: Oxford University Press.
- Kiparsky, Paul. 1972. "Explanation in phonology". In *Goals of Linguistic Theory*, edited by Stanley Peters, 189–227. Englewood Cliffs, NJ: Prentice Hall.
- Kiparsky, Paul. 1973. "'Elsewhere' in phonology." In *A Festschrift for Morris Halle*, edited by Stephen Anderson, and Paul Kiparsky, 93–106. New York: Holt, Rinehart, and Winston.
- Kiparsky, Paul. 1988. "Phonological change." In *Linguistics: The Cambridge Survey, vol. 1 – Linguistic Theory: Foundations*, edited by Frederick Newmeyer, 363–413. Cambridge: Cambridge University Press.
- Kiparsky, Paul. 1995. "The phonological basis of sound change." In *The Handbook of Phonological Theory*, edited by John Goldsmith, 640–670. Cambridge, MA: Blackwell.
- Kostakis, Andrew. 2010. "Vestige Theory: Sociolinguistic evidence for output–output constraints." *Lingua*, 120(10): 2476–2496.
- Koutsoudas, Andreas, Gerald Sanders, and Craig Noll. 1974. "On the application of phonological rules." *Language*, 50(1): 1–28.
- Kroch, Anthony. 1989. "Reflexes of grammar in patterns of linguistic change." *Language Variation and Change*, 1: 199–244.
- Kuhn, Thomas. 1962. "The structure of scientific revolutions." In *Foundations of the Unity of Science: Towards an International Encyclopedia of Unified Science* (vol. II, no. 2), edited by Otto Neurath, Rudolf Carnap, and Charles Morris, 53–272. Chicago: University of Chicago Press.
- Labov, William. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.
- Labov, William. 1994. *Principles of Linguistic Change, vol. 1: Internal Factors*. Oxford: Blackwell.
- Labov, William. 1997. "Resyllabification." In *Variation, Change and Phonological Theory*, edited by Frans Hinskens, Roeland van Hout, and Leo Wetzels, 145–179. Amsterdam: Benjamins.
- Labov, William. 2001. *Principles of Linguistic Change, vol. 2: Social Factors*. Oxford: Blackwell.
- Labov, William. 2010. *Principles of Linguistic Change, vol. 3: Cognitive and Cultural Factors*. Oxford: Blackwell.
- Labov, William, Ingrid Rosenfelder, and Josef Frühwald. 2013. "One hundred years of sound change in Philadelphia: Linear incrementation, reversal, and reanalysis." *Language*, 89(1): 30–65.
- Lloret, Maria-Rosa. 1997. "When does variability become relevant to linguistic theory?" In *Variation, Change and Phonological Theory*, edited by Frans Hinskens, Roeland van Hout, and Leo Wetzels, 181–206. Amsterdam: Benjamins.
- Markey, Thomas. 1986. "When minor is minor and major is major: Language expansion, contraction and death." Paper presented at the Third International Conference on Minority Languages, Galway, Ireland, 21–26 June 1986.
- McCarthy, John. 1999. "Sympathy and phonological opacity." *Phonology*, 16: 331–399.
- Moosmüller, Sylvia, and Hannes Scheutz. 2013. "Chain shifts revisited: The case of monophthongisation and E-merger in the city dialects of Salzburg and Vienna." In *Language Variation – European Perspectives IV*, edited by Peter Auer, Javier Caro Reina, and Göz Kaufmann, 173–186. Amsterdam: Benjamins.

- Moulton, William. 1961. "Lautwandel durch innere Kausalität: Die Ostschweizerische Vokalspaltung." *Zeitschrift für Mundartforschung*, 28: 227–251.
- Muysken, Pieter. 2014. "How can linguistics survive? The need for integrative approaches." Paper presented at the Linguistics in the Netherlands (LIN) conference, Utrecht, February 1st 2014.
- Nagy, Naomi. 2013. "Phonology and sociolinguistics." In *The Oxford Handbook of Sociolinguistics*, edited by Robert Bayley, Richard Cameron, and Ceil Lucas, 425–444. Oxford: Oxford University Press.
- Nagy, Naomi, and William Reynolds. 1997. "Optimality theory and variable word-final deletion in Faetar." *Language Variation and Change*, 9: 37–55.
- Nevins, Andrew, and Jeffrey Parrott. 2010. "Variable rules meet Impoverishment theory: Patterns of agreement leveling in English varieties." *Lingua*, 120: 1135–1159.
- Pater, Joe. 2008. "Gradual learning and convergence." *Linguistic Inquiry*, 39(2): 334–345.
- Pater, Joe. 2014. "Canadian raising with language-specific weighted constraints." *Language*, 90(1): 230–240.
- Penke, Martina, and Anette Rosenbach, eds. 2007. *What Counts as Evidence in Linguistics*. Amsterdam: Benjamins.
- Ramsammy, Michael. 2015. "The life cycle of phonological processes: Accounting for dialectal microtypologies." *Linguistics and Language Compass*, 9(1): 33–54.
- Rooryck, Johan. 2014. "Whither linguistics? Neither physics nor archaeology!" Paper presented at the Linguistics in the Netherlands (LIN) conference, Utrecht, February 1st 2014.
- Scheutz, Hannes. 1987. "Lautwandel." In *Sociolinguistics: An International Handbook of the Science of Language and Society*, vol. 2, edited by Ulrich Ammon, Norbert Dittmar, and Klaus Mattheier, 1603–1614. Berlin: de Gruyter.
- Schuchardt, Hugo. 1885. *Über die Lautgesetze: Gegen die Junggrammatiker*. Berlin: Robert Oppenheim.
- Siertsema, Berthe. 1980. "Wat is het strukturalisme?" In *Wetenschap en Taal III: Een Derde Reeks Benaderingen van het Verschijnsel Taal*, edited by Bernard Tervoort, 195–211. Muiderberg: Coutinho.
- Sloos, Marjoleine, and Marc van Oostendorp. 2012. "The relationship between phonological and geographic distance: Umlaut on the diminutive in Dutch dialects." *Taal en Tongval*, 62(2): 204–250.
- Smolensky, Paul, and Géraldine Legendre. 2006. *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar*. Cambridge, MA: MIT Press.
- Spitzer, Leo. 1921. *Hugo Schuchardt-Brevier: Ein Vademekum der Allgemeinen Sprachwissenschaft, als Festgabe zum 80. Geburtstag des Meisters* [reprinted 1976]. Tübingen: Niemeyer.
- Tagliamonte, Sali. 2011. "Variation as a window on universals." In *Linguistic Universals and Language Variation*, edited by Peter Siemund, 128–168. Berlin: de Gruyter.
- Thun, Harald. 2010. "Pluridimensional cartography." In *Language and Space: Language Mapping*, edited by Alfred Lameli, Roland Kehrein, and Stefan Rabanus, 506–524. Berlin: de Gruyter.
- Van Craenenbroeck, Jeroen. 2014. "The signal and the noise in Dutch verb clusters: A quantitative search for microparameters." Paper presented at *What Happened to Principles and Parameters?* workshop, Arezzo, Italy, 3–5 July 2014.
- Van Oostendorp, Marc. 2013. "A consumer guide to phonological evidence." *Nordlyd*, 40(1): 274–293.
- Villena Ponsoda, Juan. 2014. "Multi-levelled analysis of speech variation: Intervocalic /d/ in urban Malaga." CSLS Winter School on Language in Social Context, Kandersteg, Switzerland, 13–17 January 2014.
- Weinreich, Uriel, William Labov, and Marvin Herzog. 1968. "Empirical foundations for a theory of language change." In *Directions for Historical Linguistics*, edited by Winfred Lehmann, and Yakov Malkiel, 97–195. Austin, TX: University of Texas Press.
- Yang, Charles. 2003. *Knowledge and Learning in Natural Language*. Oxford: Oxford University Press.
- Wasow, Thomas. 2002. *Postverbal Behavior*. Stanford, CA: CSLI Publications.
- Zubritskaya, Katya. 1997. "Mechanism of sound change in optimality theory." *Language Variation and Change*, 9: 121–148.

6 Sociodialectology

TORE KRISTIANSEN

6.1 What Is Sociodialectology?

As the first sentence of his seminal 1972 book *Sociolinguistic Patterns*, which summarized his research in the 1960s and changed dialectology/linguistics forever, William Labov wrote “I have resisted the term sociolinguistics for many years, since it implies that there can be a successful linguistic theory or practice which is not social” (Labov 1972, xiii). Some years later, Jack Chambers and Peter Trudgill added a short final chapter to their important 1980 textbook, *Dialectology*, in which they discussed their own terminological dilemma. The issue concerned what to call the discipline, which, in their understanding and treatment, had incorporated the Labovian study of sociolinguistic patterns (“urban dialectology”) as a new stream that had joined in confluence with the traditional stream of dialect research (i.e., dialect geography). Chambers and Trudgill opted for the term *dialectology* (1980, 206), stating that they would use the term “to mean the study of language variety by any methodology” (1980, 17). The discipline’s unity, meanwhile, would be “provided by the theoretical underpinning of what is increasingly becoming known as ‘variation theory’” (*ibid.*). Chambers and Trudgill did consider “variation theory” as a name for the discipline, but rejected it as “too broad, since it carries no connotation about the focus of the discipline being social and spatial variation” (1980, 207). Against the backdrop of the initial concerns that were expressed by the leading figures of the swelling research stream which inundated dialectology from the 1960–1970s on, it may appear somewhat ironic that the term which has become the most commonly used to refer to this new dialectology—*variationist sociolinguistics*—includes within it both “socio-” and “variation.”

The term *sociodialectology*, which is used as the heading of this chapter, has hitherto been practically nonexistent in the literature. A pertinent question arises: if dialectology ought to be conceived of as a “social” scientific discipline, wouldn’t it be better to concur with Labov by resisting the prefixation of “socio-,” instead of embracing it? The near-identical term *social dialectology* has some currency, but is hardly used to cover anything different from “variationist sociolinguistics,” as far as I can judge from consulting a dictionary entry (Swann *et al.* 2004), a handbook chapter (Kerswill 2004), and the thus-entitled Festschrift for Peter Trudgill (Britain and Cheshire 2003). The introduction of yet another term can only be justified by offering a precise definition of how “sociodialectology” is different. The term *sociophonetics* is now frequent in the literature and is (dictionary-)defined as a discipline which “[i]nvolves the application of phonetics to sociolinguistic study” (Swann *et al.* 2004, 288). I think sociodialectology is to be defined the other way round, so to speak, as the application of

sociolinguistic study to dialectology, where dialectology is understood in its traditional sense of dialect geography. Thus, if we define “sociolinguistic study” as the scientific attempt to disentangle the complex interdependence of social, language-ideological, and linguistic processes of variation and change, the particularity of sociodialectology consists in submitting these processes to sociolinguistic study *in geographical space*. The focus on geographical space should be taken to imply that *dialect* refers to regional varieties, as per traditional terminology.

6.2 The Crucial Role of Technology

Labov (1994, 25) talks of the “natural alliance of dialect geography, sociolinguistics, phonetics, and historical linguistics,” and stresses that the possibility of establishing connections between them emerges with new technology. Indeed, it is hardly possible to overestimate the importance of new technology for the theoretical and methodological advances that developed from the 1960s onward. The potential for gathering, storing, and analyzing spoken data changed dramatically with the appearance and development of high-quality audio recording equipment and computer technology.

As a discipline that works with huge amounts of empirical data, it has always been a requisite of dialectology to construct convenient systems for storing and retrieving those data. A walk through the long central corridor of the author’s workplace, Copenhagen University’s Section of Dialectology (hence the “Danish topping” to this chapter), testifies glaringly to the immense significance of technology in this regard. At one end of the corridor, a large room contains archived records of Danish dialects covering the period 1750–1945, stored in rows of big cupboards housing innumerable drawers. These contain some 3 million paper notes organized according to both alphabetic and thematic principles, and constitute the database for the handful of editors of *Ømålsordbogen* (“Dictionary of the Insular Dialects,” where “insular” refers to Denmark’s eastern archipelago). At the corridor’s opposite end, we find a large room filled with busy computers and people working for the LANCHART (Language Change in Real Time) project, which is building an archive of spoken Danish based on audio recordings of sociolinguistic interviews carried out in the period 1970–2010. The corpus itself is stored on a server in a small room. So far, it contains some 8 million words recorded in sociolinguistic interviews and group conversations. The recordings have been transcribed and tagged for many kinds of information, allowing for a diversity of both quantitative and qualitative analyses of data that are readily retrievable, and available to any interested and authorized researcher. (Gregersen 2009 is a collection of articles that gives an impression of the breadth of possible analyses offered by the corpus).

Except for delivery time, the affordances of the postal system in terms of data gathering may not have changed much over the centuries. More recently, however, the character of long-distance mediated contact between dialectologists and their informants changed radically with the possibilities offered by the telephone system. Sitting by their phones, researchers can now contact many informants across large distances in little time. Data gathering by telephone was used as a supplementary control approach by Labov in both his New York project (Labov 1966, 118) and his Philadelphia project (Labov 2001, 69–73). It was the exclusive method of data collection in the Telsur Project, which collected data from “762 subjects in 323 communities, representing all cities with a population of over 50,000 in 1990” (Labov 2010, 9), and made it possible for the first time to provide a continent-wide *Atlas of North American English* (Labov, Ash, and Boberg 2006). The MIN project (*Modern Import Words in the Nordic Countries*) collected data on attitudes toward the influence of English by conducting telephone interviews with representative informant samples, totaling more than 5,000 informants across seven Nordic speech communities (Kristiansen and Vikør 2006;

Kristiansen 2010). With the advent of the internet and development of the appropriate software, practical constraints on asking questions and getting answers now hardly exist. Huge amounts of data can be collected quickly and without much effort. The “BBC Voices” survey, for instance, collected answers from more than 5,000 informants over 10 days in 2004 (Bishop, Coupland, and Garrett 2005, 133), whereas in 2010 the “Oslo test” reached more than 115,000 respondents in just two months (Stjernholm and Ims 2014).

Another important point to be added here concerns the internet’s role as a channel for sharing and distribution of data, and hence as a facilitator of comparisons of many kinds. This includes access not only to corpora of spoken data, but also to other sources of relevance for dialectology. By way of illustration, consider *Moths Ordbog*, which was the result of the first gathering of dialect material by correspondence in Denmark, dating back to the late seventeenth century. The task was undertaken by a high-ranking state official, Matthias Moth, who asked priests to send him lists of *ubrugelige bønderord* (“unusable peasant words”), on the basis of which Moth compiled a comprehensive dictionary that was never published until it was made available online in 2014 (http://mothsordbog.dk/moth_en?set_language=en).

Finally, we should recall that the development of transport and transfer technologies since the 1960s in itself creates new conditions for the use and distribution of both spoken and written language, not least in relation to geographical space. The latest media technology, in particular, affects the “global versus local” relationship—socially, language-ideologically and linguistically—and thus also has consequences for dialectology, in the sense of changing its object of study.

6.3 Variability, Variation

There can be no doubt that our present understanding of “variability” as a feature of language is radically different from how this feature was (or rather could be) understood before the 1960s.

Linguistic scholars have always been struggling with how to understand and describe the tension between variation and structure in language (see e.g., Gordon, this volume). When Labov entered the scene, he wrote about this issue in the introduction to *Sociolinguistic Patterns* by quoting words written by Uriel Weinreich in their groundbreaking jointly authored paper “Empirical foundations for a theory of language change”:

The facts of heterogeneity have not so far jibed well with the structural approach to language ... for the more linguists became impressed with the existence of structure of language, and the more they bolstered this observation with deductive arguments about the functional advantages of structure, the more mysterious became the transition of a language from state to state... The solution, we will argue, lies in the direction of breaking down the identification of structuredness with homogeneity. The key to a rational conception of language change – indeed, of language itself – is the possibility of describing orderly differentiation in a language serving a community. (Labov 1972, xxi; quoted with abbreviations from Weinreich, Labov, and Herzog 1968, 100, 101; some further abbreviations by the author).

It is the realization of this research program—“breaking down the identification of structuredness with homogeneity” through 50 years of empirical studies—which has changed our conception of system and variation in language beyond the point of no return. We now know that language change can be described as change in the systematicity of variation, and most of us will endorse the view that it is changes in the systematicity of variation that make language change understandable.

6.4 Linguistic Embedding

Linguistic entities of any kind come together with other linguistic entities of many kinds. These are studied by linguists in terms of form and meaning as they vary and change through interconnections and mutual influence. The study of these processes has always been a main focus of linguistic scholarship, and has provided much knowledge on how “internal factors” favor or constrain change within and across levels or subsystems of a language (sounds, grammar, words, and discourse).

It clearly stands out, not least in the area of phonetic/phonological analysis, that new technology plays an important role in the theoretical and methodological advances which have been made through variationist sociolinguistics. The phonetic transcription systems (“dialect alphabets”), which were created by the practitioners of the “new phonetics” at the end of the nineteenth century (e.g., in 1890 Otto Jespersen created *Dania* for the transcription of spoken Danish), were enthusiastically brought into play in dialectological fieldwork. But these alphabets were created to allow for the recording of perceived phonetic details, and the resulting transcriptions on reams of paper notes induce a sense of “drowning in detail” among analysts struggling with how to handle the structure-and-variation issue in the collected material (as did the editors of *Ømålsordbogen*). One may indeed wonder with Labov (2010, 149,150) if it is possible to achieve reliable inter-transcriber agreement on the 16 distinctions in vowel height that are evident in many dialect atlases. In mentioning this, Labov’s point is to stress how the acoustic measurements and scatterplots offered by present-day technology allow us to *see* what may be hard to hear. We can appreciate visually how vowel tokens are distributed in “phonological space” (a two-formant F1~F2 plane obtained by acoustic measurement, picturing the high-low and front-back dimensions of the oral cavity), and how the realizations are affected by linguistic context (for examples, see e.g., Labov 1994).

Otherwise, quantification and correlation form the basis of the variationist approach to illuminating the role of “internal factors.” Occurrences of the variants of a dependent variable are counted across groupings based on one or more independent variables. The resulting distribution is displayed in a table in terms of numbers and/or of percentages, and often also in a chart (as in Figure 6.1). Statistical significance testing may be used to

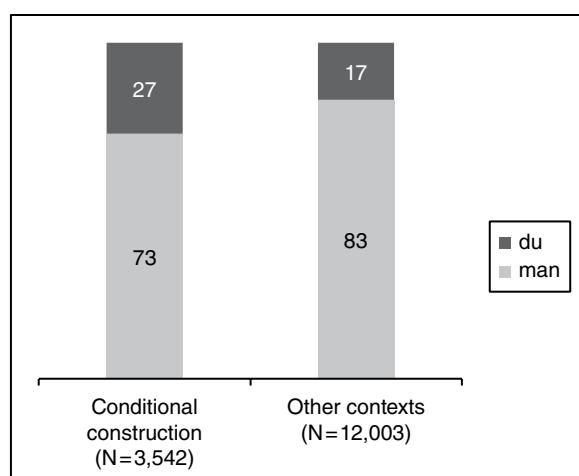


Figure 6.1 Distribution of *du* and *man* across “conditional construction” and “other contexts.”

help determine whether it is safe to generalize the observed difference(s) from the sample to a larger population. The example in Figure 6.1 presents results from Juel Jensen's analyses of the Danish generic pronoun variants *du* ("you") and *man* ("one"), and shows the distribution of these two variants across the syntactic context "conditional constructions" versus "other contexts." The sample consists of tokens of *du* and *man*, 3,542 in conditional construction context and 12,003 in other contexts, which were produced in the recorded speech of 161 informants in the LANCHART corpus. The population is the *du/man* frequency in these contexts in general. The generalizability of the observed difference (in percentages, 73% ~ 27% vs. 83% ~ 17%) was tested. Juel Jensen reports that "[a] chi square test for independence shows that the factor 'syntactic context' has a statistically significant influence on the choice of *du* versus *man* ($\chi^2 = 160.23$; 1 d.f.; $p < 0.01$)" (Juel Jensen 2009, 109).

6.5 On the "Explanatory Power" of Linguistic Embedding

With regard to the "why" questions of language variation and change, there is no doubt that the variationist explorations, as enabled by new technology, offer new perspectives and insights. For the example of the linguistic-embedding effect given in Figure 6.1, the following explanation is suggested by Juel Jensen (2009, 109): "A conditional construction is a type of context where the risk of misinterpreting the pronoun as referring specifically to the second person is reduced, and this is probably the reason for the relatively higher proportion of *du*." An "internal" (cognitive-functional) explanation of this kind sounds plausible, and is supported by the fact that the *du/man* distribution across "conditional construction" and "other contexts" has been shown to be unaffected by the "external" factors of age, gender, social class, and locality (Maegaard *et al.* 2013).

F1 ~ F2 scatterplots allowed Labov to detect for English, and for other modern West Germanic languages, two subsystems in both the front and back areas of the vowel space, for which he postulated "tracks" that vowels move along in processes of change. The distinction between a peripheral track and a nonperipheral track is defined in terms of more or less extreme F1 values, and the further conceptualization in terms of peripherality, including also the F2 dimension, has allowed Labov to expand on and develop the structuralist theories of merger and chain shift (Labov 1994, 2010). In the "meeting" between up-moving vowels (on the peripheral track) and down-moving vowels (on the nonperipheral track) it may happen that the phonemic distinction is lost—a vowel moves inward or outward to join a close neighbor on the parallel track, so to speak, and leaves a hole in the track it left, thus triggering a chain reaction, which maximizes phonological distance among the remaining members of that subsystem. As Labov puts it, "the governing principles of chain shifts operate only within subsystems and are triggered only when memberships in a subsystem undergoes change" (2010, 149).

Kerswill (2003, 229) presents scatterplots that show that changes in the vowel system of Ashford, east of London, is "a 'classical' chain shift," whereas changes in Reading, west of London, seem to be totally unsystematic until one discovers in the scatterplots that the endpoints of the vowels that move are the same in both cases. The result, we are told, "is convergence between the vowel systems east and west of the city" (2003, 230). An intriguing discovery, indeed, but one that nevertheless remains purely descriptive, as Kerswill himself points out.

In themselves, the facts of linguistic embedding cannot be taken as sufficient causes of change. But their description is an indispensable contribution to clarifying the conditions of change, and hence to sharpening the focus of our search for explanations.

6.6 Social Embedding: Rural and Urban Realities, Dialect and Standard

There is much truth in saying that sociodialectology is what we got when dialectology moved from the countryside into town, as is indicated by the label “urban dialectology.” The development of new theoretical and methodological principles was brought about by the shift of research focus to another social reality, in which the use of language was embedded in other and more complex social contexts. It should be stressed, though, that earlier dialectologists were far from blind to the social factors at work in the rural social contexts they investigated. Labov was fully aware of this when he moved dialectology to New York City: “The linguists who have contributed most to the study of language in its social context are primarily those who have worked in dialect geography” (Labov 1966, 16; see his overview in “Some earlier studies of language in its social context,” pp. 11–17). Louis Gauchat’s description of phonetic diversity in the Swiss village of Charmey (Gauchat 1905) is famous not least because Labov repeatedly makes mention of it as a work which paid attention to both *age* and *gender* (cf. indexes in Labov 1972, 1994, 2001).

The reality of *close-knit* versus *loose-knit* communities (we return to this distinction below) was reflected in traditional dialectology’s search for NORMs (Non-Mobile Older Rural Males, the “close-tie” prototypes) as informants, and in descriptions of how the close-knit communities and their traditional dialects were eroded by *mobile* people—on the one hand by outsiders who moved in, on the other by insiders who went to town for some time and then returned. Such people were “language missionaries,” in the terminology of Anders Steinsholt, who completed an early real-time study of the spreading of urban forms to the countryside north-east of Larvik, Norway (he collected and analyzed data at the end of the 1930s and again at the end of the 1960s; Steinsholt 1964, 1972). In the connections between town and countryside, the role of the outgoing and returning language missionary was typical of *young women*, according not only to Steinsholt’s report from Norway, but also according to the even earlier Danish investigations in the two rural communities of Åby (Jensen 1899) and Tvis (Skautrup 1921), both in Jutland (the continental, western part of Denmark). Being preoccupied with dedialectalization/standardization, these works describe in some detail how the local dialect was being “eroded” by incoming phonetic and lexical features, and report on how the *younger generations* and *women* were spearheading these new ways with language. Besides considering age and gender, both Jensen and Skautrup drew on official statistical data and their own insider knowledge of the communities, and offered rich quantitative and qualitative accounts of *social stratification* and *social life* in the two communities, highlighting the role of both social and geographical mobility in the process of dedialectalization/standardization.

Such traditional dialectological work, which we might classify in modern terms as dialectology of the sociology-of-language type, is of great value to contemporary efforts to describe how and explain why the traditional dialects were replaced by the standard language.

6.7 The “Primary Determinants”: Class, Sex, and Age

In his approach to language in urban community life, Labov lists six “major independent variables of sociolinguistics: sex, age, social class, ethnicity, race, and community size” (Labov 1994, 2). Chambers organizes his landmark volume *Sociolinguistic Theory* around the three first of these, stating that “[i]n modern industrial societies, these three social characteristics—class, sex and age—are the primary determinants of social roles,” and adding that “[t]hey are, of course, enormously complex, subsuming a host of social factors” (Chambers 1995, 7).

The social embedding of language variation is established by means of correlating frequencies of the studied linguistic variants with values of the social variables. The early, and classic, works in this tradition studied language variation and change in New York City (Labov 1966), Detroit (Wolfram 1969), and Norwich (Trudgill 1974).

Grouping speakers in terms of their class, sex, and age yields a picture of systematic linguistic variation at a rather abstract level of social structure. Two of the three primary variables, sex and age, have been “straightforwardly” based on biology-related information about the informants. As for social class, informants are asked for information about themselves that makes social class categorization possible, or, alternatively, they may be selected to represent a predefined social class structure established on the basis of criteria that are typically to do with income, education, or place of residence. In the Danish LANCHART project, which replicates and compares a number of sociolinguistic studies covering different types of communities from all over Denmark,¹ we chose to adopt the social class variable used in the Copenhagen study at the end of the 1980s (Gregersen and Pedersen 1991). The variable only distinguishes between Middle Class and Working Class, a dichotomy based on three indicators: education, power in the work place, and the specific content of the working process (Gregersen 2009, 9–11).

An example of the embedding of a linguistic variable in the social contexts of age, sex, and class is shown in Figure 6.2. In order to discuss it, we need to introduce the distinction between studies in apparent time and real time, a contrast that has been central to sociolinguistic study since its introduction by Labov in his New York City study (Labov 1966). Because of the lack of comparable data from the past, change has been studied in apparent time, that is, by comparing the speech of younger with the speech of older people. However, as sociolinguistic study itself has come of age, results from new studies may be compared with those of earlier ones. Thus, change can be studied in real time either via a *panel* study (based on new recordings of previous informants) or a *trend* study (based on recordings of new informants whose language can be compared with that of previous informants).

The example in Figure 6.2 shows results from the LANChART Jutland studies (Odder and Vinderup aggregated). The pictured results concern the most frequently occurring segmental difference between the western and eastern accents of modern Danish, which otherwise are now distinguished in the speech of younger people mainly by intonation (Kristiansen, Pharao, and Maegaard 2013). The eastern (Copenhagen) accent has almost exclusively [ð], whereas the western accent has [d ~ ð] variation. Young people recorded in the 1970 and 1980s showed a predominance of [d], which changed to a clear preference for [ð] among young people recorded in 2006 (real-time trend study). At the same time, recordings of the young 1970–1980s informants as adults 20 to 30 years later in 2006 (real-time panel study) showed that they too, to some extent, had followed the societal trend. Thus, these real-time studies, at the levels both of individuals (panel) and society (trend), indicate that the eastern variant [ð] is spreading at the expense of the western variant [d]. A synchronic comparison of adults and young people in 2006 (apparent-time study) indicates the same, of course. The frequency patterns displayed in respect of the class and sex variables show that [ð] is used more by middle-class people and by women, yielding the combined class + sex effect shown in the figure.

6.8 Social Networks

Sociolinguistic study has developed not least through reconsiderations of how to conceptualize and operationalize Chambers’ “primary determinants,” as well as the whole complexity of his “subsumed social factors.” The trend has been toward foregrounding social structures that are presumed to have more concrete reality in people’s lives, with theoretical

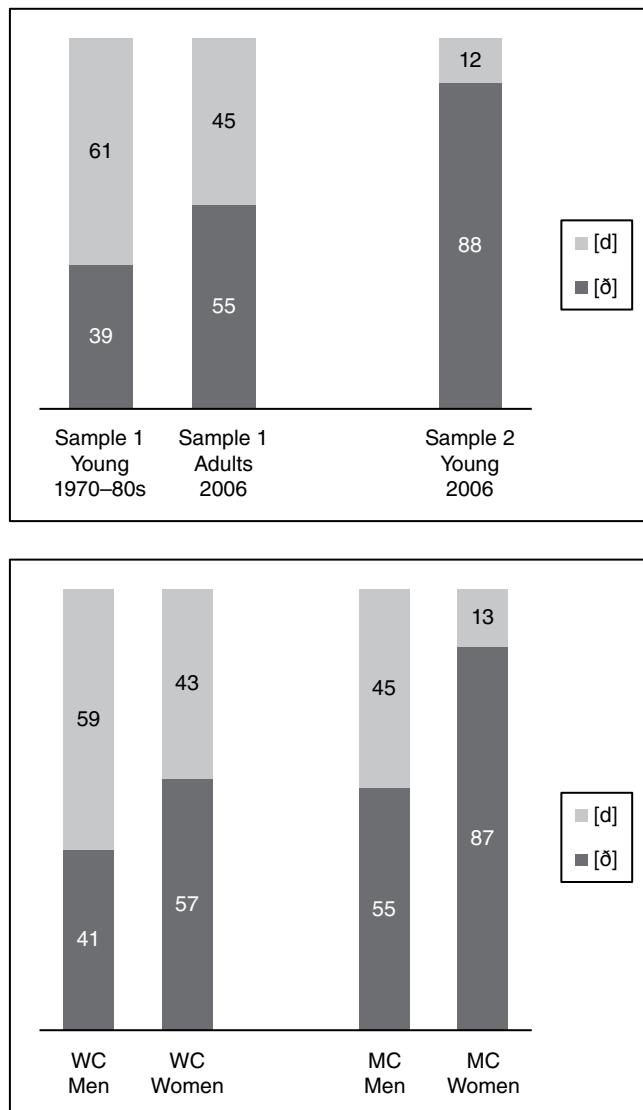


Figure 6.2 Results from the LANCHART Jutland studies. Percentages of the perfect participle morpheme *-et* pronounced with an “eastern” approximant [ð] or a “western” stop [d] by social groups distinguished in terms of *age/time* (upper), and *class/sex* (lower). The total number of tokens is 3,328, produced by 104 informants. Adapted from Maegaard and Juel Jensen (2010).

and methodological inspiration from disciplines such as sociology, anthropology, ethnography, and social psychology.

The concept of *social network* has been used in many studies, in many different ways, after its potentialities were demonstrated in Lesley and James Milroy’s work in Belfast (Milroy 1980, 2nd edition 1987a; Milroy 1992). The Milroys constructed a “network-strength scale” to be used as a measure of social cohesion and integration, and assigned network scores to each informant. The approach was particularly useful for linguistic analyses at the level of individuals, but was also used to shed light on differences in language use across three

geographical areas in Belfast (Milroy 1987a, 157ff.). In general, network studies have demonstrated an “association between a close-knit, localized network structure and adherence to a vernacular or (more broadly) nonlegitimized norm” (Milroy 1987a, 181). Inspired by the Belfast study, Inge Lise Pedersen in her study of the Danish rural community of Vissenbjerg in the mid-1980s (Pedersen 1994, reprinted in Pedersen 2009), assigned a network score to each informant based on observations to do with the following five “indicators” (cf. Milroy 1987a, 141–142): “1. Family relationship with at least two other households in the parish. 2. Work within the parish or together with at least two others from the parish. 3. Membership of a local group (organized or non-organized). 4. The cultivation of leisure activities together with at least two others from the neighborhood or with at least two local colleagues. 5. Both parents raised in the parish” (Pedersen 2009, 192). As the 15 linguistic variables in the study represented variation between “standard” and “dialect,” the general conclusion was formulated as a tendency to the effect that “the higher the network score, the more dialect-colored is the language” (2009, 193).

6.9 Communities of Practice

The concept of *community of practice* (Wenger 1998), as exploited by Penelope Eckert in her ethnographic approach to a community of high-school students in a Detroit suburb (Eckert 1989), introduced a kindred concretization of the social embedding issue, which nevertheless differed by changing the perspective from the role of social networks in language variation to the role of language variation in social meaning-making. “Meaning is made as people jointly construct relations through the development of a mutual view of, and relation to, the communities and people around them,” Eckert informs us; “This meaning-making takes place in myriad contacts and associations both within and beyond dense networks” (Eckert 2000, 34–35). It is in order to capture the process of meaning-making that we need to focus on “a level of social organization at which individual and group identities are being co-constructed,” according to Eckert, who also argues that the community of practice—“an aggregate of people who come together around some enterprise” (2000, 35)—is just such a level of social organization. In her analysis, the “Jocks” and “Burnouts” constitute communities of practice that have “emerged within, and in response to, the school’s institutional structure” (2000, 41), but they are also actors on the wider stage of the Northern Cities Chain Shift. In her own words, Eckert’s purpose was “to establish that variation carries social meaning that is very local, but embedded within a socio-geographic context, and systematically related to global patterns. [...] The link between the local and the global is the semi-local – the immediate geographic area that provides a concrete context for the local” (2000, 222). Her analyses of semi-local sociolinguistic dynamics, encompassing several high schools in the greater Detroit area, substantiate the claim that “[t]he potential for urban linguistic influence [...] is deeply embedded in the social practices of local communities of burnouts” (2000, 223).

Fundamentally, the reconsideration of social embedding in this approach is less about speaker characteristics and more about situation characteristics and how to deal with the notion of style (for discussions, see Coupland 1980, 2007; Eckert and Rickford 2001). In Labov’s work, style refers to linguistic variation across a continuum of situations that trigger different degrees of attention to speech, a continuum normally thought of in terms of formality. A formal situation, the argument goes, generates attention to speech and is therefore likely to trigger a careful style, whereas an informal situation is less likely to generate attention to speech and therefore triggers a more casual style. The “casual versus careful” distinction is linked to a distinction between a first-acquired and more systematic language (the “vernacular”) versus a later-acquired, less systematic language (the “standard”). The privileged object of study is casual style speech. In this respect, Labov’s approach is

reminiscent of theory and practice in traditional dialectology. Analyses of this kind of “stylistic variation” on dialect~standard variables will normally show the standard variants to be the more frequent in formal situations, and do in that sense shed light on processes of standardization/dedialectalization. In so-called “third-wave” variationist sociolinguistics (Eckert 2012), the linguistic variable is replaced by linguistic *resources* (or *features*) as the preferred analytical tool. Language variation is a “resource” in social meaning-making that acts in concert with a host of other resources to do with factors such as physical appearance, whereas *style* is conceived of as socially meaningful clustering of such resources.

In the first Eckert-inspired dissertation in Denmark, Pia Quist (2005, published as Quist 2012) established seven style clusters, three masculine and four feminine, during her ethnographic work at a high school in Nørrebro, a multi-ethnic neighborhood in central Copenhagen. The clusters are characterized—that is, delimited in relation to each other—in terms of sameness or difference with respect to a series of non-linguistic and linguistic features. The first masculine cluster is characterized as follows:

White skin, masculine body sign; ‘Dane’; show no to little interest in class teaching; use computers for games, chat and music; smoke; leave the school premises during lunch breaks; drink alcohol and talk about it in class; wear hip-hop clothes/baggy jeans and large T-shirts; images of naked women as computer wallpaper; listen to hip-hop and rock music.

Long (t). Frequent use of ‘lexis’— mostly slang and swear-words. ÷ use multiethnolectal features (Quist 2008, 60).

The interest lies not in how speaker categories and/or context categories “affect” a person’s use of language, but in how speakers use language to “construct” styles and personae. Quist stresses that “[t]he clusters should not be confused with groups; an individual does not belong to a cluster, but rather performs a style and that way constitutes it—through daily acts and practices” (2008, 52). This approach to style has implications for the study of language variation in geographical space: “It is not possible to describe the speech in Nørrebro as one variety (as for instance is normally done in a traditional dialect study); and it is not possible to point out a proto-typical Nørrebro-speaker” (2008, 59). Quist argues that “[t]he big city is diffuse if we look at it in terms of traditional divisions according to status, ethnicity and geography. But the practices of the citizens of the big city are not necessarily diffuse” (2012, 381; original emphasis, my translation). This is why the construct of “practice” allows us to “describe regularities in the mixed, late-modern communities, regularities that are based on practices and not on social groups or categories” (2008, 380). The way to address variation in geographical space would be to compare communities of practice in different parts of the country. This could be done, in the case of multiethnolectal practices, “for instance by comparing the three biggest cities Copenhagen, Aarhus and Odense in terms of their local and national similarities and differences” (Quist 2010, 10).

6.10 Combining Narrow and Broad Perspectives

While searching for ways to illuminate the more concrete aspects of social embedding, most sociodialectologists will strive not to lose the broader perspective. In a way reminiscent of Eckert’s semi-local analysis (see above), Quist (2012, 380) suggests combining the community of practice approach with a broader social network analysis, and also points to the possibility of combining style-cluster analysis with analysis based on the concept of *lifestyle* (inspired by Bourdieu 1984, and used in several Scandinavian dissertations, e.g., Akselberg 1995, Grönberg 2004, and Røyneland 2005).

Pedersen (1994/2009) combined an analysis based on social network theory with one based on a theory of three *life modes* (rural, worker, career) that was developed by the Danish ethnologist Thomas Højrup (1983). Pedersen did so, she argues, because “[t]here are several obvious differences between, for example, Georg’s and Knud’s views of themselves and between the way they choose to live their everyday lives that are not captured by the network analysis used here. Not all workers can be viewed as having a worker life mode” (2009, 198). Højrup’s life mode theory has also been adopted by the Milroys for linking analyses at the levels of social network with analyses at the level of socioeconomic class (Milroy and Milroy 1991; Milroy 1992, 214–220).

Labov (2010, 189), in discussing the concepts of social networks and communities of practice, admits that “[m]uch is to be learned from the study of individual variation, in seeing how individuals make use of the complex structure of community variation to evoke different social identities,” but argues that the highly regular patterns of participation in change throughout the larger speech community “call for the recognition of larger social forces operating outside of the individual’s control.”

Such considerations about the merit of combining approaches reflect a general awareness that the development of satisfactory explanations for language variation and change requires exploration of social embedding at various levels of abstraction/concreteness.

6.11 On the “Explanatory Power” of Social Embedding

Just as with linguistic embedding, the facts of social embedding cannot in themselves be taken as sufficient causes of change. Their description is nonetheless an indispensable contribution to asking the good “why” questions. As to Chambers’ “primary determinants” (age, gender, class), it is part of their “explanatory potential” that expanding linguistic variants are typically found to be used more frequently by young people, by females, and by members of the middle class (or, if more layers are included in the social stratification, by “centrally located groups as against peripherally located groups” (Labov 2001, 32). Knowing this, we may assume that it is not only the age-/time-related frequency pattern in Figure 6.2 that reveals change; the gender- and class-related pattern may also be taken as an indication that [ð] is spreading, in that it is used more by middle-class (MC) people and by women, yielding a combined class+sex effect. MC women are spearheading the change, whereas working-class (WC) men lag behind. The social embedding of the *du/man* variation shows the same pattern. Mixed-effects modeling strongly suggests an ongoing decrease in the use of *du* from the 1980s recordings to the 2000s recordings, spearheaded by younger people, middle-class people, and females (Maegaard *et al.* 2013). The descriptive patterns represent intriguing answers to the “who” question, and strong incentives to go on attempting to answer the “why” question.

The distribution of linguistic variation in geographical space is an aspect of social embedding, of course, and dialectologists have naturally always been interested in how and why isoglosses move (i.e., how and why linguistic features spread). The more detailed descriptions of how linguistic features are distributed across geographical space, as obtained by variationist frequency analyses, have given rise to several models of the process of spreading, focusing on the trajectory of linguistic change. Where is the source of an innovation, and whither is it spread? (Britain 2010 provides an overview of types of diffusion models, along with many references). Which model fits various realities best is an empirical question, of course.

The distributional facts themselves will often be suggestive of social embedding, in terms of communication lines between centers and peripheries. An isogloss which isolates small, geographically dispersed areas along the periphery of a larger area will be suggestive of

communication and spread from a center; one that unites areas divided by water will be suggestive of intense contact in earlier times by ship, as for instance the “Danish-like” lenition of plosives on the southeastern coast of Norway facing Jutland. Conversely, sparse contact and communication have been suggested as the explanation for an old dialect boundary between the eastern and western parts of Jutland. This was the result not of any significant natural obstacles but of the existence of a thinly-populated north/south belt “in the middle.” This has been called “Denmark’s most famous dialect boundary” (Thorsen 1912), with reference to a highly salient difference: the use of a post-positioned determiner on nouns in the east (as per the rest of Scandinavia, *hus-et* “the house”), versus a pre-positioned determiner in the west (as in German and English, *a hus* “the house”). Other salient differences concerning grammatical gender and the prosodic *stød*-system by and large follow the same dividing line. Independent evidence for the “sparse population” argument is found in placename research, and in old cultural differences between eastern and western Jutland.

In studies based on the concepts of social network or community of practice, the interest in weaknesses and strengths of communication lines, and their role in linguistic diffusion, is moved to a more concrete level, one at which the issues at stake concern the types of network and person involved. Who spearheads a change, and who lags behind, in terms of people acting in their daily connections with other people? Individuals who engage in many different connections, inhabiting loose-knit networks with “weak ties,” will be liable to adopt new linguistic practices; people whose daily connections form a close-knit network with “strong ties,” by contrast, are more likely to stick to existing practices. With reference to positions in network structure, different roles in the diffusion process have been theorized and discussed in terms of “leaders,” “innovators,” and “early adopters” (Labov 2001; Milroy 1992), and “sociolinguistic icons” (Eckert 2000). Speakers’ behavior with respect to the social networks they occupy is highly relevant to the theorization of geographical linguistic diffusion, in the sense that traditional rural communities are assumed to have been characterized by close-knit networks which broke down and were replaced by loose-knit networks as urbanization came along (urbanization in the sense that people began to commute or moved to town, or in the sense that urban life expanded into the countryside). The argument is hardly affected by the assumption that “the type of close-knit community that is most easily conceptualized in (close-tie) network terms is as likely to be a product of modern city life as it is to be a residue of an earlier type of social organization” (Milroy 1992, 212).

Contemporary variationist studies can sharpen the focus on the mechanisms of this process, not least by concentrating on what community as geographical “place” means to geographically mobile people. In a real-time panel study based on recordings of adolescents from the 1970/1980s and again with the same people as adults in recordings from the 2000s, Monka (2013) compared people who had moved away from one of the three Jutland towns of Odder, Vinderup and Tinglev with people who had stayed.² She found, rather expectedly, that those who had moved used less dialect, but she also found, more intriguingly, that this difference existed to begin with: it was in evidence in the early recordings, before any moving away had taken place. Moving had not enlarged the difference. It is as if an already existing difference in “mental inclination to move” had caused a difference in language use.

6.12 Language-Ideological Embedding

As human beings we are always involved in negotiating our social identities in mental, subjective processes of comparison and evaluation (Hogg and Abrams 1988), not least in terms of how we speak. Chambers (1995, 250) contends that “[t]he underlying cause of sociolinguistic differences, largely beneath consciousness, is the human instinct to establish and maintain social identity,” and that the “common motive” behind difference and uniformity in

language is “[...] the profound need for people to show they belong somewhere, and to define themselves, sometimes narrowly and sometimes generally.” Against this backdrop, it seems only natural that the facts of linguistic and social embedding are not, in themselves, felt to be good enough answers to the “why” question. This is true also for the simple reason that such facts quite obviously do not always have a bearing on language use. Moving from one place to another does not make everyone modify their speech; formal situations do not *necessarily* trigger careful speech; “natural” phonetic changes (e.g., assimilations) do not occur everywhere. Hence, scholars share a widespread feeling that we need to arrive at “another, perhaps more basic sense of why: why as a search for motivating or efficient causes” (Labov 2010, 184). James and Lesley Milroy have often pointed out that “[t]he weak-tie model is not in itself sufficient to provide a full social explanation of linguistic change. What it proposes is a set of conditions that are necessary—but not sufficient—for linguistic change to take place. [...] It is not about psycho-social attitudes to language” (Milroy 1992, 204).

Nevertheless, in spite of the widespread recognition that these efficient causes and valid explanations are to be sought for in terms of language-ideological embedding, this is certainly a less developed aspect of variationist sociolinguistics as a discipline. Developments in theorizing and operationalizing the social psychological processes involved in language variation and change are more associated with social psychology of language (Giles and St. Clair 1979; Scherer and Giles 1979; Giles and Robinson 1990) and perceptual dialectology (Preston 1989; Preston 1999; Long and Preston 2002; Niedzielski and Preston 2003). It is not that variationists have been uninterested in these developments. The Milroys have suggested, in accordance with the findings of language attitudes research by social psychologists (e.g., Brown and Gilman 1960; Ryan and Giles 1982), that an integrated model of sociolinguistic structure must take into account the competing ideologies of solidarity and status (Milroy 1987b, 208–209; Milroy 1992, 210, 213), and have furthermore stated that “models of social identity (Le Page and Tabouret-Keller 1985), accommodation (Giles and Smith 1979) and politeness (Brown and Levinson 1987) will not be irrelevant [to the further development of our social model of language change]” (Milroy 1992, 221). Accommodation theory (Giles and Powesland 1975), in particular, has been an important theoretical inspiration in some variationist work, including classics like Bell (1984), Coupland (1984), and Trudgill (1986).

However, the empirical search for independent evidence of a motivating social-psychological force behind linguistic diffusion has often been low on the agenda of variationist sociolinguistics. This may be seen as rather surprising if we consider that Labov in his NYC study had already adopted a modified form of the matched guise technique, developed by the Canadian social psychologist Wallace Lambert and colleagues (Lambert *et al.* 1960) for the study of language attitudes. Also, the “evaluation problem” figured prominently among the problems to be solved in the seminal article “Empirical foundations for a theory of language change,” which stated that “the study of the *evaluation* problem in linguistic change is an essential aspect of research leading to an explanation of change” (Weinreich, Labov, and Herzog 1968, 165). In Labov’s (1972, 162) wording, “[t]he *evaluation* problem is to find the subjective (or latent) correlates of the objective (or manifest) changes which have been observed”; and in his own projects—on Martha’s Vineyard (Labov 1963), in New York City (Labov 1966, Ch. 3), and in Philadelphia (Labov 2001, Ch. 6)—Labov has included and developed methods to “find the subjective correlates” (see also Labov 1984). Reasons for not following Labov in this respect are seldom explicitly given, but seem to be of two kinds. Either it is felt that problems of validity and reliability are too serious for attitudinal data to be of much interest to the study of language change (e.g., Milroy 1987a, 141; 1987b, 107), with the logical implication that “statistical counts of variants actually used are probably the best way of assessing attitudes” (Milroy and Milroy 1991, 19), or the various kinds of data gathered by studies taking an ethnographic approach to social networks and communities of practice are

evidence of language-ideological embedding in its own right. This seems to be the view expressed by Eckert in the conclusion of her article on the “three waves” of sociolinguistics: “The third wave locates ideology in language itself, in the construction of meaning, with potentially important consequences for linguistic theory more generally” (Eckert 2012, 98).

Whatever the consequences might be, I do not think they should include a denial of the possible benefits of collecting data on subjective correlates; on the contrary. Let me quote Irvine (2001, 24): “By foregrounding ideology I emphasize the need to investigate ideas about language and speakers independently of empirical distributions, and the need to recognize that ‘attitudes’ include participants’ basic understandings of what the sociolinguistic system consists of, not just emotional dispositions.” Irvine suggests, with reference to Silverstein, that “the best place to look for language ideology may lie in the terms and presuppositions of metapragmatic discourse, not just in assertions” (2001, 25). This is likely to be true in many cases. But in the case of linguistic diffusion, I think that experimental studies will often yield more illuminating data. Maegaard (2007) carried out an Eckert-inspired ethnography-based study of style clusters in a Copenhagen school, and subsequently designed a speaker evaluation experiment on the basis of the obtained results. She administered the latter not only to students in the same school, but also to students in another area of Copenhagen, in order to investigate whether the social values attached to the speakers were more widely shared by Copenhagen adolescents. She found that similarity dominated, but also some difference, and concluded: “This shows [...] that local and global meaning making are interrelated and that even though social meaning is experienced in the local context, it can very well draw on meaning potentials of a more global character” (Maegaard 2010, 205).

Indeed, the language-ideological studies of the LANCHART project have firmly documented that the speech variation which is relevant to social identifications among Danish youth today involves social meaning potentials which are shared “globally,” not only across different areas of Copenhagen, but across the whole geographical space called Denmark. Copenhagen is Denmark’s sole linguistic norm center. The far-reaching *linguistic* “Copenhagenization” of the country has its correlate in an even more ubiquitous *subjective* Copenhagenization: everywhere, young people not only downgrade their own local accents (which will differ from Copenhagen speech in prosody only; Kristiansen *et al.* 2013) relative to young Copenhagen speech, but they also perceive and evaluate the variation in this speech—between a conservative and a modern accent—in exactly the same way as young Copenhageners do themselves. In speaker evaluation experiments, “modern” is strongly upgraded in terms of personality traits to do with *dynamism*, whereas “conservative” does as well, or better, on traits to do with *superiority*. Importantly, this pattern emerges only when the evaluations are subconsciously offered, that is, when subjects are not aware they are revealing their language attitudes (Kristiansen 2009).

Since the Copenhagenization process is more advanced in the subjective aspect than in its linguistic aspect, we argue that ideology is the driving force of the process, rather than a concomitant of it. Furthermore, since the subconscious value system emerges with carbon-copy uniformity among adolescents across the whole country, it has to derive from some shared experience; and as we can only catch sight of the modern media universe as a likely source of this shared experience, we suggest that the modern media play an important role in the linguistic Copenhagenization of the country (Maegaard *et al.* 2013). The modern media have great influence on speakers and their speech across the Danish landscape, that is, great influence on linguistic diffusion, albeit not in a direct sense, but indirectly, by reshaping the language-ideological embedding of speakers and their speech (Kristiansen 2014). In sum, late-modern Denmark has been thoroughly Copenhagenized through variation and change in the social, language-ideological, and linguistic spheres. By submitting their complex interdependence to sociolinguistic study, sociodialectology illuminates the reality of these processes in geographical space.

NOTES

- 1 The studies replicated by LANCHART cover the geographical space of Denmark from the eastern island of Sealand (previous studies in the capital city of Copenhagen and the smaller regional center of Næstved), across the island of Funen (previous studies in the town of Vissenbjerg close to Odense, Denmark's third-largest city), to the continental part of Denmark, Jutland, in the west (previous studies in the east Jutland town of Odder close to Aarhus, Denmark's second-largest city, and in the small west-Jutland town of Vinderup).
- 2 Tinglev is a town in south-Jutland, near the German border.

REFERENCES

- Akselberg, Gunnstein. 1995. *Fenomenologisk dekonstruksjon av det labov-milroyske paradigmet i sosiolinguistikken: Ein analyse av sosiolinguistiske tilhøve i voss kommune*. Bergen: University of Bergen.
- Bell, Allan. 1984. "Language style as audience design." *Language in Society*, 13: 154–204.
- Bishop, Hywel, Nikolas Coupland, and Peter Garrett. 2005. "Conceptual accent evaluation: Thirty years of accent prejudice in the UK." *Acta Linguistica Hafniensia*, 37: 131–154.
- Bourdieu, Pierre. 1984. *Distinction. A Social Critique of the Judgement of Taste*. London: Routledge.
- Britain, David. 2010. "Language and space: The variationist approach." In *Language and Space: An International Handbook of Linguistic Variation, Vol. 1: Theories and Methods*, edited by Peter Auer, and Jürgen Schmidt, 142–163. Berlin: de Gruyter.
- Britain, David, and Jenny Cheshire, eds. 2003. *Social Dialectology: In Honour of Peter Trudgill*. Amsterdam: Benjamins.
- Brown, Roger, and Albert Gilman. 1960. "The pronouns of power and solidarity." In *Style in Language*, edited by Thomas Sebeok, 253–276. Cambridge, MA: MIT Press.
- Brown, Penelope, and Stephen Levinson. 1987. *Politeness*. Cambridge: Cambridge University Press.
- Chambers, Jack. 1995. *Sociolinguistic Theory*. Oxford: Blackwell.
- Chambers, Jack, and Peter Trudgill. 1980. *Dialectology*. Cambridge: Cambridge University Press.
- Coupland, Nikolas. 1980. "Style-shifting in a Cardiff work-setting." *Language in Society*, 9: 1–12.
- Coupland, Nikolas. 1984. "Accommodation at work: Some phonological data and their implication." *International Journal of the Sociology of Language*, 46: 49–70.
- Coupland, Nikolas. 2007. *Style: Language Variation and Identity*. Cambridge: Cambridge University Press.
- Eckert, Penelope. 1989. *Jocks and Burnouts. Social Categories and Identity in the High School*. New York: Teachers College Press.
- Eckert, Penelope. 2000. *Linguistic Variation as Social Practice*. Oxford: Blackwell.
- Eckert, Penelope. 2012. Three waves of variation study: The emergence of meaning in the study of variation. *Annual Review of Anthropology*, 41: 87–100.
- Eckert, Penelope, and John Rickford, eds. 2001. *Style and Sociolinguistic Variation*. Cambridge: Cambridge University Press.
- Gauchat, Louis. 1905. "L'unité phonétique dans le patois d'une commune." In *Aus Romanischen Sprachen und Literaturen: Festschrift Heinrich Morf*, 175–232. Halle: Niemeyer.
- Giles, Howard, and Peter Powesland. 1975. *Speech Style and Social Evaluation*. London: Academic Press.
- Giles, Howard, and Robert St. Clair, eds. 1979. *Language and Social Psychology*. Oxford: Blackwell.
- Giles, Howard, and Peter Robinson, eds. 1990 [2nd ed. 2001]. *Handbook of Language and Social Psychology*. Chichester: Wiley.
- Giles, Howard, and Philip Smith. 1979. "Accommodation theory: Optimal levels of convergence." In *Language and Social Psychology*, edited by Howard Giles, and Robert St. Clair, 45–65. Oxford: Blackwell.

- Gregersen, Frans, ed. 2009. *Acta Linguistica Hafniensia*, 41.
- Gregersen, Frans. 2009. "The data and design of the LANCHART study." *Acta Linguistica Hafniensia*, 41: 3–29.
- Gregersen, Frans, and Inge Lise Pedersen, eds. 1991. *The Copenhagen Study of Urban Sociolinguistics 1–2*. Copenhagen: Reitzel.
- Grönberg, Anna Gunnarsdotter. 2004. *Ungdomar och dialekt i alingsås*. Gothenburg: Acta Universitatis Gothoburgensis.
- Hogg, Michael, and Dominic Abrams. 1988. *Social Identifications: A Social Psychology of Intergroup Relations and Group Processes*. London: Routledge.
- Hojrup, Thomas. 1983. *Det glemt folk: Livsform og centraldirigerig*. Copenhagen: Statens Byggeforskningsinstitut.
- Irvine, Judith. 2001. "'Style' as distinctiveness: The culture and ideology of linguistic differentiation." In *Style and Sociolinguistic Variation*, edited by Penelope Eckert, and John Rickford, 21–43. Cambridge: Cambridge University Press.
- Jensen, Anker. 1898. "Sproglige forhold i Åby sogn Århus amt." *Dania*, 5: 213–231.
- Juel Jensen, Torben. 2009. "Generic variation? Developments in use of generic pronouns in late 20th century spoken Danish." *Acta Linguistica Hafniensia*, 41: 83–115.
- Kerswill, Paul. 2003. "Dialect levelling and geographical diffusion in British English." In *Social Dialectology*, edited by David Britain, and Jenny Cheshire, 223–243. Amsterdam: Benjamins.
- Kerswill, Paul. 2004. "Social dialectology / Sozialdialektologie." In *Sociolinguistics/Soziolinguistik: An International Handbook of the Science of Language and Society*, Vol. 1, 2nd ed., 22–33. Berlin: de Gruyter.
- Kristiansen, Tore. 2009. "The macro-level social meanings of late-modern Danish accents." *Acta Linguistica Hafniensia*, 41: 167–192.
- Kristiansen, Tore. 2010. "Conscious and subconscious attitudes towards English imports in the Nordic countries: Evidence for two levels of language ideology." *International Journal of the Sociology of Language*, 204: 59–95.
- Kristiansen, Tore. 2014. "Does mediated language influence immediate language?" In *Mediatization and Sociolinguistic Change*, edited by Jannis Androutsopoulos, 99–126. Berlin: de Gruyter.
- Kristiansen, Tore, and Lars Vikør, eds. 2006. *Nordiske språkhaldninger: Ei meiningsmåling*. Oslo: Novus.
- Kristiansen, Tore, Nicolai Pharao, and Marie Maegaard. 2013. "Controlled manipulation of intonational difference: An experimental study of intonation patterns as the basis for language-ideological constructs of geographical provenance and linguistic standardness in young Danes." In *Language (De)standardisation in Late Modern Europe: Experimental Studies*, edited by Tore Kristiansen, and Stefan Grondelaers, 355–374. Oslo: Novus.
- Labov, William. 1963. "The social motivation of a sound change." *Word*, 19: 273–309.
- Labov, William. 1966. *The Social Stratification of English in New York City*. Washington DC: Center for Applied Linguistics.
- Labov, William. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.
- Labov, William. 1984. "Field methods of the project on linguistic change and variation." In *Language in Use*, edited by John Baugh, and Joel Sherzer, 28–53. Englewood Cliffs, NJ: Prentice-Hall.
- Labov, William. 1994. *Principles of Linguistic Change, Vol. 1: Internal Factors*. Oxford: Blackwell.
- Labov, William. 2001. *Principles of Linguistic Change, Vol. 2: Social Factors*. Oxford: Blackwell.
- Labov, William. 2010. *Principles of Linguistic Change, Vol. 3: Cognitive and Cultural Factors*. Oxford: Wiley.
- Labov, William, Sharon Ash, and Charles Boberg. 2006. *Atlas of North American English: Phonology and Sound Change*. Berlin: de Gruyter.
- Lambert, Wallace, Robert Hodgson, Robert Gardner, and Steven Fillenbaum. 1960. "Evaluational reactions to spoken languages." *Journal of Abnormal and Social Psychology*, 60: 44–51.
- LANCHART. <http://lanchart.hum.ku.dk>. Accessed October 19th 2015.
- LePage, Robert, and Andrée Tabouret-Keller. 1985. *Acts of Identity: Creole-based Approaches to Language and Ethnicity*. Cambridge: Cambridge University Press.
- Long, Daniel, and Dennis Preston, eds. 2002. *Handbook of Perceptual Dialectology, Vol 2*. Amsterdam: Benjamins.
- Maegaard, Marie. 2007. "Udtalevariation og -forandring i Københavnsk." *Danske Talesprog*, 8: 1–277.
- Maegaard, Marie. 2010. "Linguistic practice and stereotypes among Copenhagen adolescents." In *Multilingual Urban Scandinavia: New Linguistic Practices*, edited by Pia Quist and Bente Svendsen, 189–206. Clevedon: Multilingual Matters.

- Maegaard, Marie, and Torben Juel Jensen. 2010. "Hvor er bleven blevet af?" *Danske Talesprog*, 10: 34–56.
- Maegaard, Marie, Torben Juel Jensen, Tore Kristiansen, and Jens Normann Jørgensen. 2013. "Diffusion of language change: Accommodation to a moving target." *Journal of Sociolinguistics*, 17: 3–36.
- Milroy, Lesley. 1987a. *Language and Social Networks*. 2nd ed. [1st ed. 1980]. Oxford: Blackwell.
- Milroy, Lesley. 1987b. *Observing and Analysing Natural Language*. Oxford: Blackwell.
- Milroy, Lesley, and James Milroy. 1992. "Social networks and social class: Toward an integrated sociolinguistic model." *Language in Society*, 21: 1–26.
- Milroy, James. 1992. *Linguistic Variation and Change*. Oxford: Blackwell.
- Milroy, James, and Lesley Milroy. 1991. *Authority in Language*, 2nd ed. [1st edition 1985]. London: Routledge.
- Monka, Malene. 2013. "Sted og sprogforandring – en undersøgelse af sprogforandring i virkelig tid hos mobile og bofaste informanter fra Odder, Vinderup og Tinglev." *Danske Talesprog*, 13: 1–336.
- Niedzielski, Nancy, and Dennis Preston. 2003. *Folk Linguistics*. Berlin: de Gruyter.
- Pedersen, Inge Lise. 1994. "Linguistic variation and composite life modes." In *The Sociolinguistics of Urbanization: The Case of the Nordic Countries*, edited by Bengt Nordberg, 87–206. Berlin: de Gruyter. [Reprinted as pages 180–206 in Pedersen 2009].
- Pedersen, Inge Lise. 2009. *Fra folkemål til multietnolekt*. Oslo: Novus.
- Preston, Dennis. 1989. *Perceptual Dialectology: Nonlinguists' Views of Areal Linguistics*. Dordrecht: Foris.
- Preston, Dennis, ed. 1999. *Handbook of Perceptual Dialectology*, Vol 1. Amsterdam: Benjamins.
- Quist, Pia. 2008. "Sociolinguistic approaches to multietnolect: Language variety and stylistic practice." *International Journal of Bilingualism*, 12: 43–61.
- Quist, Pia. 2010. "The sociolinguistic study of youth and multicultural practices in Denmark: An overview." In *Multilingual Urban Scandinavia: New Linguistic Practices*, edited by Pia Quist, and Bente Svendsen, 6–11. Clevedon: Multilingual Matters.
- Quist, Pia. 2012. *Stilistisk praksis: Unge og sprog i den senmoderne storby*. Copenhagen: Museum Tusculanum.
- Ryan, Ellen, and Howard Giles, eds. 1982. *Attitudes towards Language Variation: Social and Applied Contexts*. London: Arnold.
- Røyneland, Unn. 2005. *Dialektnivellering, ungdom og identitet: Ein komparativ studie av språkleg variasjon og endring i to tilgrensande dialektområde, Røros og Tynset*. Oslo: University of Oslo.
- Scherer, Klaus, and Howard Giles, eds. 1979. *Social Markers in Speech*. Cambridge: Cambridge University Press.
- Skautrup, Peter. 1921. "Om Folke- og Sprogblanding i et vestjysk Sogn." *Danske Studier*, 97–111.
- Steinsholt, Anders. 1964. *Målbryting i Hedrum*. Skrifter fra Norsk Målførerekav 19. Oslo: Universitetsforlaget.
- Steinsholt, Anders. 1972. *Målbryting i Hedrum 30 år etter*. Skrifter fra Norsk Målførerekav 26. Oslo: Universitetsforlaget.
- Stjernholm, Karine, and Ingun Indrebø Ims. 2014. "Om bruk av oslotesten for å undersøke oslomålet." *Norsk Lingvistisk Tidsskrift*, 32: 100–129.
- Swann, Joan, Ana Deumert, Theresa Lillis, and Rajend Mesthrie. 2004. *A Dictionary of Sociolinguistics*. Tuscaloosa, AL: University of Alabama Press.
- Thorsen, Peder. 1912. "Den berømteste Dialektgrænse i Danmark." *Afhandlinger og Breve II*: 107–135. (Volumes I–III edited and published by Jens Byskov and Marius Kristensen, 1927–1930).
- Trudgill, Peter. 1974. *The Social Differentiation of English in Norwich*. Cambridge: Cambridge University Press.
- Trudgill, Peter. 1986. *Dialects in Contact*. Oxford: Blackwell.
- Weinreich, Uriel, William Labov, and Marvin Herzog. 1968. "Empirical foundations for a theory of language change." In *Directions for Historical Linguistics: A Symposium*, edited by Winfred Lehmann, and Yakov Malkiel, 95–188. Austin, TX: University of Texas Press.
- Wenger, Etienne. 1998. *Communities of Practice: Learning, Meaning, and Identity*. Cambridge: Cambridge University Press.
- Wolfram, Walt. 1969. *A Sociolinguistic Description of Detroit Negro Speech*. Washington DC: Center for Applied Linguistics.
- Ømålsordbogen 1–, 1992–. (Edited and published by Copenhagen University's Department of Scandinavian Research/Section of Dialectology).

7 Dialectometry

HANS GOEBL

7.1 Introduction

In my treatment of dialectometry in this chapter, I will consider the topic from the standpoint of a Romance geolinguist. It seems necessary to emphasize this point, because we can observe a certain dissimilarity in the linguistic geographies of Romance, German, and English linguistics, a fact which sometimes complicates understanding. I would also like to emphasize that my contribution relies completely and solely on data from linguistic atlases. Whenever I subsequently speak of “geolinguistic variation,” the reader should be aware that my ideas are based on linguistic atlas data.

Dialectometry might thus more properly be called *atlantometry*. The fundamentally *inductive* character of dialectometry must also be pointed out. The preponderant aim of dialectometry consists in discovering, via the numerical analysis of many concrete patterns,¹ more abstract patterns which would otherwise remain hidden. We do so in order to obtain a systematic insight into the problem of, so to speak, the “basilectal management of space by *Homo loquens*.” The primary motives for research in dialectometry are thus linguistic, and have initially nothing to do with quantity. Many dialectometrists have adopted a similar position, including in particular, Jean Séguy, the original creator of the term and the method of “*dialectométrie*” (Séguy 1973: 1).

Historically speaking, the position of Séguy (1914–1973) is that of a single link in a long chain. Obviously, this metaphorical chain is located in France, and it refers to some very important chapters of French spiritual and scientific evolution. Séguy thus has many fore-runners whose contributions and importance will be presented below (see also the historical overviews given in Goebel, 2006b, 2013d).

7.2 Historical Background A: From the *Ancien Régime* up the End of the Nineteenth Century

Our historical retrospective starts with the gradual emergence of a “geodetic” conception of the territory of France under the *Ancien Régime*, the political system in force during the period from the sixteenth to the late eighteenth centuries. During this time, a large number of geographers, economists, tax collectors, and military engineers tried not only to measure the size of the territory of France but also to grasp the importance of certain socio-economic

variables in order to inform the king about the riches of his realm. Witness, for instance, the activity and writings of four generations of the Cassini family (from Jean Dominique Cassini, 1625–1712, until Jean Dominique, comte de Cassini, 1748–1845).

As a result, the general idea developed—obviously in a slow and rather subconscious manner—that the space of France was a kind of machine capable of producing a certain quantity of goods, with gears that meshed according to certain inner principles.

The many reforms of the French Revolution (1789–1799) and the specifically centralistic spirit of the Napoleonic administration (1799–1814) considerably accelerated new concepts of space that were emerging at the time. The most important event in this respect was the abolition of the old subdivision of France into its historical provinces, and the establishment of the Republican grid of departments. The central philosophy of the new grid had two characteristics of key importance from the *metrological* point of view: (a) the equal size and the even distribution of the different departments (whose number initially was 83) over France, and (b) the definition of a canonical *spatial sampling* for further research.

The new grid was in fact used immediately not only by the Napoleonic administration but also by the subsequent regimes (Monarchies of Bourbon and Orleans 1814–1848, Second Republic 1848–1852, Second Empire 1852–1870, and Third Republic 1871–1940).

From the linguistic point of view, the most interesting activity of the Napoleonic initiatives was the standardized collection of a great number of dialectal translations of the *Parable of the Prodigal Son* undertaken by Charles Etienne Coquebert de Montbert (1755–1831) and his son Barthélémy Eugène (1785–1847; see further, Pop, 1950, *passim*). Charles Etienne should not be considered a linguist as such, but rather a statistician in the spirit of the eighteenth century who was inspired by a wide-ranging empirical curiosity.

Painstaking collection of a wide range of empirical data continued further under the re-established monarchy and gave rise to a very sophisticated investigation of the national space of France. The collection of the data was, naturally, based on the departmental grid.

However, the most crucial problem was the quantitative processing and subsequent visualization of the data collected. In this respect, from throughout the nineteenth century one can find scholarly treasures, which unfortunately are little known outside France: see for instance Gilles Palsky's (1996) superb *Des Chiffres et des Cartes*, which not only contains a good description of the general evolution of these investigations but also a great number of excellent photographic reproductions of the respective visualizations.

As a result, one can assume that at the end of the nineteenth century practically all political and intellectual figures in France had good familiarity with graphics and visual representations giving evidence of the geographical arrangement of a very great number of social, economic, and demographic *mechanismes* of national interest.

7.3 Historical Background B: Jules Gilliéron and ALF (*Atlas linguistique de la France*)

Jules Gilliéron (1854–1926) was born in the French-speaking part of Switzerland and moved permanently to Paris at the age of 22 (Pop and Pop 1959: 5–19). He rapidly came into contact with eminent personalities of the Parisian academic scene. From 1883 onward he was charged with the teaching of “*dialectologie de la Gaule romane*” at the École Pratique des Hautes Études, a position in which he continued until his death. What he encountered in Paris from a geo-linguistic point of view was a great number of unsolved questions referring to the linguistic subdivision of France in the present and in the past, and many complaints about insufficient data in this connection. His chief intellectual mentor was the great French philologist Gaston Paris (1839–1903). One of the major concerns in the 1870s and 1880s was the question of whether the intellectual concept of dialect had a real counterpart in reality and could

therefore “exist.” In France, the general opinion denied the existence of dialects but fully recognized the real existence of a great number of single linguistic features. Gaston Paris summarized this fact in 1888 as follows: “Il faut faire la géographie non pas des *dialectes*, mais des *traits linguistiques*” (“We must make the geography not of dialects, but of linguistic traits”). Unfortunately, there was no precise knowledge about the exact geographical extent of such a great number of different dialectal features. This gap would be filled by ALF.

Before starting the work for ALF, Gilliéron exercised his methodological ideas in the Swiss canton of Valais, publishing in 1881 a small phonetic atlas, which could be considered, to some extent, as a forerunner of ALF.

However, what he prepared between 1881 and 1897 was quite different from this little test piece. When the real fieldwork for the future ALF started in 1897, Gilliéron had already elaborated a precise research agenda:

- *Theoretically*: the great challenge was to determine the geographical range of a large number of basilectal geolinguistic features belonging to different linguistic categories (phonetics, morphology, vocabulary, etc.).
- *Practically*: the inquiries should be done *in loco* by contacting bilingual people (French and local dialect) and observing two strict principles. The fieldworker, Edmond Edmont, was told (a) to transcribe only the first answer given by the interviewee, and (b) to avoid any “extortion” of further (multiple) responses. The challenge was therefore to elicit only the *basilectal component* of the multiple competence of the interviewees.

These two constraints had a profound impact on the quality of the ALF data, and guaranteed its perfect commensurability. In just four years (1897–1901), Edmont succeeded in visiting 638 localities spread evenly over the Romance-speaking parts of France, Belgium, and Switzerland, and neighboring areas (the Channel Islands and Piedmont, Italy) where Galloromance varieties were spoken. In his peregrinations he used three questionnaires, starting with a set of 1,421 questions, which was later further enlarged to 1,920 items.

Edmont’s astonishing accomplishment was enhanced by the rapid publication of the collected data between 1902 and 1910. As a result, the completed ALF comprised 10 in-folio volumes with full-text maps in which the reader could find, for three geographically coherent sections (series A: whole grid; series B: southern part of the grid; series C: south-eastern part of the grid), the transcriptions produced by Edmont and minimally corrected by Gilliéron.

Another issue of great importance, as we shall see in the next section, was the parallel publication of blank maps (or *cartes muettes*, “mute maps”) of the ALF grid (with 638 localities, sites, or points) and their diffusion among interested scholars.

7.4 The Practical and Theoretical Importance of ALF

ALF’s contribution to Romance linguistics and philology all over Europe was immediate and substantial. An important factor in this success was Gilliéron’s teaching at the École Pratique des Hautes Études. His classes were attended by a large group of upcoming Romanists from all parts of Europe. The courses Gilliéron gave became legendary (Pop and Pop 1959: 53–63). This was true also of his publications, which resonate with a great personal commitment and are full of innovative linguistic ideas (for an excellent example, see Gilliéron 1918).

One of the main pillars of Gilliéronian linguistics was the systematic study of geolinguistic feature areas. While working with the aforementioned mute maps in order to exploit the raw data transcriptions of single ALF maps, Gilliéron (like other Romance scholars, e.g., Jaberg, 1908) noticed that the diffusion areas of different linguistic features could vary

considerably according to size, shape, and geographic location. So as to understand and explain this (initially very strange) variability, Gilliéron developed a special methodology called *aréologie*, in which the different *aires* are regarded as results of processes of diffusion, retraction, and resistance, all of which result from the metalinguistic actions of dialect speakers. We re-encounter here the old idea that relations in space are the outcomes of specific human behavior.

In according “metalinguistic responsibility” to dialect speakers, Gilliéron developed a psycholinguistic theory, whereby factors such as linguistic creativity and the management of homonymy had pivotal roles. Note that Gilliéron utilized and analyzed the ALF data by looking exclusively at single atlas maps and avoiding any data synthesis.

Let us, however, return to the use of mute maps, which also played a central part in the new geolinguistic conceptions. The use of mute maps always required the following steps:²

- the choice of a classification criterion in order to extract and visualize some specific properties from the raw data of a given atlas map,
- the choice of the visualization mode: signatures that are either *spatial* (i.e., using areas) or *linear* (using isoglosses),
- the projection of the selected signatures (in color or in black and white) onto the blank form of the mute map.

Obviously, the graphic quality of such cartographic exercises could vary considerably: the maps could be produced for personal study or for publication, and they might assume different cartographic forms depending on the geolinguist’s drawing ability. What is really important, however, is the fact that following the publication of ALF it was clear that no one would be able to avoid intelligent work with mute maps. In Germanic philology, incidentally, linguistic atlases never offer their data in raw form but instead present them in the form of symbol maps. These are just a particular classification of the raw data,³ which often remain completely invisible or inaccessible to the user. In such circumstances, it is obviously difficult to develop clear ideas about the classification of atlas data.

Whereas Gilliéron’s teaching systematically neglected diachrony, many of his Swiss, German, and Austrian followers used the ALF data for diachronic studies, starting from two questions: (a) what were the sizes, shapes, and geographic locations of ALF feature areas in the past? and (b) is it possible to extract, from medieval data, comparable areal information so as to reconstruct diachronic evolution over the course of two or three centuries?

From our dialectometric viewpoint, Gilliéron, his ALF, and the newborn ALF geolinguistics are highly important for several reasons: for the excellent metrological quality of the ALF data, for the revitalization of the idea that spatial relations depend upon human behavior, and for data classification based on mute maps.

Unfortunately, all these advantages are scarcely known outside Romance philology. Another particularity of Romance linguistics and philology is worth mentioning, that of the tight link established between linguistic geography and a wide range of sub-branches of Romance philology (text philology, historical grammar, lexicology, and etymology).

ALF’s example was rapidly imitated (see Chambers’ chapter on written surveys, this volume) and applied to other great Romance-speaking domains such as Italy (AIS, 1928–1940, created by the Swiss scholars Jakob Jud and Karl Jaberg), Romania (ALR, 1938–1942, under the responsibility of Sextil Pușcariu, Sever Pop, and Emil Petrovici), and Catalonia (ALC, 1923–1964, under the direction of Antoni Griera). Fortunately, the basic methodological principles and assumptions of ALF were not altered in the process.

7.5 From ALF to Jean Séguy's *Dialectométrie*

After World War II, regrettably, the empirical guidelines observed by Gilliéron when compiling the ALF were slowly forgotten in France. The metrological status of ALF as a kind of “glotto-geodesy” of France was replaced by the rather uncritical search for new, “naturalistic,” and linguistically attractive data. The nationwide perspective of ALF stimulated the appetite for analogous but regional perspectives, but the rather unspecific character of the ALF questionnaire engendered the desire for regional questionnaires with (much) more specific items. Furthermore, standardized inquiries were renounced for the sake of more authenticity and naturalism.

One of the first Parisian scholars of Gilliéron, Albert Dauzat (1877–1955), tried at the end of the 1930s to meet all of these requirements in a new research project called *Nouvel Atlas Linguistique de la France* (NALF), attempting above all to maintain the Gilliéronian demand for commensurability in the collected data. For the sake of NALF—which, from the beginning, meant a “family” of linguistic atlases each related to single historical regions such as Normandy, Provence, Gascony, and so on—Dauzat recommended that the different questionnaires should have approximately the same extent, and that a third of the items they contained should be identical.

Owing to the appearance of a new generation of geolinguists who no longer respected (or understood?) either Gilliéron’s or Dauzat’s principles, the new regional atlases, whose compilation started at the beginning of the 1950s, evolved in a completely different direction. They were characterized by the following principles: a highly regionalized choice of items for (NALF) questionnaires; the establishment of new inquiry grids, very often to the complete exclusion of the old ALF sites; the de-standardization of the data collection by applying guided or completely free conversation with informants, instead of standardized questioning; the conscious elicitation of multiple responses by “squeezing” the multiple competences of the informants; and the use of local dialects during the inquiries, instead of normal French.

It is obvious from the viewpoint of Gilliéronian glotto-geodesy that these data collection methods no longer had the same quality as those of ALF. The new regional linguistic atlases did not evolve toward the production of a second, more detailed layer of glotto-geodetic analyses of different parts of France that would allow comparative insights into the geolinguistic dynamics of a period of about 50 years. Instead, they went in the direction of compiling large-scale geographically stratified regional vocabularies.

At the end of the twentieth century the complete list used for NALF (which was later renamed to the *Atlas Linguistique de la France par Régions*) contained 25 titles, the investigation grids of which cover the whole territory of France.⁴ Without doubt, they all have considerably enlarged our knowledge of many French regional dialects. But they cannot be used for any large-scale comparison. Across the 25 questionnaires there are not even ten items which they all have in common. In short, the NALF atlases reflect a completely different geolinguistic methodology that abandons the requirement that the data it produces should be commensurable. It is no longer glotto-geodesy as per ALF, but data collection for its own sake.

One of these new regional atlases was the *Atlas Linguistique de la Gascogne* (ALG) compiled by Jean Séguy (1914–1973). ALG was elaborated and published in six folio volumes between 1954 and 1974. Séguy started ALG fully aware of the directives given by Dauzat, but on the way yielded further and further to what we might call “data collection syndrome.” Finally, while trying to exploit the inner structures of ALG’s data, he neared desperation, confronted as he was by the endless variability in the size, shape, and geographic location among the different feature areas on ALG’s mute maps. This was the moment of his conversion to quantitative thinking and the beginning of his efforts to bring order to this apparent chaos.⁵

Séguy's truly dialectometrical writings consist of two articles (Séguy 1971, 1973) and some interesting pages and maps at the end of the sixth volume of ALG. In these writings he deals—always in a very elementary way, and using rudimentary graphic devices—with some problems of quantification with respect to measuring linguistic distances (not similarities!) between two neighboring sites, for which he analyzed data taken not only from ALG but also from other Romance linguistic atlases. In spite of these methodological weaknesses, Séguy's position is nevertheless seminal, in the following sense. Behind (or below) the apparent chaos of the ALG data and that of other atlases, he conjectured the existence of an underlying primary order which could be explored and expressed in quantitative terms. Séguy assumed the existence of a quantitative relation between geographic and linguistic distances, and conducted simple experiments with it, attempting to apply distance formulae taken from general statistics to ALG data and those from other linguistic atlases. In 1973, he created the neologism *dialectométrie*.

7.6 Regensburg-Salzburg Dialectometry (RS-DM): Theoretical Assumptions and Practical Achievements

My own dialectometric work began at Regensburg University, to which I was affiliated between 1973 and 1982, and continued later in Salzburg. Unfortunately, Séguy's early death in 1973 prevented any collaboration. I started from the following theoretical assumptions.

The continuous changing of geolinguistic feature areas according to their size, shape, and geographic location seems to be a linguistic universal. Moreover, this fact accounts for the (often bemoaned) non-coincidence of isoglosses. It also seems to be the motivation for the claim made by many linguists that "each word has its own history,"⁶ which, for the sake of linguistic geography, should be translated "each *feature* has its own *area*." In a given geolinguistic dataset the total number of areas seems to play the same role as the totality of words in a given text. It would thus be interesting to compare the respective findings in quantitative linguistics.

The inner structure of different linguistic atlases (like ALF, AIS, and others) depends on the interplay of a large number of single feature areas belonging to different linguistic categories (phonetics, morphology, lexis, etc.). If there are regularities in the spatial distribution of these areas, they can only be found via the scrutiny of a large set of areas. Given the very visual character of (Romance) linguistic geography, the central heuristics of the new investigations must continue to be essentially cartographic, obviously using quantitative instead of qualitative cartography.

At the same time (1973 ff.), the practical considerations were as follows:

1. There were some new calculation machines called *computers*, which seemed to be useful for the treatment of mass data.
2. Are there methods of quantitative classification in other fields with similar statistical needs and experiences?
3. Is there previous experience with the quantitative mapping of a great amount of geographically-dispersed numerical data?
4. Can the new *computers* also be applied to cartography?

With respect to (1), I succeeded very quickly in establishing collaborations with computer specialists by avoiding any *bricolage* and the relative shortcomings of do-it-yourself solutions. Secondly, I soon learned of "numerical classification" (*numerische Klassifikation, classification automatique*), and could even come into direct contact with some of its representatives (e.g., H. H. Bock, R. R. Sokal).⁷ It was also very useful to examine interdisciplinary

case studies, mainly in the field of biology, economics and quantitative geography (see Haggett 1965). In addition, I came into contact with the German quantitative linguist Gabriel Altmann at Bochum University, and some of his students.

The answer to (3) was positive. In the early 1970s there already existed good handbooks and reviews on quantitative cartography written in German, French and English (e.g., Dickinson 1973), in which the statistical, visual, and cartographic problems associated with the making of quantitative maps were extensively discussed.

As for (4), the use of computers in cartography did not begin until the end of the 1970s. The integration of shadings, hatchings and colors into this process did not proceed without problems.

7.7 Some Methods and Findings of the RS-DM⁸

I will now briefly discuss four important methods used by RS-DM⁹ using four cartographic examples.

7.7.1 Step A: Taxatation and Presentation of Plate 1

The raw data consist of 626 original ALF maps from which we have extracted—by means of phonetic, morphologic, and lexical taxatation—1,681 “working maps” (WM) (See Figures 7.1 and 7.2, and Plate 1 (map)).

The main goal of taxatation is the areal classification of the original maps on the basis of specific linguistic criteria. In our case, these criteria belong to Romance historical phonetics and lexicology. In the 1,681 WMs we find 18,047 feature areas (and therefore as many linguistic elements called *taxates*) that vary greatly in size, shape, and geographic location. As to their inner fragmentation, see the histogram in Figure 7.2., which shows that there are many roughly structured WMs but few WMs of fine granularity.

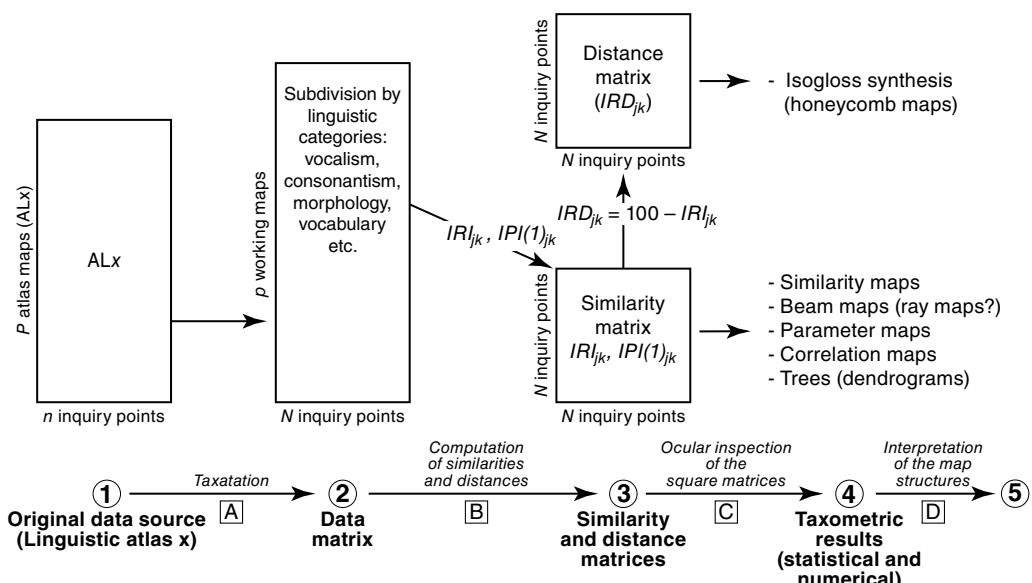


Figure 7.1 Flow chart of the methods used by the Regensburg-Salzburg-dialectometry.

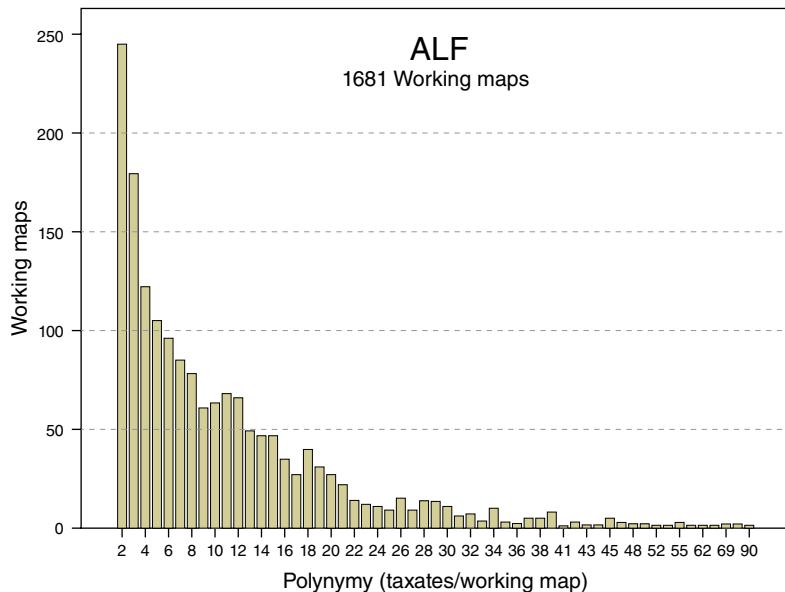


Figure 7.2 Histogram of the total ALF-corpus, showing the granulation and frequency of 1681 working maps (WMs) belonging to all linguistic categories. For better understanding: the granulation of the WMs starts with 2 taxates/WM (valid for 245 WMs) and ends with 90 taxates/WM (valid for 1 WM).

The exponential-like regularity of the decreasing curve has been found in all our dialectometric analyses.¹⁰ It seems to be a direct consequence of the collective dialectal behavior of the French speakers. As early as 1985, Altmann described this situation as the result of the antagonistic interplay of birth and death processes, which were obviously related to the adoption and atrophy of linguistic features, which, in the present case, are associated with areas.

We conjecture that these spatially related regularities have the same law-like status as the well-known *Lautgesetze* (sound laws), discovered by the Neogrammarians of the end of the nineteenth century, whose regularities are related to time.

Plate 1 was established on the basis of the ALF map 18, *l'aile* ("the wing"). It shows 11 different outputs (*taxates* and their *areas*) of the final Latin -A (in the Latin etymon ÁL[A]). This etymon occurs at all 641 ALF sites. The most frequent output (at 350 ALF sites) is the taxate *zéro* (Ø, i.e., the Latin final -A vanishes completely). The other ones are -à (at 70 ALF sites) and -ó (at 69 ALF sites), and so on, as shown in the legend.

The changing granularity of the WMs is called *polynymy*. Theoretically, it ranges from 2 to N (the maximal number of localities in a given grid). Plate 1 is therefore 11-nym: it is a map of medium granularity from which we can derive, in the present ALF taxatation, 67 other specimens.

The geographic structure of the map in Plate 1 clearly shows the north-south division of the Gallo-Romance domain. It suggests that the Northern taxate 1 (zero) expanded at the expense of -a or -o, which are closer to their Latin origin. By including diachronic information derived from the different parts of the Gallo-Romance domain, the linguistic interpretation of the map profile can be further refined and completed.¹¹

The data matrix in Figure 7.3 shows that every character vector of an atlas site consists of areas of different sizes. It could thus be useful to analyze this variation quantitatively. It is

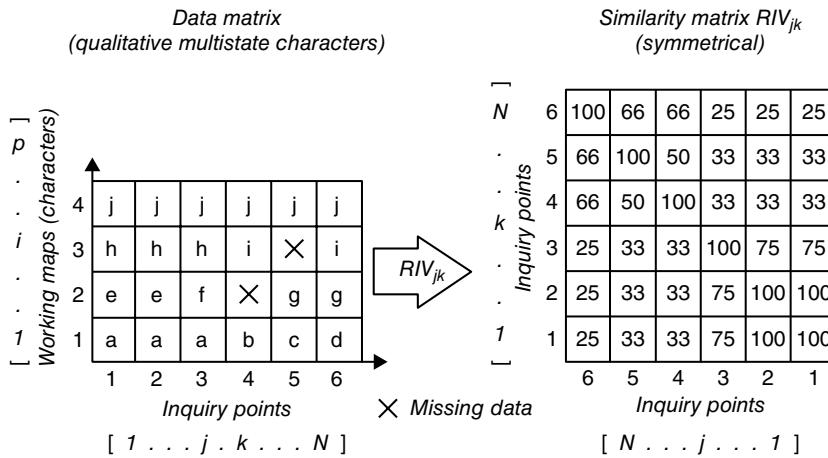


Figure 7.3 Data matrix and similarity matrix. Scheme of calculation of the interdialectal similarities via RIV_{jk} (Relative Identity Value).

necessary, obviously, to organize the data matrix according to different linguistic categories if we wish to compare the different computations with one another.

Since the 1990s we have used VDM (*Visual DialectoMetry*), a computer program created and further improved by Edgar Haiderl, for managing all the steps of the numerical and visual processing of our data.¹² The dialectometric basics implemented in VDM are drawn from our handbook (Goebl, 1984).

7.7.2 Step B: Computation of Similarities and Distances

The main result of step B is the change in the ontological nature of the investigated data, namely from *qualitative* to *quantitative* (see Figure 7.3). Statistically speaking, this happens by measuring the similarities between the N locality vectors. As the handbooks of numerical classification offer a variety of similarity indices, one should select an index that fits one's concepts of interdialectal similarity.

In this instance, the Relative Identity Value (RIV_{jk}) has proven very successful. It is calculated using the number of pairwise matchings (also called co-identities, or COI) and the number of pairwise mis-mッチings (co-differences, COD) of taxates. RIV_{jk} values range between 0 and 100%, according to the formula at (1):

$$\text{RIV}_{jk} = 100 \sum \text{COI}(i)_{jk} / \sum \text{COI}(i)_{ik} + \sum \text{COD}(i)_{jk} \quad (1)$$

Table 7.1 shows the meaning of these symbols.

The calculated similarity values will be stored in the (square) similarity matrix (see Figure 7.1, point 3), which can easily be converted into a distance matrix by applying the following formula: $distance(RDV) = 100 - similarity(RIV)$. The two matrices are depositories of the totality of the pairwise relations (be they similarities or distances) that exist between the N investigated sites.

Because the square matrices consist of two symmetrical halves and the scores located along the diagonal are irrelevant, the number of the valid similarity or distance values is $N/2(N-1)$.

In the case of ALF, we are faced with 638 original sites that have been augmented by three artificial sites corresponding to the standard languages French, Italian, and Catalan. Thus, we have 641 sites and 205,120 similarity or distance values for further processing.

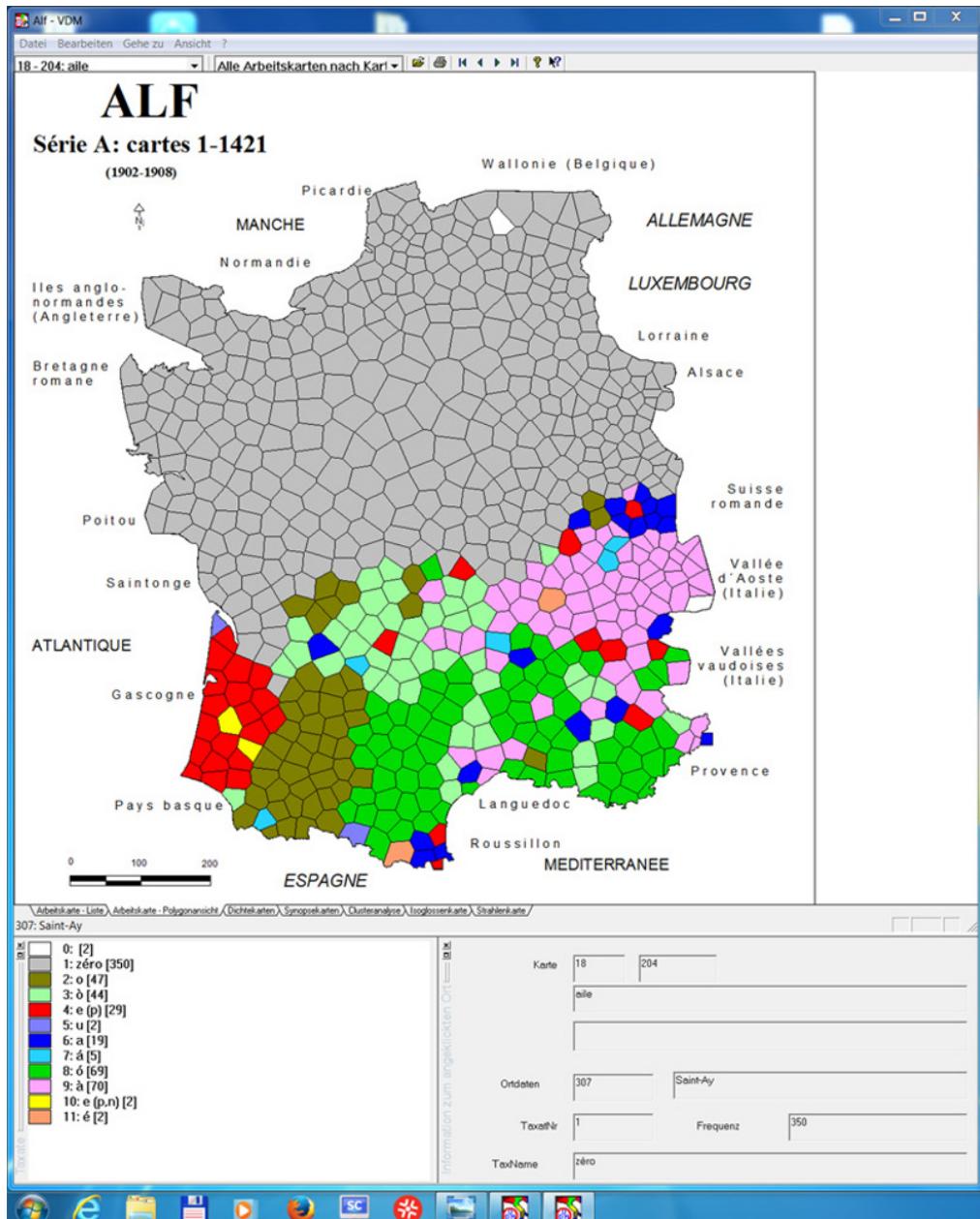


Plate 1 Sample of a *phonetic working map*: spatial distribution of the Gallo-Romance results of final -A in the Latin etymon ÁLA (< Fr. *aile*) ‘wing’ (following ALF 204 *aile*). Cartographic status: qualitative choropleth map. See Section 7.7.1, pp. 129–131. (See insert for colour representation of the figure.)

Table 7.1 Meanings of the symbols used in the formula at (1).

Symbol	Meaning
$COD(i)_{jk}$	co-difference between two taxates (on the map i and for the sites j and k)
$COI(i)_{jk}$	co-identity between two taxates (on the map i and for the sites j and k)
i	one of p working maps
j	reference site
k	site (atlas point) to be compared with the reference site j
p	total number of the working maps in the data matrix
N	total number of the sites (atlas points) in the data matrix
RIV_{jk}	Relative Identity Value (between the attribute vectors of the sites j and k)
RDV_{jk}	Relative Distance Value (between the attribute vectors of the sites j and k)

7.7.3 Step C: Visualization

According to the cartographic tradition of linguistic geography, the subsequent processing of the calculated data has to be visual. The data output thus moves from numerical to visual. The cartographic status of the new maps or schemes will no longer be *qualitative* (as in Plate 1), but henceforth *quantitative* (as in Plate 2, Plate 3, and Plate 4).

Obviously, this change should be done algorithmically. In this respect, the quantitative branch of thematic cartography offers a series of very useful solutions, which have been partly incorporated into VDM (see Dickinson 1973 *et passim*, and Goebel (1984, 86, 113)).

VDM provides different choropleth and isopleth maps and some dendrographic schemes (trees), which all use colors. Here we discuss only the cartographic details and the dialectometric status of the following three map types: similarity maps (see Plate 2, and Goebel 1984 I, 114), parameter maps (see Plate 3, and Goebel 1984 I, 136) and isogloss synthesis (see Plate 4, and Goebel 1984 I, 183, *et passim*). For the discussion of the remaining map types (beam maps, trees, correlation maps), see Goebel (1983; 1984, 172; 2005a).

7.7.4 Step D: Presentation and Discussion of Plate 2, Plate 3, and Plate 4

7.7.4.1 Similarity Map

From the statistical point of view, every similarity map relies on one of the N (=641) vectors of the similarity matrix (see Plate 2). There are consequently 641 similarity maps that may be generated and compared to one another. Of the 641 RIV_{jk} values, 640 have been visualized. The reflexive score $RIV_{307,307}$ (= 100) has been disregarded: the polygon in question belongs to the reference site (here, ALF P. 307), and it therefore remains blank.

The cartographic prerequisites are as follows. Cartographically speaking, similarity maps are quantitative choropleth maps, which allow us to visualize a quantitative pseudo-continuum. The inner numerical variation of the N vectors of the similarity matrix should be adequately represented by the correspondingly variable visualization.

The spatial repartition of the different graphic steps should be done by appropriate interval algorithms. For optimizing the visual recognition of spatial patterns, having free choice over the number of graphic intervals is indispensable. The base map should be polygonized according to Voronoi geometry (Okabe, Boots, and Sugihara 1992).

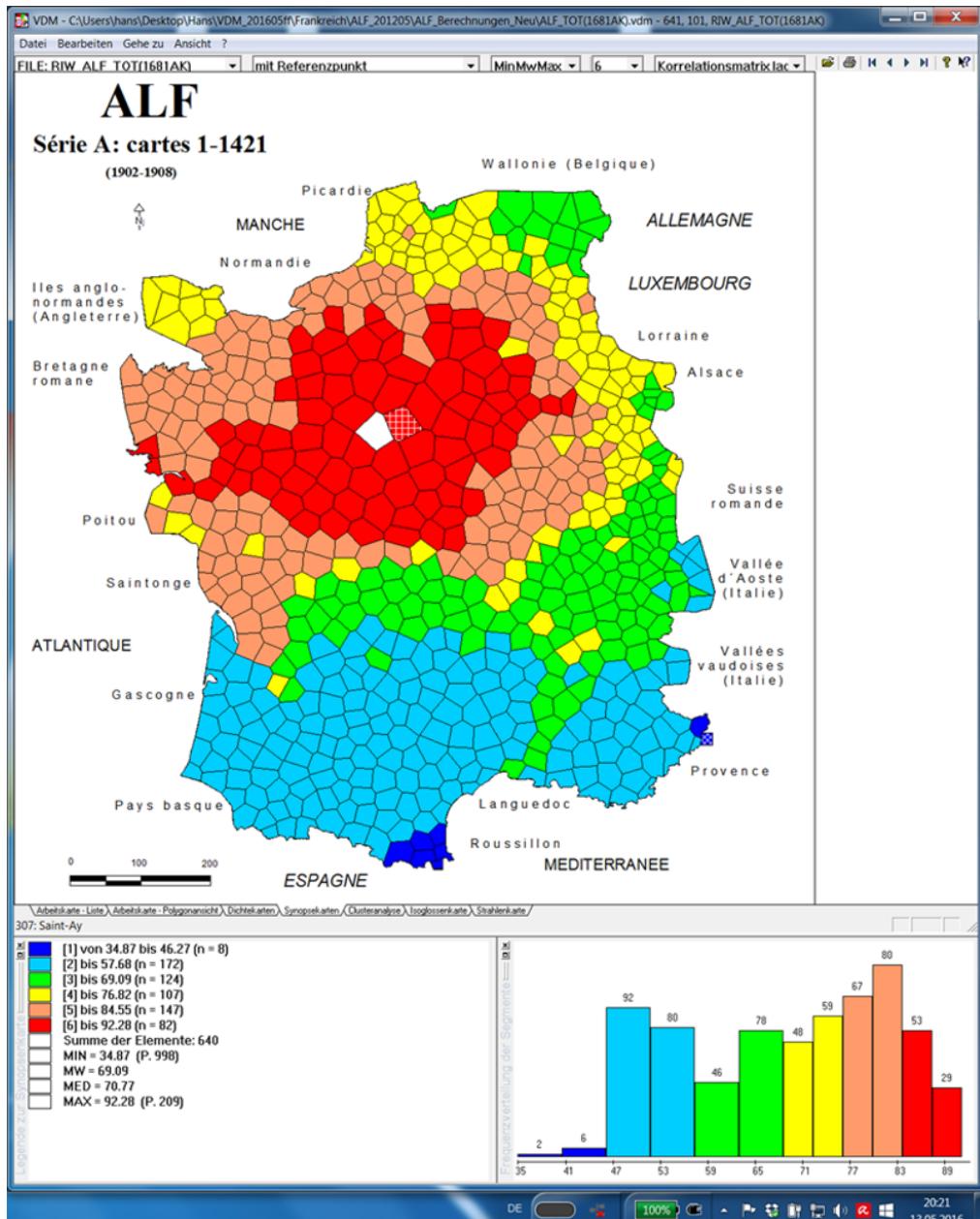


Plate 2 Sample of a similarity map: spatial distribution of the similarity values referring to ALF-point 307 (Saint-Ay, Département Loiret). Similarity index: RIV_{307,k}; corpus: 1681 working maps, all linguistic categories; algorithm of visualisation: MINMWMAX 6-tuple. Cartographic status: quantitative choropleth map. See Section 7.7.4, pp. 133–139. (See insert for colour representation of the figure.)

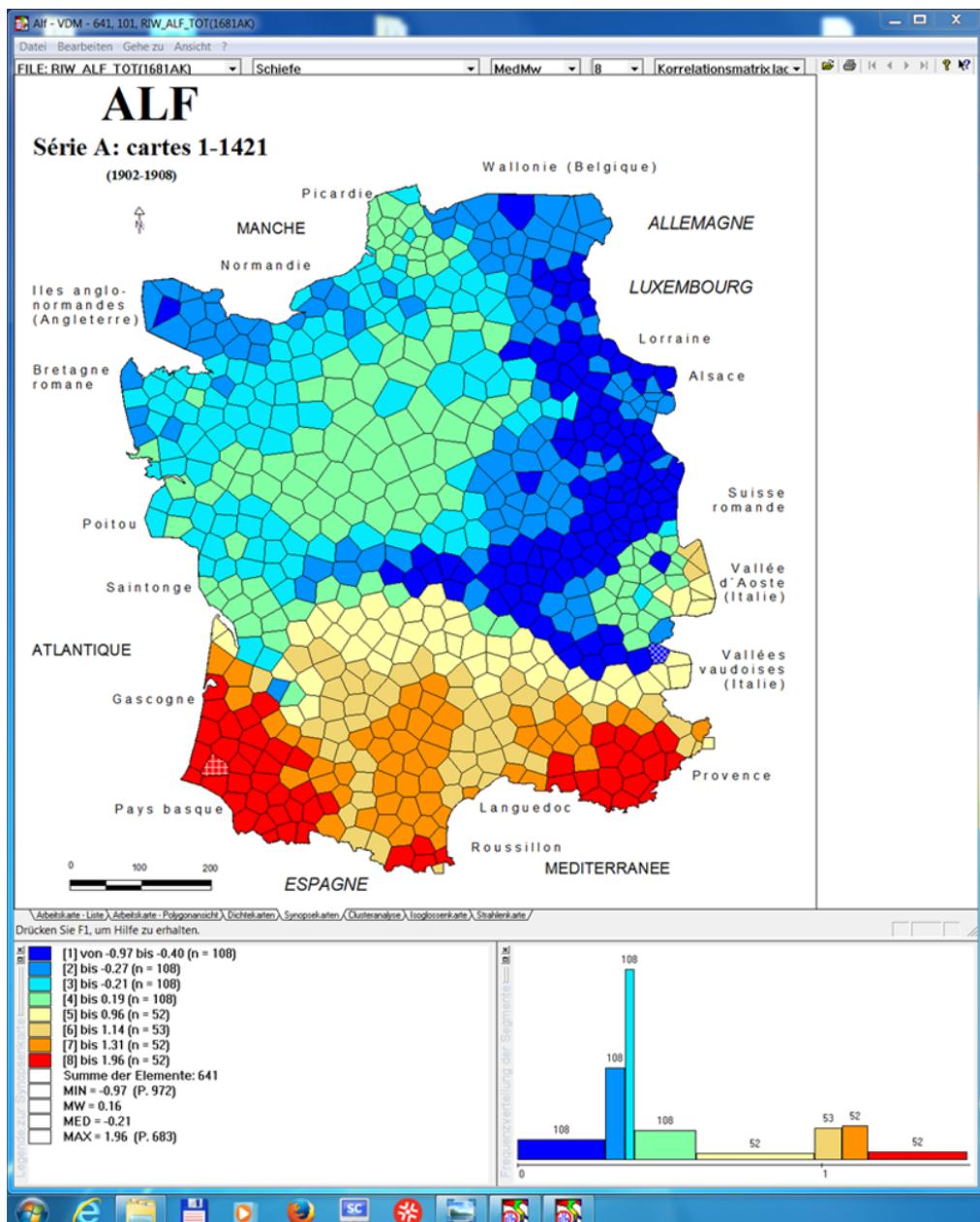


Plate 3 Sample of a parameter map: synopsis of 641 skewness values (according to the asymmetry index of R. A. Fisher). Similarity index: RIW_{jk}; corpus: 1681 working maps, all linguistic categories; algorithm of visualisation: MEDMW 8-tuple. Cartographic status: quantitative choropleth map. See Section 7.7.4, pp. 133–139. (See insert for colour representation of the figure.)

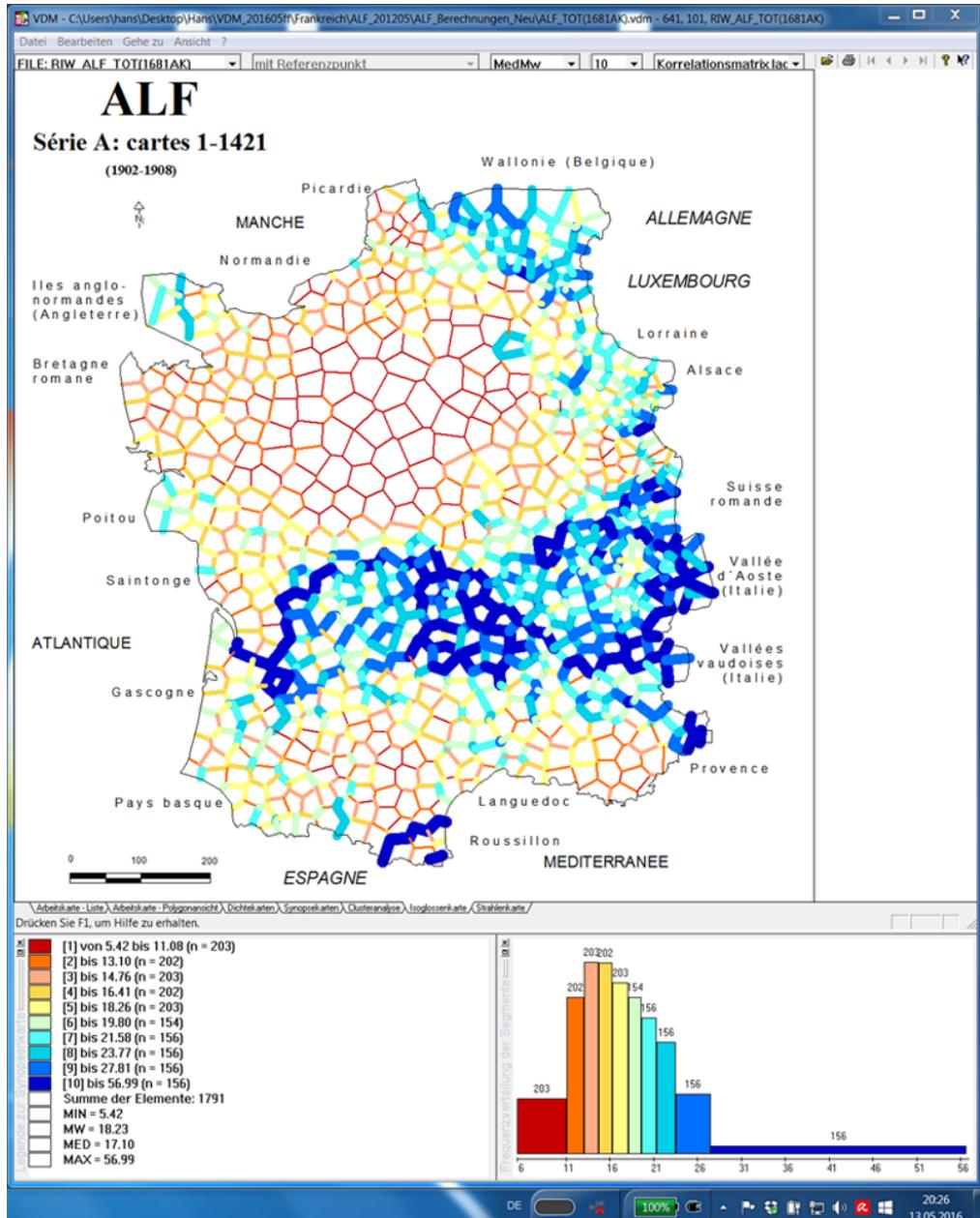


Plate 4 Sample of a interpoint map (honeycomb mode): synopsis of 1791 distance values (according to RDV_{jk}). Distance index: RDV_{jk}: corpus: 1681 working maps, all linguistic categories; algorithm of visualisation: MEDMW 10-tuple. Cartographic status: quantitative isarithmic (or: isopleth) map. See Section 7.7.4, pp. 133–139. (See insert for colour representation of the figure.)

For all these requirements the VDM program offers adequate solutions. The number of intervals is always even. The interval algorithms used in this article (MINMWMAX and MEDMW) always distribute N or N-1 dialectometric scores on the two sides of the arithmetic mean (cf. Goebel 1984 I, 93). MINMWMAX gives the intervals located above and below the arithmetic mean the same numerical width, whereas MEDMW gives the intervals located above and below the arithmetic mean the same size (i.e., number of polygons). MINMWMAX has been applied to Plate 2, MEDMW to Plate 3 and Plate 4.

In practice, the current use of visualization techniques depends deeply on visual training and experience, and belongs to what is nowadays called “imaging.” It is always recommended that the user begin the contemplation of a choropleth map with the polygons of the highest-ranking intervals before proceeding to the polygons of lower-ranking intervals. The choropleth structure of the map shows a regular decrease in the dialectal similarities in space. The same phenomenon can be observed in all the remaining N-1 similarity maps of the same similarity matrix, as well as when using other similarity indexes.

Obviously, the regularity of the decrease in similarity from the reference site to the borders of the map is independent of the personal linguistic experience of the speakers themselves. This decrease must be explained with reference to the synergy of the forces of all language laws that determine the linguistic nature of the 641 locolects of the ALF grid.

The similarity maps have different (geo)linguistic meanings. To a Romance linguist, the stratification of the (warm and cold) colors on Plate 2 is characteristic: the sum of the zones in intervals 4–6 (warm colors) corresponds (with the exception of Wallonia in the North) to the influence area of the *Langue d’Oil*, whereas the sum of the polygons in intervals 1–3 (cold colors) constitute the catchment area of the *Langue d’Oc* (or Occitanian). Moreover, a similarity map indicates the (relational) position of a given locolect in the middle of a given grid.

Note that these analyses can be repeated for different linguistic categories. In this case, the shapes of the various similarity profiles can vary somewhat. However, these differences are mostly rather slight (see Goebel, 2002, 2003, 2006a, 2010, and 2013a).

We have also observed that the overall structures of all our dialectometric visualizations remain practically unaltered once we have amassed a set of 200–300 randomly selected WMs. The same is true by using WM corpora with *little*, *middle* or *big* polynomy (see Figure 7.2). Similarity maps can also be interpreted from a *diffusionist* perspective, as they thereby show—metaphorically speaking—the results of diffusional effects of one element of the whole group.

Another metaphorical interpretation would be a *communicative* one. The algebraic logic of RIV_{jk} corresponds to the network of connections of telephone calls. Because any single co-identity (COI) would represent a single telephone contact and any co-difference (COD) the opposite (i.e., no contact), it follows that every similarity map relying on RIV_{jk} can be interpreted as analogous to the telephone activity of a given subscriber to a telephony service. From this perspective (see Plate 2), one can see that the intensity of the ‘telephone connections’ of ALF point 307 is rather circular, and decreases rapidly to the North (Wallonia), the South, and the South-East of the grid.

7.7.4.2 Parameter Map

The legend and histogram in Plate 2 give all the information necessary for the statistical understanding of the respective similarity distribution (minimum 34.87; arithmetic mean 69.09; maximum 92.28, etc.; see Plate 3). The shape of the histogram shows a rather symmetrical frequency distribution with two modes. Of course, for the remaining 640 similarity distributions all of these values may be different.

A systematic check of these numerical variations in fact shows that it is possible to interpret them linguistically. In this regard, the study of the symmetry of the similarity distributions

is of particular interest. The legend of Plate 2 shows that 336 (= 107 + 147 + 82, or 52.5%) of 640 RIV values lie above the arithmetic mean (69.09). Again using a communicative metaphor, one could argue that the “dialecticity” of ALF P. 307 is related well to the rest of the grid, whereas the contrary would be true if, say, 60% of the RIV values lay below the arithmetic mean. In this respect, one of the most suitable indices we can use to grasp the symmetry of a frequency distribution is its skewness (for the formula, see Goebel 1984 I, 150).

The construction of Plate 3 depends upon the following procedure: (a) computation of the 641 skewness values, (b) mapping of these values, and (c) linguistic evaluation of the new choropleth pattern. What is immediately striking about Plate 3 is its self-explanatory structure: in the North, East, and West the polygons in intervals 1 and 2 (dark and middle blue) form two circular, or pincer-shaped, patterns with a very clear spatial distribution. The large “circle” surrounds the whole *Domaine d’Oil*, whereas the domain of *Franco-Provençal* (on the South-eastern periphery) is located between the two “jaws” of the pincer, which has its pivot to the West of Lyon. In the South, we have three “bulwarks” (Gascony, Languedoc, and Provence, all in red (interval 6), which are linked together by polygons in intervals 4 and 5. This geographical structure, together with our knowledge of the linguistic evolution of France and some statistical ideas, allows us to confer the following linguistic meaning on the different hatchings and shadings of the map.

- *Polygons in interval 1 or 2 (dark and middle blue)*: these are zones of great “linguistic compromise.” That is, they are dialecticities with a high percentage of many large-sized (“mega-choric”) linguistic attributes (and their areas);
- *Polygons in interval 6 (red)*: these represent zones of minimal linguistic compromise. They are, in other words, dialecticities with a high percentage of small- (“oligo-choric”) and medium-sized (“meso-choric”) linguistic features (and their areas).

Note that by the term “linguistic compromise” (German *Sprachausgleich*) we mean a specific ratio of the entanglement of oligo-, meso- and mega-choric areas as a result of local or regional language contact and conflict.

The two circular configurations in intervals 1 and 2 are, on the one hand (referring to the *Domaine d’Oil*), the visible result of the secular expansion of the linguistic type of the *Langue d’Oil*, and on the other hand (in relation to *Franco-Provençal*), the visible result of the retreat of the old Latin linguistic heritage of Lugdunum/Lyon, with all its concomitant linguistic consequences, pressed as it was on two sides by the *Langue d’Oil* in the North and by the *Langue d’Oc* in the South. By contrast, “aggregates” in interval 6 (red) in the South are real zones of linguistic isolation, whereas between them there are some weak flows of linguistic compromise (see the polygons of intervals 4–5).

The choropleth structure of Plate 3—which very clearly indicates dynamic irradiation in the North and a punctual (block-like) resistance in the South—is of great importance for the history of the whole Gallo-Romance domain. Similar structures can be found in all our dialectometric analyses.

7.7.4.3 Interpoint Map

Cartographically speaking, Plate 2 and Plate 3 were based on *areas* (polygons). Their iconic message depends upon the visual interplay of continuous surfaces, whereas the oldest cartographic schemes in linguistic geography, viz. the drawing of isogloss bundles, is based on the combining of *lines* crossing the space discontinuously. The dialectometric modeling of the combined drawing of isoglosses is quite straightforward. Cartographically, the drawing of isoglosses follows the edges of the polygons in the Voronoi base map (see Plate 4). The drawing of isogloss bundles is done by changing the thickness and coloring

of the polygon sides. From the statistical point of view, the values to be visualized are *distances*, rather than *similarities* ($RDV_{jk} = 100 - RIV_{jk}$). They can be extracted from the distance matrix according to the neighborhood geometry of the Voronoi base map.¹³

However, the line-based isopleth map presents new visual challenges, because the eye has to capture a discontinuous image syntax. In Plate 4 we applied a visualization using ten graphic steps, the interval algorithm MEDMW, and a medium-sized thickness span for the polygon edges. The number of visualized distance values, and therefore polygon edges, is 1,791, a number which corresponds to 8.73% of the total content (205,120 RDV values) of the distance matrix. The taxometrical impact of this visualization is thus rather small. The line-based pattern of Plate 4 is nevertheless highly suggestive.

With regard to the spatial compactness of thick and dark blue polygon edges, it appears rather clearly that there are four hotspots on the map. One is in the South, between the Roussillon and the Languedoc. Another is in the middle of the map, between the Domaines d’Oc and d’Oil (note the curvature). A third is located at the Eastern border of the map, and the fourth on the Northern border of the Franco-Provençal area, between the Northern domains of Picardian and Wallonian. By contrast, there are zones of little compartmentalization in the central parts of the Domaines d’Oil and d’Oc (i.e., Languedoc and Provence).

A visual comparison with older French isogloss syntheses (see Rosenqvist 1919 and Ettmayer 1924) reveals a striking resemblance between those pioneering maps (realized, of course, via manual work) and the interpunctual message of Plate 4.

7.8 Final Remarks

The dialectometrical procedures introduced in the present chapter have an exclusively diagnostic and exploratory character, and represent a hybrid compound with elements taken from linguistics, statistics, and cartography. The concurrence of these methods is intended to enhance our knowledge of the structure and function of geolinguistic networks using quantitative means.

Dialectometry, as it has been documented here, is defined programmatically as the *quantitative* branch of classical, atlas-borne and essentially *quality-oriented* linguistic geography. It could be demonstrated, with dialectometrical support, that there exist hitherto hidden and unexpectedly complex spatial patterns, which would imply the existence of genuine “spatial laws” in linguistic atlas data. These patterns exemplify the postulate expressed at the beginning of the chapter concerning the “basilectal management of space by *Homo loquens*,” which is one of the numerous semiotic behaviors of our species.

NOTES

1 This is exactly what we call “area” later on.

2 See Jaberg (1908), who gives an excellent introduction in this kind of analysis.

3 The DSA (*Deutscher Sprachatlas*) is representative in this regard. Its author, Georg Wenker, himself converted a good proportion of his data into maps, which later became canonical.

4 See the historical overview in Winkelmann (2001).

5 One should not overlook the intellectual support given to Séguy by his friend and colleague Henri Guiter (1909–1994); see Guiter (1973).

6 See Malkiel (1967) and Christmann (1971).

7 See Sneath and Sokal (1973), Bock (1974), and Chandon and Pinson (1981).

- 8 See my contributions on dialectometry published between 1981 and 2013. I refer the reader also to the web-based bibliography of my DM writings at https://www.sbg.ac.at/rom/people/prof/goebl/dm_publi.htm (accessed 7 March 2017).
- 9 Nowadays there are two further dialectometrical research centers worth mentioning, these being the Groningen school (see the contributions of Nerbonne, Heeringa, Prokić, and Shackleton quoted below, and Chapters 20 and 23 of this volume), and the Athens (Georgia) school (see the work of Kretzschmar and Schneider, as well as Chapter 3 of this volume). Szemrecsanyi (2013) and Chapter 18 of this volume are also relevant. Note that in Groningen and Athens the basic assumptions of dialectometry differ from those of Regensburg-Salzburg, mainly by omitting the concept of area and stressing the method of sequence comparison by means of Levenshtein distance.
- 10 See our DM contributions related to the following domains: France (1984, 2002, 2003, 2004, 2005b, 2006a, 2007, 2010, 2013a), Italy (1981, 1983, 1984), Iberia (2013b), Catalonia (2013c), England (2007, and Goebel and Schiltz, 1997), and German-speaking Switzerland (Goebel, Scherrer, and Smečka, 2013).
- 11 See Jaberg's (1908) pioneering booklet, and the major synthesis of Brun-Trigaud, Le Berre, and Le Dû (2005), which represents the sum of the Gallo-Romance *aréologie* based on ALF.
- 12 See <http://www.dialectometry.com/dmdocs/index.html> (accessed 7 March 2017).
- 13 Note that the geometric principle of dialectal neighborhood ("interpunctual contiguity") had already been defined by Carl Haag in 1898.

REFERENCES

- Altmann, Gabriel. 1985. "Die Entstehung diatopischer Varianten: Ein stochastisches Modell." *Zeitschrift für Sprachwissenschaft*, 4: 139–155.
- Bock, Hans Hermann. 1974. *Automatische Klassifikation: Theoretische und praktische Methoden zur Gruppierung und Strukturierung von Daten (Cluster-Analyse)*. Göttingen: Vandenhoeck und Ruprecht.
- Brun-Trigaud, Guylaine, Yves Le Berre, and Jean Le Dû. 2005. *Du temps dans l'espace: Explication linguistique de la France (Essai d'interprétation des cartes de l'Atlas linguistique de la France de Jules Gilliéron et Edmond Edmont augmenté de quelques cartes de l'Atlas Linguistique de la Basse-Bretagne de Pierre Leroux)*. Paris: Comité des Travaux Historiques et Scientifiques (CTHS).
- Chandon, Jean-Louis, and Suzanne Pinson. 1981. *Analyse typologique: Théories et applications*. Paris: Masson.
- Christmann, Hans Helmut. 1971. "Lautgesetze und Wortgeschichte: Zu dem Satz *Jedes Wort hat seine eigene Geschichte*." In *Sprache und Geschichte: Festschrift für Harri Meier zum 65. Geburtstag*, edited by Wolf-Dieter Stempel, and Eugenio Coseriu, 111–124. Munich: Fink.
- Dickinson, Gordon. 1973. *Statistical Mapping and the Presentation of Statistics*. London: Arnold.
- DSA. 1927–1956. *Deutscher Sprachatlas, aufgrund des von Georg Wenker begründeten Sprachatlas des Deutschen Reiches in vereinfachter Form begonnen von Ferdinand Wrede, fortgesetzt von Walther Mitzka und Bernhard Martin*. Marburg/Lahn: Elwert, 23 fascicles with 128 maps.
- Ettmayer, Karl von. 1924. *Über das Wesen der Dialektbildung erläutert an den Dialekten Frankreichs*. Vienna: Denkschriften der Akademie der Wissenschaften in Wien.
- Gilliéron, Jules. 1881. *Petit atlas phonétique du Valais Roman (Sud du Rhône)*. Paris: Champion.
- Gilliéron, Jules. 1918. *Généalogie des mots qui désignent l'abeille d'après l'Atlas Linguistique de la France*. Paris: Champion.
- Gilliéron, Jules, and Edmond Edmont, eds. 1902–1910 [1968]. *Atlas linguistique de la France* (10 vols.). Paris: Champion. [ALF]
- Goebel, Hans. 1981. "Éléments d'analyse dialectométrique (avec application à l'AIS)." *Revue de linguistique Romane*, 45: 349–420.
- Goebel, Hans. 1983. "Parquet polygonal et treillis triangulaire: Les deux versants de la dialectométrie interponctuelle." *Revue de linguistique Romane*, 47: 353–412.

- Goebel, Hans. 1984. *Dialektometrische Studien: Anhand Italoromanischer, Rätoromanischer und Galloromanischer Sprachmaterialien aus AIS und ALF* (3 vols.). Tübingen: Niemeyer.
- Goebel, Hans. 2002. "Analyse dialectométrique des structures de profondeur de l'ALF." *Revue de linguistique Romane*, 66: 5–63.
- Goebel, Hans. 2003. "Regards dialectométriques sur les données de l'Atlas linguistique de la France (ALF): Relations quantitatives et structures de profondeur." *Estudis romànics*, 25: 59–120.
- Goebel, Hans. 2004. "Sprache, Sprecher und Raum: Eine kurze Darstellung der Dialektometrie – Das Fallbeispiel Frankreich." *Mitteilungen der österreichischen geographischen Gesellschaft*, 146: 247–86.
- Goebel, Hans. 2005a. "La dialectométrie corrélative: Un nouvel outil pour l'étude de l'aménagement dialectal de l'espace par l'homme." *Revue de linguistique Romane*, 69: 321–67.
- Goebel, Hans. 2005b. "Dialekte und Familiennamen in Frankreich: Ein interdisziplinärer Vergleich mit den Mitteln der Dialektometrie." In *Gene, Sprachen und ihre Evolution: Wie verwandt sind die Menschen – Wie verwandt sind ihre Sprachen?*, edited by Günter Hauska, 68–99. Regensburg: Universitätsverlag Regensburg.
- Goebel, Hans. 2006a. "Recent Advances in Salzburg Dialectometry." *Digital Scholarship in the Humanities* (formerly *Literary and Linguistic Computing*), 21(4): 411–35.
- Goebel, Hans. 2006b. "Warum die Dialektometrie nur in einem roman(ist)ischen Forschungskontext entstehen konnte." In *Was kann eine vergleichende romanische Sprachwissenschaft heute (noch) leisten? Romanistisches Kolloquium XX*, edited by Wolfgang Dahmen, Günter Holtus, Johannes Kramer, Michael Metzeltin, Wolfgang Schweickard, and Otto Winkelmann, 291–317. Tübingen: Narr.
- Goebel, Hans. 2007. "A bunch of dialectometric flowers: A brief introduction to dialectometry." In *Tracing English through Time – Explorations in Language Variation: In Honour of Herbert Schendl on the Occasion of his 65th birthday*, edited by Ute Smit, Stefan Dollinger, Julia Huettner, Gunther Kaltenböck, and Ursula Lutzky, 133–71. Wien: Braumüller.
- Goebel, Hans. 2010. "Dialectometry: Theoretical prerequisites, practical problems, and concrete applications (mainly with examples drawn from the 'Atlas linguistique de la France', 1902–1910)." *Dialectologia*, Special Issue 1, 63–77.
- Goebel, Hans. 2013a. "Dialectometry and quantitative mapping." In *Language and Space: An International Handbook of Linguistic Variation*, Vol. 2. – *Language Mapping*, edited by Alfred Lameli, Roland Kehrein, and Stefan Rabanus, 433–57. Berlin: de Gruyter.
- Goebel, Hans. 2013b. "La dialectometrización del ALPI: Rápida presentación de los resultados." In *Actas del XXVI Congreso Internacional de Lingüística y de Filología Románicas*, vol. 6. (Valencia 2010), edited by Emili Casanova Herrero, and Cesáreo Calvo Rígual, 143–54. Berlin: de Gruyter.
- Goebel, Hans. 2013c. "La dialectometrització dels quatre primers volums de l'ALDC." *Estudis romànics*, 35: 87–116.
- Goebel, Hans. 2013d. "Le Baiser de la Belle au bois dormant où: des péripéties encourues par la géographie linguistique depuis Jules Gilliéron." In *Dialectologie: corpus, atlas, analyses*, edited by Rita Caprini, 61–84. Alessandria: Edizioni dell'Orso [= *Corpus* 12, 2013].
- Goebel, Hans, Yves Scherrer, and Pavel Smečka. 2013. "Kurzbericht über die Dialektometrisierung des Gesamtnetzes des 'Sprachatlases der deutschen Schweiz' (SDS)." In *Vielfalt, Variation und Stellung der Deutschen Sprache*, edited by Karina Schneider-Wiejowski, Birte Kellermeier-Rehbein, and Jakob Haselhuber, 153–75. Berlin: de Gruyter.
- Goebel, Hans, and Guillaume Schiltz. 1997. "A dialectometrical compilation of CLAE I and CLAE II: Isoglosses and dialect integration." In *Computer Developed Linguistic Atlas of England (CLAE II)*, edited by Wolfgang Viereck, and Heinrich Ramisch, 13–21. Tübingen: Niemeyer.
- Griera, Antoni. 1923–1964. *Atlas Lingüístic de Catalunya* (8 vols.). Barcelona: Edicions Polígrafa. [ALC]
- Guiter, Henri. 1973. "Atlas et frontières linguistiques." In *Les dialectes romans de France à la lumière des atlas régionaux*, edited by Georges Straka, and Pierre Gardette, 61–109. Paris: Centre National de la Recherche Scientifique.
- Haag, Carl. 1898. *Die Mundarten des Oberen Neckar- und Donaulandes (Schwäbisch-alemannisches Grenzgebiet: Baarmundarten)*. Reutlingen: Hutzler.
- Haggett, Peter. 1965. *Locational Analysis in Human Geography*. London: Arnold.
- Heeringa, Wilbert. 2004. *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. Groningen: Groningen Dissertations in Linguistics, 46.

- Heeringa, Wilbert, and John Nerbonne. 2001. "Dialect areas and dialect continua." *Language Variation and Change*, 13: 375–400.
- Jaberg, Karl. 1908. *Sprachgeographie: Beitrag zum Verständnis des Atlas linguistique de la France*. Aarau: Sauerländer.
- Jaberg, Karl, and Jakob Jud, eds. 1928–1940 [1971]. *Sprach- und Sachatlas Italiens und der Südschweiz* (8 vols.). Zofingen: Ringier. [AIS]
- Kretzschmar, William, and Edgar Schneider, eds. 1996. *Introduction to Quantitative Analysis of Linguistic Survey Data: An Atlas by Numbers*. Thousand Oaks, CA: Sage.
- Malkiel, Yakov. 1967. "Each word has a history of its own." *Glossa*, 1: 13–149.
- Nerbonne, John, and William Kretzschmar. 2003. "Introducing computational techniques in dialectometry." *Computers and the Humanities*, 37: 245–55.
- Nerbonne, John, and William Kretzschmar. 2006. "Progress in dialectometry: Toward explanation." *Digital Scholarship in the Humanities* (formerly *Literary and Linguistic Computing*), 21(4): 387–97.
- Nerbonne, John, and William Kretzschmar, eds. 2013. "Special Issue 'Dialectometry++'." *Digital Scholarship in the Humanities* (formerly *Literary and Linguistic Computing*), 28(1).
- Okabe, Atsuyuki, Barry Boots, and Kokichi Sugihara. 1992. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. Chichester: Wiley.
- Palsky, Gilles. 1996. *Des chiffres et des cartes: Naissance et développement de la cartographie quantitative française au XIX^e siècle*. Paris: Comité des travaux historiques et scientifiques (CTHS).
- Paris, Gaston. 1888. "Les parlers de France [1888]." In *Mélanges linguistiques*, edited by Gaston Paris, 432–48. Paris: Champion 1909.
- Pop, Sever. 1950. *La dialectologie: Aperçu historique et méthodes d'enquêtes linguistiques* (2 vols.). Gembloix: Duculot.
- Pop, Sever, and Rodinca Doina Pop. 1959. *Jules Gilliéron: Vie, enseignement, élèves, œuvres, souvenirs*. Louvain: Centre International de Dialectologie Générale.
- Prokić, Jelena. 2010. *Families and Resemblances*. Groningen: Groningen Dissertations in Linguistics, 88.
- Pușcariu, Sextil, Sever Pop, and Emil Petrovici, eds. 1938–1940. *Atlasul lingvistic român* (4 vols.). Cluj: Muzeul Limbii Române. [ALR]
- Rosenqvist, Arvid. 1919. "Limites administratives et division dialectale de la France." *Neophilologische Mitteilungen*, 20: 87–119.
- Séguy, Jean, ed. 1954–1974. *Atlas linguistique et ethnographique de la Gascogne* (6 vols.). Paris: Centre National de la Recherche Scientifique. [ALG]
- Séguy, Jean. 1971. "La relation entre la distance spatiale et la distance lexicale." *Revue de Linguistique Romane*, 35: 335–57.
- Séguy, Jean. 1973. "La dialectométrie dans l'Atlas linguistique de la Gascogne." *Revue de Linguistique Romane*, 37: 1–24.
- Shackleton, Robert. 2010. *Quantitative Assessment of English-American Speech Relationships*. Groningen: Groningen Dissertations in Linguistics, 81.
- Sneath, Peter, and Robert Sokal. 1973. *Numerical Taxonomy: The Principles and Practice of Numerical Classification*. San Francisco, CA: Freeman.
- Szmrecsanyi, Benedikt. 2013. *Grammatical Variation in British English Dialects: A Study in Corpus-Based Dialectometry*. Cambridge: Cambridge University Press.
- Viereck, Wolfgang, and Heinrich Ramisch, eds. 1991–1997. *Computer Developed Linguistic Atlas of England* (2 vols.). Tübingen: Niemeyer. [CLAE]
- Winkelmann, Otto. 2001. "Romanische Sprachatlanten." In *Lexikon der romanistischen Linguistik* (vol. 1/2), edited by Günter Holtus, Michael Metzeltin, and Christian Schmitt, 1004–68. Tübingen: Niemeyer.

8 Dialect Contact and New Dialect Formation

DAVID BRITAIN

8.1 Introduction

Dialect contact approaches to language change examine the linguistic consequences of interaction between speakers of different but mutually intelligible varieties of a language. Most of the foundational research in this field has examined contact that has resulted from acts of distant and/or large-scale migration or other significant acts of mobility, by individuals or groups, thereby bringing together speakers of often quite radically different varieties, although the approach is also used to explain contact as a result of more mundane and everyday forms of mobility. Researchers argue that certain types of temporary linguistic change typically occur when people interact, even briefly, but contend that if the contact is prolonged enough it is possible for permanent linguistic changes to result. At the level of the individual, one possible outcome is the acquisition of a second dialect. If contact occurs at a community level, with many individuals bringing different varieties together, new dialects can eventually emerge as a result of these linguistic changes, dialects that are different in systematic ways from all of those that had originally interacted.

Dialect contact approaches therefore rely on a robust link between, at one end of the scale, the linguistic outcomes of often fleeting and temporary encounters, and on the other, the emergence of new dialects that can, in some cases, come to function as national language varieties. This chapter aims to explain and exemplify that link by examining the linguistic consequences of brief, short-term contacts, of more prolonged and longer-term contacts, of linguistic accommodation and second dialect acquisition at the level of the individual, and of new dialect formation at the level of the community.

To a certain extent, all interactions between speakers of a language represent cases of dialect contact, since everyone's dialect is at least subtly different to everyone else's even within the same community, family, or peer group. Possibly as a consequence of this, researchers sometimes adopt dialect contact approaches to explain all forms of linguistic change or to challenge or reinterpret traditional explanations of specific changes. Contact approaches have been used, for example, to demonstrate the importance of examining *all* speakers in a community, rather than, as is often the case, a rather sanitized and selective sample of "authentic" and autochthonous long-term residents. They have also been used to counter claims of how certain varieties have come to diverge from others, and even to demonstrate that there has been no linguistic change at all, despite claims to the contrary from within other models of change.

8.2 The Urge to Converge

Dialect contact approaches support the view that there is a “general and seemingly universal (and therefore presumably innate) human tendency to ‘behavioural co-ordination’” (Trudgill 2004, 27–28). Consequently, when two people interact, there is an underlying, though perhaps not conscious, motivation to “make the interaction work” and, at least to some extent, there is a degree of interactional convergence. People can override this underlying orientation and actively diverge, but essentially, the model would claim, the default is for speakers to reduce the linguistic distance between themselves. If such interactional synchronisation becomes regular and routine to speakers of different dialects, it is possible that the convergent linguistic consequences may be regularly adopted as variants in that person’s repertoire, or even—over time, and in the right conditions—become permanent features of that person’s dialect. However, such an urge to convergence does not and should not imply accuracy, totality or success in that convergence on the part of the interacting speakers. As a result, this linguistic coordination does not result in harmonisation, but often in the emergence of innovative linguistic forms as well as innovative constellations of linguistic forms.

There is a very significant literature examining behavioral synchronisation, which, according to Ackerman and Bargh, “emerge[s] in virtually all situations involving more than one person” and is a “fundamental property of social interaction” (2010, 336, 357). At a linguistic level, there is evidence for convergent linguistic behavior even among the very youngest humans. In a now well-known study, Lieberman (1967, 60) showed that a 10-month-old boy lowered the fundamental frequency of his babbling when playing with his father more than when playing with his mother, as did a 13-month-old girl.

Within the field of speech and communication accommodation theory, a number of studies have demonstrated convergent linguistic behavior at different levels of linguistic structure. Figure 8.1 is based on Coupland’s (1985, 63) work examining how a Cardiff travel agent accommodated to the varying phonologies of customers of different social classes. There is, with the exception of interactions with customers in the professional occupation group, a clear relationship between the customers’ levels of non-standardness and those of the travel agent, but nevertheless the agent was rarely totally accurate in that convergence (and total accuracy is almost never socially desirable or acceptable (Auer and Hinskens 2005, 342)). Audience design models accounting for linguistic style have also clearly demonstrated how speakers attend to and linguistically shift toward their audiences, whether real or imagined, with directly addressed interlocutors triggering a more significant linguistic shift than other nearby hearers. Bell has provided positive evidence of such shift at the levels of phonology and morphology (1984), as well as discourse (Bell and Johnson 1997), as have Rickford and McNair-Knox (1994).

These accommodation studies are important for demonstrating the pervasiveness of convergence, even during fleeting encounters. More critical for the dialect contact model, however, is how these shifts, on a larger scale and over the ever longer term, can lead to enduring linguistic changes to a speaker’s repertoire. While much of the work in the speech accommodation theory paradigm is experimental, people’s everyday encounters in the world and the linguistic consequences of those events are much harder to define, control and measure. Later, the chapter will examine what can happen in the longer term—second dialect acquisition and new dialect formation, for example. But what evidence is there of the nature of the intermediate stage, the stage between fleeting accommodation and permanent linguistic change? What happens after a year of accommodation to dialects that are distinct from your own? After two years? After 10? As the time depth of such investigations becomes greater, so too do the methodological challenges, because the possibility of tracking groups of migrant individuals over ever longer periods of time becomes impractical (though see

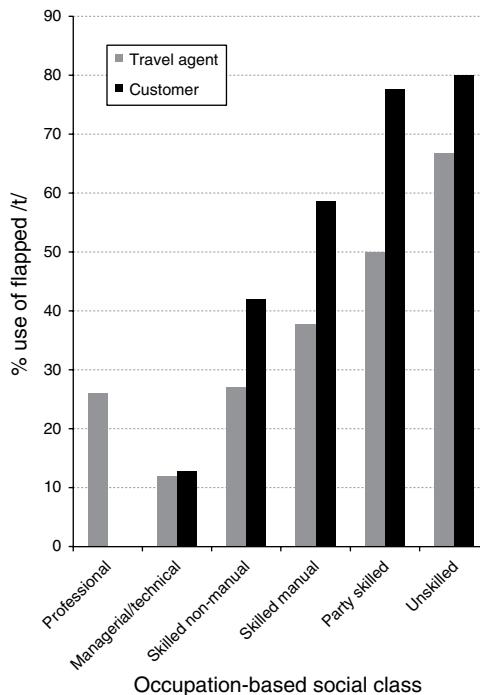


Figure 8.1 A travel agent's linguistic accommodation to her customers: the use of flapped /t/ in Cardiff (Coupland 1984, 63).

Sankoff (2004) and Rhodes (2012), who examined the real-time linguistic changes that took place in the speech of mobile individuals starring in the panel-survey TV series *7-Up*). Often, the accommodatory behavior of individuals is studied post-hoc, without researchers having been able to examine the specific variable linguistic profiles of those specific individuals *before* they migrated (see Britain 2012 for a discussion of methodologies in dialect contact research).

Two studies shed interesting light, however, on the intermediate stage between accommodation and acquisition. Mæhlum (1996) investigated the linguistic consequences of partial community membership in Longyearbyen on the Norwegian island of Spitsbergen in the Svalbard archipelago, halfway between mainland Norway and the North Pole. Svalbard is an important mining center, drawing workers predominantly, but not at all exclusively, from Northern Norway. Because of the inhospitable conditions it does not support a long-term permanently resident population, but has instead what Mæhlum calls a periodical semi-migrant community—the average length of residence has fluctuated, but in general it has been between 3 and 12 years (1996, 315). Also, with few exceptions, nobody spends the whole year on the island, but nearly all migrants return to their homes in mainland Norway for two to three months per year. For many children of the workers of Longyearbyen, a good deal of their dialect socialization period is spent in this annual rotation between the mainland and the island. In these rather socially turbulent conditions, it is not surprising that “no stabilisation toward one collective linguistic norm” (1996, 318) has occurred.

Instead, however, Mæhlum reports “a number of different norms, which are in part individually conditioned” (1996, 318), and presents some of these different norms by examining the language of four young people who have spent considerable periods of time, every year,

for most of their lives, in Longyearbyen. The youngest, aged 7, has adopted a dialect with many features from Northern Norway, the demographically dominant area of migrant origin. Another, aged 11, has developed a dialect that most resembles that of Bergen, where her mother is from. The oldest, aged 19, has developed “an individual variety which it is hardly possible to identify geographically” (1996, 318). Mæhlum interprets his dialect as a “synthesis of different elements from different regional varieties of mainland Norway” (1996, 318), but with Northern and Eastern Norwegian being especially influential (1996, 320). Despite the community-wide heterogeneity, Mæhlum suggests that there has nevertheless been partial convergence because each of the children studied has made some steps toward adopting variants from their social milieu. Her research therefore supports Trudgill’s (1986) earlier claims that individuals do not necessarily follow the same linguistic route as they acquire a new dialect.

Hirano (2013) investigated whether Anglophones of different nationalities temporarily resident in Japan accommodated to each other over a period of one year. American, English, and New Zealand language teachers on a Japanese graduate English-teaching program were recorded speaking to a compatriot shortly after arriving in Japan at the start of their teaching contract, and then again after a year. Along with the second recording, Hirano collected detailed information about the teachers’ social network ties with other Anglophones in Japan as well as with Japanese people, and developed a complex set of social network indices to measure these ties. These network scores were correlated with changes in linguistic behavior between the first and second interviews one year apart. Her data analysis showed subtle convergence after a year, convergence that strongly correlated with the strength of network ties with relevant speakers. She found, for example, that those teachers from England who had developed in Japan strong social network ties with North Americans and Australasians increased their use of word-final intervocalic /t/ flapping, whereas those who had the strongest ties with fellow English compatriots actually decreased their already low levels of flapping (Hirano 2013, 152). She also found that the teachers generally *decreased* their levels of -t/d deletion, and this shift was especially marked among those who had strong social network ties with Japanese teachers (2013, 192–194). Hirano’s study is important not just because it shows how important social network strength is in shaping levels of accommodation and acquisition, but also because she could base her claims of shift not on hypothetical estimations of how people spoke *before* the period of accommodation began, but on actual strictly comparable datasets of data recorded before and after.

8.3 Acquiring a Second Dialect

The two studies of medium-term change are unusual because they capture a snapshot of the accommodation-cum-acquisition process in rather diffuse dialect situations. Work in second dialect acquisition has tended to focus on individual migrants moving to new communities where a relatively distinct and focused ambient dialect prevails. The now classic study in this field is Chambers’ (1992) investigation of the dialects of six Canadian youngsters who moved to southern England. Through a careful examination of a number of lexical and phonological variables, Chambers was able to propose a series of linguistic principles that characterize the dialectological consequences of moving to a new dialect area. The two most important, which have received a good deal of subsequent empirical support, are:

- Lexical acquisition is faster than phonological (1992, 677). Consequently, the Canadian children picked up, for example, the “British” lexemes “trousers” (*vis-à-vis* “pants”), “nappy” (*vis-à-vis* “diaper”) and “bonnet” (*vis-à-vis* “hood” of a car) more quickly than they did non-rhoticity, lack of flapped /t/, THOUGHT-LOT split, and so on.

- “‘Simple’ phonological rules progress faster than complex ones” (1992, 682). The former are “automatic processes that admit no exceptions,” whereas the latter “have opaque outputs, that is they have exceptions or variant forms or... they have in their output a new or additional phoneme.” So Chambers contrasts, for example, the relative ease with which the younger Canadian children lose their Canadian flapped /t/ (1992, 682), subject to a clearly definable phonological rule, with their general failure to appropriately assign a back vowel to the BATH lexical set, given that Canadians “typically have the vowel [æ] in these words as in other words with ME /ã/” (1992, 683), and that the BATH lexical set is only roughly phonologically defined and subject to a number of exceptions. The difficulty speakers with mergers have in acquiring splits is further exemplified in Chambers’ data by the children’s poor success at acquiring the THOUGHT-LOT split (1992, 688; see also Vousten 1995; Labov 1994; Rys 2007).

Watts (2000) examined the variable acquisition of Cheshire English (south of Manchester) by 14 American children. Some had been in Cheshire for around two years, and others for only six months, enabling Watts to make comparisons between “early” and “late” expatriate children respectively. She compared them, as Chambers had, with similarly aged local controls. She found significant differences in the acquisitional success of the two groups. More than half of the early expatriate children, for example, had reduced their /t/-flapping to less than 50% of all tokens. Only one of the late arrivals had done so, however, and three of the others hadn’t shifted at all. Watts found lexical variants being acquired faster than phonological ones for both early and late expatriates (2000, 41), but, interestingly, found little progress in the acquisition of (Northern) Cheshire English’s use of [ʊ] in the STRUT lexical set. Instead, many of the early expatriate children used an /ʌ/ vowel intermediate between [ʌ] and [ʊ], suggesting the sort of partial convergence we saw earlier when discussing accommodation. As we will see, such intermediate forms often become fossilized in longer-term dialect contact situations too. Watts, also like Chambers, is wary of over-explicit determinations of the age at which success at second dialect acquisition begins to tail off, demonstrates considerable variability in acquisition rates amongst children of the same age in her study, and seeks social network explanations for individual differences. In the case of one child who was a slow acquirer relative to his age-mates in the survey, Watts (2000, 49) explains that he lived well away from the school he attended, had no contact with his classmates outside school, and didn’t tend to mix with kids in his neighborhood. Not every study has found social network to influence second dialect acquisition rates, however; Werlen *et al.*’s (2002) research on the acquisition of Bernese Swiss German by migrants from the Canton of Wallis did not find that social networks accounted for levels of second dialect adoption.

Studies of migration within Arabic speech communities also present supportive evidence for Chambers’ principles. Al-Dashti (1998) investigated second dialect acquisition among Egyptian migrants to Kuwait, recording male and female, long-term and short-term migrants, as well as local controls. The women in his survey had all married local men, but the men did not have local family ties. The women, consequently, had much closer ties with Kuwaiti (Arabic) than the men did, and even though the longer-term male migrants had arrived in the community at a much younger age (on average, aged 12) than the longer-term females (on average aged 23), and had spent longer in the community than the women (men on average 19 years, women 13 years), adoption success was linked more closely with social network ties than with time spent in the community and age of arrival. As other studies have suggested, he found that the longer-term migrants had acquired the new dialect more successfully than the more recent ones, who had been in Kuwait on average eight months, and that more lexis had been acquired than phonology. Many of the phonological variables Al-Dashti analyzed would be considered as “complex” in Chambers’ terms. One feature investigated was the variable (θ). Here, Egyptian Arabic has variation between [t] and [s],

but Kuwaiti Arabic has [θ] as the sole variant of (θ). It does, however, have [t] and [s] elsewhere in the sound inventory, that is, not as variants of (θ). As a result of the complexity of correspondences and non-correspondences between the two systems—for example, the fact that the Egyptians would have to add an entirely new sound to their inventory—the migrants fared relatively poorly at acquiring Kuwaiti (θ), with the long-term migrants showing evidence of having acquired it only marginally (as 5% of tokens), and the more recent migrants not at all. Other “complex” variables were acquired somewhat more successfully by the migrants, and Al-Dashti’s findings, and that of others on second dialect acquisition (e.g., Rys 2007), lends support to Kerswill’s (1996, 200) proposed cline of acquisition difficulty from vocabulary at the “easy to acquire” end to lexically unpredictable phonological rules at the “difficult to acquire” end as a way of classifying linguistic constraints on acquisition in a more fine-grained way.

So studies of longer-term migration by individuals have shown that new features can be acquired (semi-)permanently, that often complex forms are not acquired or only poorly so (rather, simpler forms are retained), that linguistically intermediate forms often result from second dialect acquisition, and that the process tends to be more successful among younger rather than older migrants, with social factors such as strength of social network ties often intervening to accelerate or hinder rates of adoption of new features.

8.4 Old Ingredients, New Dialects

In the case of second dialect acquisition research, the target variety at which migrants are aiming is generally the overwhelmingly dominant ambient variety, but nevertheless, speakers do not usually perfectly adopt it. Researchers have extended their investigations of contact to examine long-term outcomes of accommodation and acquisition in communities where there is *no* straightforwardly dominant variety and often, indeed, no indigenous dialect of that language spoken at all before migration has brought different varieties together. Such communities include those that have resulted from colonial settlement (e.g., Trudgill (2004) for (mainly Southern Hemisphere) English, Mougeon and Beniak (1994) for Canadian French, Penny (2000) for Latin American Spanish, Matsumoto and Britain (2003) for Micronesian Japanese, etc.), from mass movements of indentured labor (e.g., Barz and Siegel 1988), from planned urbanization (e.g., New Town creation (Kerswill and Williams 2000), from land reclamation (e.g., Britain 1997a, 2015), and so on. In each case, the speech community resulting from the migration is a dialectally mixed one, bringing together different social and regional dialects of British and Irish English in the case of New Zealand English (e.g., Britain 2008), different dialects of Hindi from across Eastern and Northern India in the case of Guyanese Bhojpuri (Gambhir 1988), different dialects of French in the case of Québécois (Mougeon and Beniak 1994), and so on. It is often the case, of course, that some regions are better represented than others in the mix—southern and eastern England for New Zealand, Bihar and Uttar Pradesh for Guyanese Bhojpuri. But almost always the dialect that results from such mixture ends up being different in systematic ways from any of the original migrant dialects, even in cases where there is an overwhelmingly dominant input variety.

Dialect contact theorists argue that the same sociolinguistic principles apply in these situations as occur in the shorter-term and/or individual migrant forms of contact that we saw earlier. In the case of the new communities just mentioned, however, the linguistic outcomes of accommodation and first-generation second dialect acquisition among the new migrants serve as the “linguistic target” variety for the next generation of speakers in that community (Trudgill 1986). We can track the emergence of a new dialect as follows:

- A community is formed from large-scale migration from a number of different dialect areas;
- As a result of the “urge to converge” discussed earlier, adult speakers in the community accommodate linguistically to each other, and, over time, some of this accommodatory behavior becomes (semi-)permanent. As we have seen, this can result in “complex” forms not being acquired and other forms being partially acquired. It is also argued here that very rare forms, shared by very few people in the community, are more likely to be leveled away as the forms succumb rather readily to the effects of prolonged accommodation. Some dialect forms, on the other hand, will be especially well represented in the dialect mix, and so may well be more often accommodated to than more minority dialect forms, and thereby be especially influential. As a result of all these subtle changes, the community dialect is diffuse, but nevertheless, somewhat less diffuse than it was at the point of migrant arrival in the new community. The earliest stages of linguistic focussing have begun.
- This somewhat simpler, somewhat leveled variety, incorporating some partial, incomplete shifts serves as the admittedly still diffuse ambient dialect for children being born in the community. As they grow older, children set about continuing the process of accommodation and acquisition, trying to derive a system from the linguistic mêlée around them. They focus the variety further. As we saw in the case of the Norwegian children partially resident in Svalbard, the individual linguistic routes people take as they accommodate can differ. Consequently, Trudgill (2004) argues, people at this stage—natives of the new area—demonstrate extreme inter- and intra-individual variability, and also occasionally combine input dialect forms in unusual and idiosyncratic ways. He provides the example from his work on the early New Zealand English of Mrs. Ritchie, who had Scottish parents, and who combined characteristic features of varieties of Scottish English such as rhoticity (realized as a partially unvoiced tap), the presence of /m/ in words such as “which,” and FOOT ~ GOOSE merger, with PRICE and MOUTH diphthongs typical of the south-east of England. “It is perhaps not too fanciful,” Trudgill suggests (2004, 105), “to suppose that Mrs Ritchie may be the only English speaker ever in the history of the language to have said things like *out here* [æət hiəf].”

And so the process goes on. New generations of children accommodate further, systematize more, and focus the dialect, reducing the high levels of variability. It is debatable how long it will take for a fully focused new dialect to emerge (see Kerswill and Trudgill 2005), and this will be partly constrained by factors such as how divergent the forms were that entered the mix in the first place, and opportunities for social mixing in the new community. So while in some communities new dialects can focus relatively quickly within two or three generations, others can take longer (or at least some especially complex characteristics of those varieties can take longer to fully focus; cf. Britain 1997b).

As the examples of accommodation and second dialect acquisition have suggested, a number of different linguistic processes are typical of this focussing process, usually bundled together under the label of koineization. These are as follows.

8.5 Leveling

This is the eradication of marked linguistic features, marked (a) in the sense of being in a minority in the ambient feature pool after the contact “event,” (b) in the sense of being overtly stereotyped, or (c) marked in the sense of being found rarely in the world’s languages and/or learned late in child language acquisition (Britain 2013, 176). This is by far the most

frequently attested outcome of long-term dialect contact generally. It is exemplified in detail in Kerswill and Williams's (2000) well-known study of the British New Town of Milton Keynes, in which they demonstrate how minority dialect input forms for several variables have failed to survive into the new dialect of the city. Particularly affected by the leveling in this case was the traditional local dialect spoken in the area before sudden urbanization. Kerswill and Williams note, for example, the leveling of [ɔɪ] forms of the PRICE diphthong, [ɛʊ] forms of MOUTH, relatively non-front [a] variants of TRAP, and the use of /ʌ/ instead of initial /wʊ/ in words such as "woman" (2000, 81, 86), all of which are now characteristic of the pre-New Town local dialect.

Leveling is also a characteristic of the Ban Khlong Sathon (BKS) new town in North-Eastern Thailand, studied by Prompapakorn (2005). BKS was established in the 1960s, one of a number of villages in the Khorat region settled by people moving out of areas that had achieved National Park designation, by farmers encouraged to shift from logging to arable agriculture, by political protesters fleeing conflict in Bangkok, and by many others seeking greater economic stability. It saw an influx, therefore, not only from elsewhere in Khorat, but also of speakers of Central Thai dialects (from around Bangkok, to the west) as well as from Isan dialects (from further north and east). Prompapakorn shows how one of the Isan variants [h] and an archaic relic Central Thai variant [r] of the (r) variable (e.g., /ron/ "hot," /rak/, "love") have been leveled over time, in favor of the [l] variant common to and dominant in all the varieties of Thai in the mix. Similarly, she finds that the second consonant in the clusters /Cl/ and /Cr/, which are minority forms in Central Thai and Khorat and totally absent from Isan, have been completely eradicated among her younger speakers. In each case there were clear majority forms in the input dialects, and the leveling process has "tidied up," getting rid of remnant minority variants.

Finally, Matsumoto and Britain (2003, 57), examining the variety of Japanese that developed in Palau in Micronesia during Japan's 30-year control of the islands between 1914 and 1945, found that the negating suffix "-nai"—the dominant variant from the dominant source region of migrants from Japan, the (north)east—had leveled away other variants from that region, as well as variants from other parts of Japan, such as "-n" from the (south)west.

8.6 Simplification

Simplification refers to the process by which a new dialect becomes more regular, having fewer categories (such as gender, case, honorifics), fewer person/number inflections, or fewer complex constraints on variation than the dialects in the original mix (Britain 2013, 177). One set of new dialect contexts in which simplification is very visible is the numerous dialects of Hindi-Bhojpuri that have emerged as a result of indentured labor migrations from India to, for example, Fiji, Mauritius, South Africa, Trinidad, and Guyana in the second half of the nineteenth century (e.g., Barz and Siegel 1988). Gambhir (1981, 259) compares the present tense system in Indian varieties of Hindi (with over two dozen morphemes, demonstrating sensitivity to person, number, and gender) with that of the Guyanese variety, which has just two. Similarly, Siegel (1988, 131) shows that Fiji Hindi retains just 2 definite future suffixes, whereas Indian Hindi-Bhojpuri, again sensitive to person, number and gender, has 15. Fiji Hindi uses *egā*, which originates in Bazaar Hindustani, for first and second person plural, male and female, whereas the generic third person form in Fiji is *t*, Indian Hindi-Bhojpuri's third person singular form (and its only form not marked for gender). Bhatia (1988, 190), similarly, compares the system of future tense marking in Trinidad Hindi, which has 3 distinct forms, with that in Indian Hindi-Bhojpuri, which has 18. While the Trinidadian variety only marks for person, the Indian varieties mark also for honorifics, number and gender.

8.7 Interdialect

We noted earlier in the discussions of accommodation and second dialect acquisition that often the linguistic consequences of these processes are not necessarily complete or fully accurate. Interdialect forms, in the context of koineization and new dialect formation, result from a fossilization of this partial, incomplete accommodation. The outcome, therefore, is that variants emerge that were present in none of the input varieties, but clearly result from the convergence of such varieties because the novel form is somehow linguistically intermediate. One clear example of such interdialect can be found in the accent of the English Fens, an area of former marshland that was reclaimed gradually between the mid-seventeenth and late nineteenth centuries, triggering migration from outside to farm the rather fertile new lands (Britain 1997b). As a result of contact between dialects to the north and west of the area which had [ʊ] variants of STRUT (and consequently had no STRUT ~ FOOT split) and those to the south and east, which had [ʌ] variants, a phonetically intermediate [ɤ] variant developed and stabilized.

8.8 Reallocation

Reallocation is said to have occurred when two (or more) variants rather than just one survive the focussing process. This, like interdialect, is relatively rare, but it tends to occur when the two (or more) eventually victorious variants come into contact in relatively equal numbers. Focussing, however, refunctionalizes the variants to serve new and different roles—linguistic or social—from those they performed in the original input variety. In Britain (1997a), I showed how contact-induced reallocation could account for the emergence of a Canadian-Raising-like allophonic distribution for /ai/ in the Fens, mentioned above. Variants of /ai/ with open nuclei [aɪ~a:] for example, [naɪ?ta:m] “night time” from the West of the Fens and variants with centralized nuclei [əɪ], [nəɪ?taɪm] from the East interacted to produce an allophonic system sensitive to following voicing in the Central Fens, with [əɪ] before voiceless consonants and [aɪ~a:] elsewhere, giving examples such as [nəɪ?ta:m]. Taeldeman (1989), examining a border contact area between East and West Flanders in Belgium, found another such example. He noted that dialects in the western parts of Eastern Flanders delete intervocalic /g/ (so *liegen*, “to lie” [li:ən], *brugge*, “bridge” [broe:ə]), but the dialects of Western Flemish retain it as a laryngeal—[li:hən], [brøehə]. In villages at the border between the two regions, Taeldeman reports that the Western laryngeal forms are used before <-e> [brøehə] and the Eastern preference for deletion is found before <-en> [li:ən] (1989, 156). Two regional forms have therefore been reallocated to different phonological positions.

Siegel (1997) and Prompapakorn (2005) provide examples where the reallocated forms come to perform social rather than linguistic work. In Fiji Hindi, the originally Indian Bhojpuri third person possessive form *okar* has rural, rustic connotations (relative to the more socially neutral *uske*), and is, according to Siegel (1997, 127), often used in comedic portrayals of “country bumpkin characters.” In Ban Khlong Sathon, Prompapakorn reports how contact between the Central Thai negator *mai* and the (north eastern) Isan form *bor* has led to stylistic reallocation. Among her first-generation speakers, those from the Central Thai area categorically used *mai* and those from Isan equally categorically used *bor*. Amongst the third generation there was a mix, with *bor* being used 70% of the time. *Mai*, however, is in good health, and Prompapakorn demonstrates that it is used when discussing more “formal” topics, such as visits to the doctor, school work, and so on, whereas *bor* is used for more “informal” themes (2007, Ch. 4) (see further Britain and Trudgill 2005).

These processes, and especially leveling, have been shown to occur in many different contact situations from many different locations and many different languages. It is

important to remember, however, that these are just strong tendencies, and sometimes new variants can succeed which do not fit into these types (Trudgill 1986, 102).

The dialect contact model was originally applied mostly to account for situations in which quite radically divergent dialects have come into contact. However, widespread evidence of convergence, koineization and especially leveling beyond such specific contexts has led some to extend the model to attempt to account for patterns of dialect variation that appear to have resulted from more mundane and everyday forms of mobility. Consequently, researchers have used contact models to provide explanations for urban linguistic change (e.g., Milroy *et al.* 1994, Watt 2002), to explain the gradual disappearance of traditional rural variants (e.g., Britain 2009, Piercy 2010), and to explain convergence, at the expense of local dialects, at the regional level (e.g., Al-Wer 1997, Torgersen and Kerswill 2004, Vandekerckhove 2005, Hornsby 2009). This process has come to be known as supralocalization or regional dialect leveling (see Britain 2010, 2011).

8.9 Challenging Change

As well as extending the model to account for more everyday forms of mobility-induced dialect mixing, contact approaches have also been used to question and reinterpret traditional accounts of certain linguistic changes, as well as to challenge approaches to sociolinguistic methodology that shunned mobility, treating it as a threat to our ability to access descriptions of “authentic” dialects.

One example of how contact models have been used to reinterpret existing explanations of change (see also Britain 2008) is Kuo’s (2005) work on Taiwanese Mandarin. Taiwan acquired a significant Mandarin-speaking community only after the Chinese Civil War in the late 1940s, when Mao’s Communists pushed the Chinese Nationalists out of Mainland China. The Nationalists found a population on Taiwan that spoke either an indigenous Austronesian language or another Chinese language but very few speakers of Mandarin, which was dominant on the Mainland. The numerically most important language, spoken by around three-quarters of the population of Taiwan, was Southern Min. The new Nationalist rulers from the Mainland, however, imposed Mandarin, their own variety, as the language of education and administration. Gradually, a Taiwanese variety of Mandarin emerged. Previous research on Taiwanese Mandarin (e.g., Chien 1971, Kubler 1985) had claimed that it was different from Standard Beijing Mandarin because of the second language acquisition failure of the Southern Min-speaking population—essentially, that Taiwanese Mandarin was a result of the failure of the Taiwanese people to learn Standard Mandarin accurately. They pointed to the fact that whilst Standard Beijing Mandarin had four retroflex consonants, /tʂ tʂʰ ʂ ʐ/, in its inventory, Southern Min had none. Thus, when learning Mandarin, the Southern Min speakers merged the retroflexes with their corresponding non-retroflex sounds /ts tsʰ s dz/, and diffused the merger to the population at large, including to the children of original mainlanders (see Kuo 2005 for a review of such claims in the literature).

Kuo conducted a careful analysis both of mid-twentieth century dialectological descriptions of Chinese regional dialects, and census information on the regional origins of the mainland migrant population. She thereby demonstrated that while retroflexes were common in central Beijing they were rarely found elsewhere in China, and were almost entirely absent amidst the varieties of different Chinese languages spoken by the mid-twentieth century migrants to Taiwan (Kuo 2005, Ch. 6). Overall, she estimates that non-retroflex forms were at least ten times more common than retroflexes in the migrant population (2005, 142). The merger of retroflex and non-retroflex proposed by earlier researchers seems untenable, she argues, given that retroflex consonants were barely used at all by the Mandarin-speaking population of Taiwan, let alone the Southern Min speakers (and she cites claims by Zhang

(1974) that there was no guarantee that the teachers in Taiwanese schools could speak Standard Beijing Mandarin in any case). In Taiwan, there were simply no—or rather far too few—retroflexes to merge with. So it seems very likely that the lack of retroflexes (and many other features of Standard Beijing Mandarin, examined by Kuo (2005)) in Taiwanese Mandarin was a simple result of them not having been brought to Taiwan in sufficient numbers in the first place, and the few that were brought, being highly marked, were swiftly leveled away. Kuo (2005) additionally conducted an analysis of these features in the contemporary Taiwanese Mandarin of the city of Keelung, and finds that of over 12,000 tokens of the relevant variables there were no fully retroflex tokens at all, and just 21 occurrences of “near-retroflexes” (2005, 136).

It would be quite appropriate to argue at this point that, since Southern Min was the overwhelmingly dominant variety spoken on Taiwan when the Mainlanders arrived, and since the variety does not have the retroflexes, it is likely that language interference played a significant role in accounting for the lack of them in Taiwanese Mandarin, even if the formal explanation for their lack is not one of merger. But Kuo found other linguistic variants in present-day Taiwanese Mandarin that are *not* found in local Southern Min, but which were the dominant variants among the migrant Mainlander population. These include the presence of /f/, /y/, /ie/, the “dental apical” vowel /i/, and the preservation of a distinction between /n/ and /l/. Furthermore, [an] variants of /aj/ and [on] variants of /uj/—usual today in Keelung—were dominant in mainland dialects of Chinese languages but not present in Southern Min, and whereas the former was the main form used in Beijing, the latter was not. In each of these cases the majority forms survived and the minority ones were leveled away, as the contact model would predict. Kuo’s evidence (see 2005, 184 for a summary) suggests that the overwhelmingly dominant determiner of present-day Taiwanese Mandarin is not the influence of Southern Min, nor the influence of the Beijing standard, but what the majority dialect form was amongst the migrants from the mainland. Taiwanese Mandarin is, Kuo argues, largely a Chinese *koiné* shaped by processes of dialect leveling.

Dialect contact approaches have also been influential methodologically. Whereas traditional dialectological and variationist approaches have tended to exclude newcomers in favor of speakers born and socialized within particular communities, contact approaches recognize the influential role of migrants in triggering changes that ultimately may affect the “indigenous” locals. In earlier variationist work it was usual to exclude residents who had not fully been brought up in the local community, and methodological introductions to variationism weighed up what an acceptable cut-off age for “nativeness,” beyond which people will never successfully acquire the local dialect, might be (e.g., Milroy and Gordon 2003, 27). Some argued that children can completely acquire the local dialect if they arrive in the community by the age of 8 to 10, whereas others pointed to evidence that only children whose parents had been raised in the community gained full competence (e.g., Payne 1980). The focus was not so much on the potential influence of migrants, but about the acquisition process and the likelihood that people in the sample represented “authentic,” situated and localized dialects. Labov’s pioneering study of New York (2006, 110–111) excluded those who had arrived after the age of eight. Calculations by Kerswill (1993, 35) suggest that excluding mobile individuals and late arrivals from his study in Bergen, Norway, meant that “well over 50 percent of the original sample are excluded by various nativeness-related criteria.”

Horvath (1985) was one of the first to demonstrate the important role of migrants in her work in Sydney, Australia, where migrant Greek and Italian communities were leading changes that diffused to the majority Anglo-Irish community. Other studies, couched more explicitly within a dialect contact framework, have made similar findings. Recent work on change in the Englishes spoken in London, for example, has provided ample evidence of migrant Englishes diffusing changes to the local “Anglo” population (see, for example,

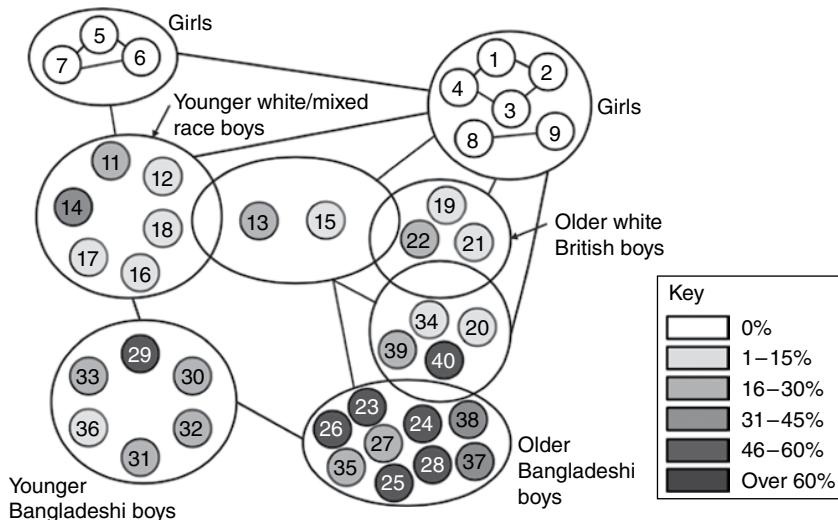


Figure 8.2 [ei] variants of /ei/ among different friendship subclusters within an East End London youth club (adapted from Fox 2007). Each speaker is represented by a circle; circled groups of speakers represent small friendship groups; the lines between the friendship groups show intergroup connections.

Cheshire *et al.* 2011). Particularly enlightening in this respect is Fox's work (2007, see also 2015) on the transmission of phonological and morphological features from the large Bangladeshi community to Anglos in the east of the UK capital. She conducted long-term research in an East End youth club, frequented by Anglo boys and girls and Bangladeshi boys and, as a result of her ethnography, was able to construct sociograms showing the clusterings of closer network ties among the adolescents in the club. Figure 8.2 demonstrates how one phonological variant—the (for London) innovative use of [ei] for FACE (cf. traditional London Cockney [æɪ ~ aɪ] (Wells 1982, 307)) has percolated from the older Bangladeshi boys in the club down to Anglo boys and younger Bangladeshi boys, but has not yet reached the girls. Fox's work demonstrates that sampling the whole community, including migrant speakers, is essential if we wish to understand the origins and direction of local linguistic change.

8.10 Conclusion

Dialect contact models of language change emerged largely through attempts to account for the linguistic consequences of rather specific acts of long-distance migration, for example, European settlement colonisation in the Americas and Australasia (e.g., Trudgill 1986, 2004), Indian indentured labor in colonial plantations (e.g., Barz and Siegel 1988), and the individual migration of expatriates and their children between different speech communities (e.g., Chambers 1992). Today these models are routinely used to address questions of contemporary language change more generally. We can account for this more holistic engagement in a number of ways. Firstly, there has been a recognition of the pervasiveness of contact. All interaction between speakers is, in some sense, dialect contact: the coming together of different (albeit possibly only subtly so) language varieties. This, together with humans' seemingly innate behavioral convergence, has allowed dialect contact models to explain

change in contexts of contact much less dramatic than those it had engaged with earlier. Secondly, contact approaches have taken a much more inclusive stance than some other approaches to dialect change with respect to who is a “relevant” speaker: an individual who, in adding to the pool of variants circulating in a community, may be contributing toward the trajectory of the local dialect. Thirdly, there has, across the social sciences, been an increasing recognition of the need to challenge the a-mobile focus of much earlier research and to present a new sociology that engages with (but does not fetishize) mobility (see, for example, Cresswell 2006, Urry 2000). Dialect contact models are built on an assumption of mobility, and have consequently been at the forefront of contemporary attempts to understand traditional dialect leveling, dialect supralocalization and other linguistic consequences of a world that appears to be increasingly on the move (Britain 2013). Such issues had, for a long time, been put to one side in a variationist sociolinguistics that was engaged primarily with community-internal linguistic change (see Labov 2001, 20, and Britain 2016, for a critique of this focus).

Challenges, of course, remain. There is an ongoing debate about the extent to which the outcomes of contact are deterministic and about the role played by social identity in accounting for the shape of new dialects (see, for example, Trudgill 2008, and the suite of papers in response to it) as well as about the extent to which the media—as opposed to dialect contact—can account for the propagation of some changes (see Sayers 2014, and ensuing set of responses). Dialect contact approaches are likely, nevertheless, to continue to find fertile ground in a dialectology that more and more is embracing the reality of everyday mobility.

REFERENCES

- Ackerman, Joshua, and John Bargh. 2010. “Two to tango: Automatic social coordination and the role of felt effort.” In *Effortless Attention: A New Perspective in the Cognitive Science of Attention and Action*, edited by Brian Bruya, 335–372. Cambridge, MA: MIT Press.
- Al-Dashti, Abdelmohsen. 1998. *Multilingualism in Kuwait: A Sociolinguistic Analysis*. Unpublished PhD dissertation, University of Essex.
- Al-Wer, Enam. 1997. “Arabic between reality and ideology.” *International Journal of Applied Linguistics*. 7: 251–265.
- Auer, Peter, and Frans Hinskens. 2005. “The role of interpersonal accommodation in a theory of language change.” In *Dialect Change: Convergence and Divergence in European Languages*, edited by Peter Auer, Frans Hinskens, and Paul Kerswill, 335–357. Cambridge: Cambridge University Press.
- Barz, Richard, and Jeff Siegel, eds. 1988. *Language Transplanted: The Development of Overseas Hindi*. Wiesbaden: Harrassowitz.
- Bell, Allan. 1984. “Style as audience design.” *Language in Society* 13: 145–204.
- Bell, Allan, and Johnson, Gary. 1997. “Towards a sociolinguistics of style.” *University of Pennsylvania Working Papers in Linguistics* 4: 1–22.
- Bhatia, Tej. 1988. “Trinidad Hindi: Its genesis and generational profile.” In *Language Transplanted: The Development of Overseas Hindi*, edited by Richard Barz, and Jeff Siegel, 179–196. Wiesbaden: Harrassowitz.
- Britain, David. 1997a. “Dialect contact and phonological reallocation: ‘Canadian Raising’ in the English Fens.” *Language in Society* 26: 15–46.
- Britain, David. 1997b. “Dialect contact, focusing and phonological rule complexity: The koineisation of Fenland English.” *University of Pennsylvania Working Papers in Linguistics*. 4: 141–170.
- Britain, David. 2008. “When is a change not a change? A case study on the dialect origins of New Zealand English.” *Language Variation and Change* 20: 187–223.
- Britain, David. 2009. “One foot in the grave? Dialect death, dialect contact and dialect birth in England.” *International Journal of the Sociology of Language* 196/197: 121–155.
- Britain, David. 2010. “Supralocal regional dialect levelling.” In *Language and Identities*, edited by

- Carmen Llamas and Dominic Watt. 193–204. Edinburgh: Edinburgh University Press.
- Britain, David. 2011. “The heterogenous homogenisation of dialects in England.” *Taal en Tongval*, 63: 43–60.
- Britain, David. 2012. “Koineization and cake baking: Reflections on methods in dialect contact research.” In *Methods in Contemporary Linguistics*, edited by Andrea Ender, Adrian Leemann, and Bernhard Wälchli, 219–238. Berlin: de Gruyter.
- Britain, David. 2013. “The role of mundane mobility and contact in dialect death and dialect birth.” In *English as a Contact Language*, edited by Daniel Schreier and Marianne Hundt, 165–181. Cambridge: Cambridge University Press.
- Britain, David. 2015. “Between North and South: The Fenland.” In *Researching Northern Englishes*, edited by Raymond Hickey. 417–435. Amsterdam: Benjamins.
- Britain, David. 2016. “Sedentarism, nomadism and the sociolinguistics of dialect.” In *Sociolinguistics: Theoretical Debates*, edited by Nikolas Coupland. 217–241. Cambridge: Cambridge University Press.
- Britain, David, and Peter Trudgill. 2005. “New dialect formation and contact-induced reallocation: Three case studies from the English Fens.” *International Journal of English Studies*, 5: 183–209.
- Chambers, Jack. 1992. “Dialect acquisition.” *Language*, 68: 673–705.
- Cheshire, Jenny, Paul Kerswill, Sue Fox, and Eivind Torgersen. 2011. “Contact, the feature pool and the speech community: The emergence of multicultural London English.” *Journal of Sociolinguistics*, 15: 151–196.
- Chien, Ching-gwo. 1971. *A Contrastive Study of the Phonological Systems of Mandarin Chinese and Taiwanese*. Unpublished MA dissertation, Fu Jen University.
- Coupland, Nikolas. 1984. “Accommodation at work: Some phonological data and their implications.” *International Journal of the Sociology of Language*, 46: 49–70.
- Cresswell, Timothy. 2006. *On the Move: Mobility in the Modern Western World*. London: Routledge.
- Fox, Sue. 2007. The Demise of Cockneys? Language Change in London’s ‘Traditional’ East End (PhD). University of Essex.
- Fox, Sue. 2015. *The New Cockney: New Ethnicities and Adolescents’ Speech in the Traditional East End of London*. Basingstoke: Palgrave Macmillan.
- Gambhir, Surendra. 1981. *The East Indian Speech Community in Guyana: A Sociolinguistic Study with Special Reference to Koine Formation*. Unpublished PhD dissertation, University of Pennsylvania.
- Hirano, Keiko. 2013. *Dialect Contact and Social Networks: Language Change in an Anglophone Community in Japan*. Frankfurt: Peter Lang.
- Hornsby, David. 2009. “Dedialectalization in France: Convergence and divergence.” *International Journal of the Sociology of Language*, 196/197: 157–180.
- Horvath, Barbara. 1985. *Variation in Australian English: The Sociolects of Sydney*. Cambridge: Cambridge University Press.
- Kerswill, Paul. 1993. “Rural dialect speakers in an urban speech community: The role of dialect contact in defining a sociolinguistic concept.” *International Journal of Applied Linguistics*, 3: 33–56.
- Kerswill, Paul. 1996. “Children, adolescents and language change.” *Language Variation and Change*, 8: 177–202.
- Kerswill, Paul, and Peter Trudgill. 2005. “The birth of new dialects.” In *Dialect Change: Convergence and Divergence in European Languages*, edited by Peter Auer, Frans Hinskens, and Paul Kerswill, 196–220. Cambridge: Cambridge University Press.
- Kerswill, Paul, and Ann Williams. 2000. “Creating a new town koine.” *Language in Society*, 29: 65–115.
- Kubler, Cornelius. 1985. *The Development of Mandarin in Taiwan: A Case Study of Language Contact*. Taipei: Student Book.
- Kuo, Yun-Hsuan. 2005. *New Dialect Formation: The Case of Taiwanese Mandarin*. Unpublished PhD dissertation, University of Essex.
- Labov, William. 1994. *Principles of Linguistic Change, Vol. 1: Internal Factors*. Oxford: Blackwell.
- Labov, William. 2001. *Principles of Linguistic Change, Vol. 2: Social Factors*. Oxford: Blackwell.
- Labov, William. 2006. *The Social Stratification of English in New York City*, 2nd ed. Cambridge: Cambridge University Press.
- Lieberman, Philip. 1967. *Intonation, Perception and Language*. Cambridge, MA: MIT Press.
- Mæhlum, Brit. 1996. “Semi-migration in the Arctic – A theoretical perspective on the dialect strategies of children on Spitsbergen.” In *Language Contact across the North Atlantic*, edited by Per Sture Ureland, and Iain Clarkson, 313–331. Tübingen: Niemeyer.

- Matsumoto, Kazuko, and David Britain. 2003. "Contact and obsolescence in a diaspora variety of Japanese: The case of Palau in Micronesia." *Essex Research Reports in Linguistics*, 44: 38–75.
- Milroy, Lesley, and Matthew Gordon. 2003. *Sociolinguistics: Method and Interpretation*. Oxford: Blackwell.
- Milroy, James, Lesley Milroy, Sue Hartley, and David Walshaw. 1994. "Glottal stops and Tyneside glottalisation: Competing patterns of variation and change in British English." *Language Variation and Change*, 6: 327–358.
- Mougeon, Raymond, and Edouard Beniak, eds. 2004. *Les Origines du Français Québécois*. Sainte-Foy, QC: Les Presses de l'Université Laval.
- Payne, Arvilla. 1980. "Factors controlling the acquisition of the Philadelphia dialect by out-of-state children." In *Locating Language in Time and Space*, edited by William Labov, 143–178. New York: Academic Press.
- Penny, Ralph. 2000. *Variation and Change in Spanish*. Cambridge: Cambridge University Press.
- Piercy, Caroline. 2010. *One/a/or Two? The Phonetics, Phonology and Sociolinguistics of Change in the trap and bath Vowels in the Southwest of England*. Unpublished PhD dissertation, University of Essex.
- Prompapakorn, Praparat. 2005. *Dialect Contact and New Dialect Formation in a Thai New Town*. Unpublished PhD dissertation, University of Essex.
- Rhodes, Richard. 2012. *Assessing the Strength of Non-contemporaneous Forensic Speech Evidence*. Unpublished PhD dissertation, University of York.
- Rickford, John, and Faye McNair-Knox. 1994. "Addressee- and topic-influenced style shift: A quantitative sociolinguistic study." In *Sociolinguistic Perspectives on Register*, edited by Douglas Biber, and Edward Finegan, 235–276. Oxford: Oxford University Press.
- Rys, Kathy. 2007. *Dialect as Second Language: Linguistic and Non-linguistic Factors in Secondary Dialect Acquisition by Children and Adolescents*. Unpublished PhD dissertation, University of Gent.
- Sankoff, Gillian. 2004. "Adolescents, young adults and the critical period: Two case studies from Seven Up." In *Sociolinguistic Variation: Critical Reflections*, edited by Carmen Fought, 121–139. Oxford: Oxford University Press.
- Sayers, David. 2014. "The mediated innovation model: A framework for researching media influence in language change." *Journal of Sociolinguistics*, 18: 185–212.
- Siegel, Jeff. 1988. "The development of Fiji Hindustani." In *Language Transplanted: Development of Overseas Hindi*, edited by Richard Barz and Jeff Siegel, 121–149. Wiesbaden: Harrassowitz.
- Siegel, Jeff. 1997. "Mixing, levelling and pidgin/creole development." In *The Structure and Status of Pidgins and Creoles*, edited by Arthur Spears, and Donald Winford, 111–150. Amsterdam: Benjamins.
- Taeldeman, Johan. 1989. "A typology of dialect transitions in Flanders." In *New Methods in Dialectology: Proceedings of a Workshop held at the Free University Of Amsterdam, December 7–10, 1987*, edited by Bert Schouten, and Pieter van Reenen, 155–163. Dordrecht: Foris.
- Torgersen, Eivind, and Paul Kerswill. 2004. "Internal and external motivation in phonetic change: Dialect levelling outcomes for an English vowel shift." *Journal of Sociolinguistics*, 8: 24–53.
- Trudgill, Peter. 1986. *Dialects in Contact*. Oxford: Blackwell.
- Trudgill, Peter. 2004. *New Dialect Formation: The Inevitability of Colonial Englishes*. Edinburgh: Edinburgh University Press.
- Trudgill, Peter. 2008. "Colonial dialect contact in the history of European languages: On the irrelevance of identity to new-dialect formation." *Language in Society*, 37: 241–254.
- Urry, John. 2000. *Sociology beyond Societies: Mobilities for the Twenty-First Century*. London: Routledge.
- Vandekerckhove, Reinhild. 2005. "Interdialectal convergence between West-Flemish urban dialects." In *Perspectives on Variation: Sociolinguistic, Historical, Comparative*, edited by Nicole Delbecque, John van der Auwera, and Dirk Geeraerts, 111–127. Berlin: de Gruyter.
- Vousten, Rob. 1995. *Dialect als Tweede Taal: Linguistische en Extra-linguistische Aspecten van de Verwerving van een Noordlimburgs Dialect door Standaardtalige Jongeren*. Amsterdam: Thesis Publishers.
- Watt, Dominic. 2002. "'I don't speak with a Geordie accent, I speak, like, the Northern accent': Contact-induced levelling in the Tyneside vowel system." *Journal of Sociolinguistics*, 6: 44–63.

- Watts, Emma. 2000. *Acquisition of the Cheshire Dialect by American Expatriate Children*. Unpublished MA dissertation, University of Essex.
- Wells, John. 1982. *Accents of English* (3 vols.). Cambridge: Cambridge University Press.
- Werlen, Iwar, Barbara Buri, Marc Matter, and Johanna Ziberi. 2002. *Projekt Üsserschwyz: Dialektloyalität und Dialekt-akkommodation von Oberwalliser Migranten*. Bern: Institut für Sprachwissenschaft, Universität Bern.
- Zhang, Boyu. 1974. *Taiwan Diqu Guoyu Yundong Shiliao*. Taipei: Taiwan Commercial Press.

9 Dialect Change in Europe—Leveling and Convergence

PETER AUER

9.1 Introduction

There can be no doubt that the main development in the European dialects over the past hundred years or so has been convergence, both among the dialects themselves and toward the standard variety.¹ Usually, these two developments go hand in hand, leading to leveling, although horizontal convergence between dialects is occasionally also observed without convergence toward the standard variety. In some parts of Europe, the influence of the standard variety on the dialects has been much more pronounced than in others. For instance, the dialects in most of Denmark and France have almost disappeared, while in German-speaking Switzerland the dialects have, by and large, remained uninfluenced by the standard variety (although they may have leveled out horizontally toward more prestigious dialectal varieties). Some varieties lack a standard roof, and hence have no chance to converge to it.

This overall tendency contrasts sharply with the situation in the Americas and Australasia, where leveled varieties of the European colonial languages already existed during the establishment of colonial settlements; some of these leveled varieties then developed into new standard varieties ("New Zealand English," "Argentinian Spanish," "Brazilian Portuguese," etc.), whereas some remained non-standard but underwent divergence from each other, such that new regional ways of speaking emerged (as in the USA). Some traditional settlers' dialects have survived in isolation in so-called *language enclaves* (mainly dialects of (Low) German, Dutch, Swedish, Danish, and Norwegian) surrounded by a different colonial (standard) language, and have also diverged from their European ancestor dialects.

The overall picture of convergence in Europe versus divergence in the former colonies, although mostly a fitting one, obscures some smaller counteracting trends in Europe that point to divergence: owing to certain political and social processes, recent developments in some European regions have boosted regional ways of speaking ("New Regionalism"). Additionally, massive immigration from southern Europe, the Middle East and increasingly also other parts of the world has led to the establishment of (poly-)ethnic ways of speaking that have introduced new non-standard variants into (mainly) young people's linguistic repertoires, a process which is, however, orthogonal to the development in the dialects.

9.2 Dialect and Standard

From the nineteenth century onward, terms such as leveling, loss, mixture, split, convergence, and divergence have been used by linguists and dialectologists. Since their meanings are not always well defined, they will be discussed here briefly.

Firstly, it is of course important to clarify what we mean by *dialect*. In continental European dialectology and sociolinguistics, the term is always used in opposition to a standard language/variety. Hence, the standard is not considered a dialect. This view contrasts with the Anglo-American tradition, in which “dialect” is often equated with “variety,” and where the term “standard dialect” is not uncommon. This oppositional definition, which goes back at least to Coseriu (1980), implies that there is no dialect without an overarching standard. Germanic varieties such as Alsatian, which has lost its standard German roof, are therefore no longer considered dialects. In some European sociolinguistic research traditions, particularly in Italy, the term “dialect” is avoided altogether in order to refer to non-standardized varieties such as Sicilian, Calabrian, and Venetian that are much older than the standard and should not be seen as being derived from it, as the term “dialect (of Italian)” might suggest. However, the historical status of Bavarian in Germany, Brabantian in Belgium, and Northumbrian in England is no different from that of Sicilian or Calabrian: they all pre-date the respective standard languages (in some form or other).² The relational definition of dialect is, by definition, linked to a standard variety. The terminology suggests—correctly—that once the particular relationship between standard and dialect is established, it will result in the two varieties impacting on each other, usually with the standard influencing the dialect more than the reverse.

The lay usage of the term “dialect” in Europe oscillates between a reference to a reified construct which on the one hand has gone through a multitude of processes of enregisterment and sometimes even codification, and a reference to any regional way of speaking (which dialectologists would call a regional dialect, or even a regionalized standard), on the other. In the first case, the “dialect” is usually believed to have disappeared from everyday life but to have existed in its “authentic,” “pure” form in the (distant) past. This nostalgic (re-)construction of the dialect as a “thing lost” corresponds with the fact that only few speakers consider themselves—and are considered by others to be—“real” dialect speakers. These same speakers exhibit a clear penchant for restricting such authenticity claims to very old people with a rural background. The reification of dialects as things lost has been supported by professional research in traditional dialectology, which began at the same time as standard varieties became codified as overarching roofs over the dialects, often as early as the seventeenth century. This codification appeared first in the format of *idiotica* (vocabulary collections of local linguistic peculiarities), then in the format of dialect dictionaries, atlases and grammars, even though linguists found out very early that giving precise definitions of such dialects is difficult. These enregistered varieties can be called “traditional dialects.”

Many people in Europe who consider themselves dialect speakers will readily concede that they cannot any longer speak the “real” (traditional) dialect, but may still use the term “dialect” to refer to their own way of speaking, as long as it displays regional features that are salient to them. The linguistic study of regional speech forms in this second sense of “dialect” has received much less attention in European dialectology and sociolinguistics than research on the traditional dialects, arguably because it lacks the prestige of the latter. Nevertheless, research on regionalized modes of speech has increased over the last decades at least for certain parts of Europe, so that some generalizations are possible. It is in research on these “modern” dialects that the terms “convergence” and (to a lesser degree, since less relevant) “divergence” have been employed most often.

Although the geographical boundaries between dialects are often difficult to draw, they are highly salient constructs for the speakers themselves. The base category level on which the prototypes are identified by lay speakers is often not that of the local dialect (of a village,

or another narrowly defined area), but rather that of a larger politically or ethnically defined spatial entity. It is therefore often based on language-external considerations. For instance, lay terms for German dialects such as "Swabian," "Franconian," or "Bavarian" refer to the Germanic tribes who populated these areas in the time of the "Migration Period" (*Völkerwanderung*) mostly before the year 1000, and lay terms such as Dutch/Belgian "Limburgian" or French "Provençal" refer to (former) political units. Needless to say, the enregisterment of these dialect areas is heavily based on media representations that typically select highly salient features to portray speakers from a certain region. In Germany, these prototypes are linked to the dialects of the urban centers (e.g., Stuttgart, Nurnberg, and Munich for Swabian, Franconian, and Bavarian, respectively). Below this base level of categorization, speakers may be able to name and describe more narrowly defined dialect areas if they know them from everyday experience (cf. Stöckle 2014).

The definition of "standard" is more controversial. I have suggested elsewhere (Auer 2005) that a standard variety in modern Europe has three features, not all of which emerged at the same time historically: a standard language is (a) an H-variety (i.e., minimally used for writing and on formal occasions, and therefore carrying some official prestige), (b) a common language (i.e., overarching a language area often defined by national political borders in which speakers, regardless of which other regional or local linguistic resources they use, orient to this common variety for communication) and (c) a codified variety for which some kind of externalized norm exists. In some European societies, the standard variety developed from one particular highly prestigious regional variety (often that of the capital, such as Paris, London, or Stockholm) which is considered to be closest, or even identical, to the standard. In others it was actively constructed by linguistically interested laymen (sic) and (later) linguists in such a way as to avoid salient regional features and to be acceptable to speakers from many regions, as was the case for German, Finnish, or Czech. Particularly in the latter type, the standard starts out as an exclusively written variety, often constructed against the background of language history, and is used for oral communication only after a long process of standardization. For instance, it took the German standard language, although it was firmly established as a written norm in the eighteenth century, until the early twentieth century to become the relatively uniform spoken variety we know today, when general education and the development of mass media led to at least a passive knowledge of the oral standard among the masses of the population, and when its acceptance as a spoken variety became undisputed by the elites. Before that, regional varieties were spoken in formal contexts, even by educated speakers, and dialects in informal ones (cf. Mihm 2000). Nonetheless, even in the first half of the twentieth century, the oral standard variety was beyond the active reach of most Germans (and Austrians and German-speaking Swiss people). The most important process that changed this situation in the second half of the twentieth century is known as the demotivization of the oral standard (cf. Auer and Spiekermann 2011, following Mattheier 1997): the standard gained currency as a spoken variety among the masses, and it was used in more and more situations in everyday life to the degree that in present-day Germany it is the only spoken variety available to many speakers, particularly in the north and parts of the middle of the country. The fact that the standard variety is now available to everyone has led to changes in what is accepted as standard, as well as an increased tolerance for variation within the standard (see below for examples).

9.3 Convergence, Divergence, Diffusion, and Leveling

We are now in a position to define the processes by which two varieties become more similar (convergence) or less similar (divergence). Obviously, these processes can take place between dialects, or between dialect and standard. The terms "horizontal" and "vertical" are often

used to capture this difference. In the first case, the converging/diverging varieties are of approximately the same status or prestige. In the second, the two varieties are of unequal status. But given the description of the standard variety above, the difference is not only due to prestige, but also to social functions, different backing by codification, and availability of/reliance upon a written version of the standard. Nonetheless, the distinction between horizontal and vertical processes is not always easy to draw. Often, even dialects do not have the same status/prestige, and convergence toward the more prestigious dialect of a local center (for instance, an urban variety) can resemble convergence toward the standard. The distinction is particularly difficult when the standard variety is associated with the language of one particular region, even though it may not be identical with it (as in the case of Tuscany in Italy).

Horizontal (inter-dialect) convergence may lead to leveling, that is, the loss of distinctions (simplification) in the respective varieties. In this case, the most marked features of the dialects in contact disappear. The complexity of the overall repertoire of linguistic forms decreases, since the most divergent forms are lost, although internal variation within a dialect may *increase* due to competing features. When several dialects converge simultaneously, the term “koineization” (Berruto 1995, 226–227) seems fitting. Inter-dialect leveling is different from dialect mixture, which is a rarer process that can be considered an extreme case of borrowing. In this case, overall complexity is preserved or even increases, since the new mixed variety incorporates forms from all contributing varieties.

Horizontal convergence (without vertical convergence) is most frequent when no standard roof exists. There are several quite well-attested historical examples of koineization of the first type (e.g., the Andalusian or Upper Saxonian koinés, both of which were triggered by migration). However, horizontal convergence can also occur in a standard/dialect repertoire in the course of the emergence of regional dialects (see below). The best examples are those in which inter-dialectal convergence/leveling implies divergence from the standard.³ Thus, North Bavarian non-vocalizing varieties increasingly vocalize /l/ in coda position, although this implies divergence from the non-vocalizing standard. The reason is that the dominant Bavarian regiolect, which is based on Middle Bavarian and the Munich area (see below), has /l/-vocalization as one of its most salient features (hence, Middle Bavarian /fnøi/ wins over North Bavarian /fnɛl/, cf. Std.G. /fnɛl/ “fast”). Along the same lines, the same North Bavarian varieties replace their diphthong /ou/ with the Middle Bavarian—but also regiolectal—form /ua/ (in reflexes of Middle High German (MHG) /ue/), although both are distinct from std. German /u:/ (as in std. *tun* /tu:n/ “to do” ~ North Bavarian /dou/ ~ Middle Bavarian /dua/). Svahn and Nilsson’s (2014) study of West Swedish dialects near Gothenburg finds leveling and an orientation toward (old and new) standard features associated with Stockholm, as well as leveling and orientation toward Gothenburg, the nearby urban center, depending upon the social and economic structure of each place.⁴

Note that sometimes what looks like pure horizontal convergence is actually an indirect process of standardization. For instance, when dialectal /i:/ is replaced by equally dialectal diphthongal /ɛi/ or /ɛi/ (in reflexes of MHG ī, as in *Eis* “ice”) in the Low Alemannic dialect of Freiburg in southwest Germany, this looks like convergence with neighboring Swabian Alemannic at first glance, but is more fruitfully explained as a halfway convergence to the standard variant /ai/ (cf. Auer 1988). Another example of an indirect approximation of the standard variety is described by Leinonen (2010) for Sweden. Standard Swedish has an allophonic variation between an open and an extra-open front mid vowel, since /ɛ:/ and /œ:/ are lowered before /r/ (as in *dör*), whereas many dialects do not lower these vowels in this context. Leinonen shows that a new system has evolved recently, which is establishing itself in the demoticized standard: in this system, the long front mid vowels are lowered irrespective of their phonological context. She interprets this innovation as a compromise between the phonological system of the dialects and the (intended) standard variety.

The leveling of existing dialects is one of the important historical processes on the horizontal level. The other is *diffusion*, that is, the spread of innovations across (parts of) a language area (cf. Haas 2010). Horizontal diffusion is widely documented for periods of the European languages in which no oral standard existed or was still weak (cf. Kloek 1927) on the spread of diphthongization in sixteenth-/seventeenth-century Dutch, or Herrgen (1986) on the spread of coronalization ([ç]>[s]) in nineteenth-century Middle German), whereas it seems to be less frequent in modern periods. The spread of glottalization in England is a likely counterexample; see Milroy *et al.* (1994).

Taelde man (2005, 263) distinguishes between “contagious diffusion” via personal contact and “hierarchical” (or “parachuting”) diffusion according to a hierarchical pattern in which the larger urban centers take the lead (cf. also Chambers and Trudgill 1980, 196–202). The second type of diffusion is deeply linked to prestige. Contagious diffusion is presumably less important in a modern society in which communication is not restricted to face-to-face networks, and social stratification plays an important role. One of many examples of hierarchical diffusion is the borrowing of the more prestigious Leipzig features into the city dialect of Berlin in the seventeenth and eighteenth centuries (cf. Auer 2013a), from where it spread into Berlin’s surrounding areas.

9.4 Dialect-to-Standard Convergence: Regional Dialects and Regional Standards

Much more common in Europe than inter-dialectal convergence is leveling due to convergence in the standard ~ dialect dimension. The local (traditional) dialects are replaced by varieties with a larger geographical reach (so-called “regional dialects” or “regiolects,” sometimes also called “supralocalization”; cf. Britain 2009, Hinskens 1996). These regional dialects—taken as a whole—are considered to be closer to the standard variety than the local, traditional dialects. In the formation of regional dialects, horizontal and vertical convergence go hand in hand: as the traditional dialects lose more local features and replace them with those closer to or identical with the standard forms, the non-standard varieties become more similar to each other.

Note that although one can claim that regional dialects are ideologically and structurally “in between” the traditional dialects and the standard, this is only correct for regional varieties taken as a whole. For some localities and features, the shift from the traditional dialect to the regional dialect may also imply *divergence* from the standard (when the local dialect happens to coincide with the standard form), or simply the replacement of one dialectal form by another, more prestigious one (cf. the examples from Bavarian mentioned above).

While the traditional dialects are no longer used in many parts of Europe, the regional dialects show high vitality in many regions. In Germany, they seem to be backed by the old regional standard varieties of spoken German that existed until the nineteenth century (cf. Schmidt 2009, Lenz 2010, 204–206). For instance, the dialects of Upper Saxony (East Middle German) are today extinct; the regional variety that speakers nowadays consider the dialect of the area is a leveled variety, which, 200 years ago, still functioned as the regional standard, but has now been re-evaluated as a (regional) dialect.

By its very nature, dialect-to-standard convergence and the concomitant process of inter-dialectal leveling also implies certain processes of divergence, including dialect split. In contemporary societies, these are due to the fact that the reach of the standard languages usually—and increasingly—ends at the national borders, which are not always isomorphic with the traditional dialect borders (Auer 2013b). When an old dialect continuum is cross-cut by a state border, the standard will only influence the traditional dialects in the state where it roofs these dialects; on the other side of the border, another standard, if there is one, will

exert its influence. The result is dialect divergence at the state border. A case in point is the French/German border separating Alemannic-speaking Alsace from the German Upper Rhine area, where very similar dialects used to be spoken. But it is only in Germany that the German standard language has led to the emergence of a regional dialect, whereas the Alsatian varieties have preserved their traditional shape since they are no longer roofed by standard German, and since the exoglossic standard (French) cannot trigger dialect~standard convergence (apart from via loanwords). The same, but with inverted roles, holds for the Dutch/German border, where a former Low German/Dutch dialect continuum has been broken up and a dialect border has emerged that coincides with the state border. Here, Low German has no roof—the standard language in Germany is based on High German—and the Low German dialects, if spoken at all, cannot level into a regional dialect. On the Netherlands side of the border, meanwhile, the Dutch standard language has a huge impact on dialects, and has led to standard convergence.

Another example of a dialect group that left the German standard roof, as the Alsace did, is Luxembourg. Here, an endoglossic standard has emerged—standard Lëtzebuergish—which is used in addition to French and German. The new standard language is closely linked to the variety of the capital, Luxembourg City, and has already started to trigger dialect~standard convergence in the dialects most distant from the capital (Gilles 1999). All three of these examples show how political borders have shaped the dialectal landscape over the recent decades in Europe.

In sum, regional dialects have emerged in many European regions, taking over the function of the traditional dialects as the most regional way of speaking, with the traditional dialects surviving, if at all, mainly as artefacts. The resulting repertoire type is conveniently termed *diaglossic* (Bellmann 1997) as opposed to *diglossic* (Ferguson 1959). In a diaglossic repertoire, the gap between standard and traditional dialects is filled by intermediate forms, such as regional dialects. In a diglossic repertoire, by contrast, the speakers can only choose between the H ("high") and L ("low") varieties, without the possibility of compromise.

In a diaglossic repertoire, leveling through the loss of the most dialectal (and most local) forms is of course a continuous process of dialect change; it does not necessarily stop with the emergence of regional dialects. Rather, the process can persist and lead to the elimination of the regional dialectal forms as well, resulting in an even more homogenized language area. What is left is a form of the standard, which may still show regional—substrate—features (i.e., regional standards).

It is useful here to introduce van Coetsem's (1988) distinction between source and recipient linguistic systems, depending on which variety (or rather, which group of speakers) takes the active role. The emergence of regional dialects and regional standards is due to dialect speakers' attempts to sound less local—often, less rural—and to approximate the standard. Here, the dialects, or rather their speakers, are the agents. However, when regional standards emerge, it can also be the standard (or its speakers) in the *agens* role, leading to "destandardization." Here, the standard speakers try to sound more regional or even dialectal, in a kind of downward convergence.

How the "intermediate space" in a diaglossic repertoire is to be theorized is a matter of dispute. The gradual process of dialect-to-standard convergence described above would suggest a continuum of forms. However, several authors have argued for a two-way process in which the regionalization of the standard and the standardization of the dialects lead to a double continuum (see Berruto 1989 for Italian, as well as Lenz 2003 for West Middle German).

9.5 Dialect Loss

From the preceding discussion, it follows that the traditional dialects have lost vitality in most parts of Europe, with Norway and German-speaking Switzerland being the two notable exceptions, or have already disappeared from most people's everyday lives. However, there

are two different scenarios for dialect loss that should be distinguished. One is based on a diglossic repertoire in which standard and dialect stand in sharp contrast and there is no alternative to speaking either dialect or standard—there are no intermediate forms. It is clear that in such a situation, dialect loss can only proceed in the form of *dialect shift*: the traditional dialect is simply not passed onto the next generation. In this case, the dialect does not change at all, or the only changes are due to attrition.

Dialect shift has occurred, for instance, in most parts of the Low German dialect area. Depending on the time and the socio-economic circumstances under which the shift took place, imperfect acquisition of the more prestigious (standard) variety accompanying the shift has led to different results. In regions where the shift took place early (in the late nineteenth or early twentieth century), it led to the formation of a distinct regional/urban non-standard variety of High German with many Low German substrate features. This is the case in the Ruhr region and in Berlin, where the shift from Low to High German was accompanied and partly triggered by massive migration in the era of industrialization. Alternatively, when the shift occurred late (after World War II), the only remnants of Low German are a colloquial, slightly regionalized standard with a few substrate features (as in the case of Hamburg, where the shift to High German was completed after 1950).

The second scenario for dialect loss starts out from a diglossic repertoire in which standard and dialect are related by a (quasi-)continuum of intermediate forms. In a diglossic repertoire, dialect loss is merely the endpoint of a long continuous process of eliminating the most regional features, until only the standard, which may show internal variation, remains. By its nature, the loss does not always (or even regularly) reach completion. During the long process of gradual dialect loss, the dialect changes, that is, it loses certain features and replaces them with others drawn from regional dialects or the standard language. Sometimes, compromise forms emerge. But despite these changes, the chances of a regional way of speaking surviving and establishing itself at some point during this process are much higher than in the case of dialect loss through shift from a diglossic starting point.

Let us consider two examples of dialect change starting from a diglossic repertoire in Europe, since in both cases large empirical studies have investigated the nature of the process in detail.

The first case is the southwest of Germany, where Alemannic dialects are traditionally spoken. This region is exceptionally well documented, allowing a comprehensive reconstruction of developments over the twentieth century. Auer, Schwarz, and Streck (2008) investigated these developments on the basis of spontaneous interview and questionnaire data. The data were based on interviews carried out between the 1950s and 1970s with 583 traditional dialect speakers in 360 locations. In a version of the apparent time paradigm, these spontaneous data were compared with (largely) the same speakers' elicited questionnaire data, which formed the basis of the *Südwestdeutsche Sprachatlas* (SSA), covering 579 locations in the same area. The elicited SSA data, in turn, were compared to the data collected by Wenker in the *Deutscher Sprachatlas* (DSA) toward the end of the nineteenth century in southwest Germany in even more locations.⁵ The design of the study therefore implied a double comparison in real and apparent time, with the latter operationalized as the difference between dialect knowledge and dialect use. The model predicts that features will change when the questionnaire data of Wenker and the SSA are compared, and when questionnaire answers and the spontaneous speech of the SSA informants are compared (Figure 9.1).

The results of the study allow us to reconstruct the main developments in a traditional Upper German dialect in the last century. The following conclusions can be drawn:

1. Isolated (archaic) features with a small reach disappeared from the map during this time.
2. Dialect features with a larger reach sometimes disappeared as well; sometimes they became even stronger. Whether the first or the second happened depends, among other

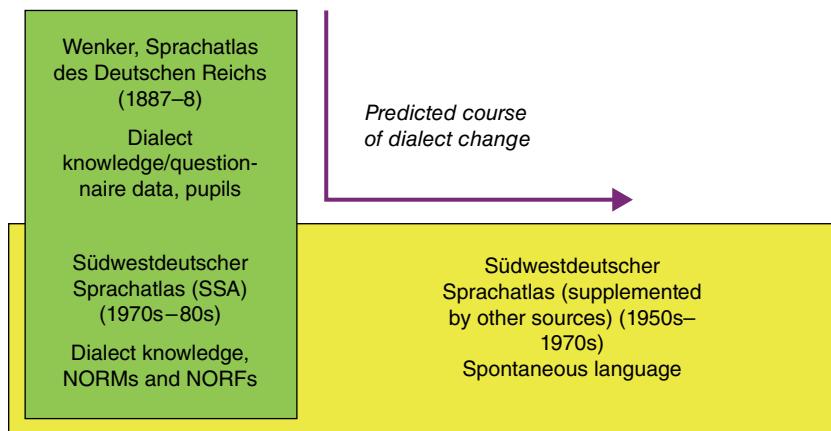


Figure 9.1 Design of the southwest German dialect study (NORM=non-mobile older rural male; NORF=non-mobile older rural female).

things, on the phonological status of the variable, the divergence of the dialectal realization from the standard, and (most importantly) whether the features had become part of a regional dialect.

3. Even where regional dialects emerged, they show internal variation between regiolectal and standard forms.

All three developments are shown in Figures 9.2 and 9.3. The map in Figure 9.2 represents the reflexes of MHG /ei/, here in the word *heim* "home," in the traditional dialects at the oldest stage, that is, the traditional dialect isoglosses of Wenker in the late nineteenth century (black) and the traditional, older informants of the SSA (gray). As can be seen, there are three large dialect areas: the MHG diphthong is realized as /ai/ (in the west, which also corresponds to the standard pronunciation), /oa/ (in the middle), and /oi/ (in the east). In addition, there are smaller dialect areas in which the same MHG diphthong was traditionally realized as /e:/ or /ei/ (checked and dotted areas respectively near the border with France/Alsace), the monophthong /ɔ/ or even /a/ (gray shading west and northwest of Lake Constance, shown as the solid gray area at bottom center), and /ua/ (in a small area northeast of Lake Constance, checked gray and white).⁶ A comparison between Wenker's and the SSA data does not show a major displacement of the isoglosses separating the three larger dialect areas. Apparently, when asked for the most traditional dialect pronunciation, the interviewees had—even in the second half of the twentieth century—no problem reproducing the same forms that Wenker's teachers had recorded as normal dialect usage in the late nineteenth century. This holds for the /ai/, /oi/, /oa/ and /ɔ/ areas. The very small /ei/, /e:/ and /ua/ areas (/heim, he:m, huam/) were not identified as areas by Wenker at all, due to a mixture of these pronunciations with the more dominant /ai/ and /ua/ forms. In the somewhat archaizing SSA maps, these responses were drawn as separate, small dialect areas.

Figure 9.3 shows that actual *usage* by the same informants in spontaneous speech does not correspond to the *knowledge* patterns elicited in the 1970s. (In this map, usage data are superimposed on the traditional SSA maps). We note the following changes:

1. the smallest areas (those for /heim, he:m, huam/) have almost disappeared in spontaneous speech, although the informants still remember them. They are replaced by the forms of the corresponding larger area (/heim, he:m/ > /haim/; /huam/ > /hoim, hom/) or the standard realization /haim/.

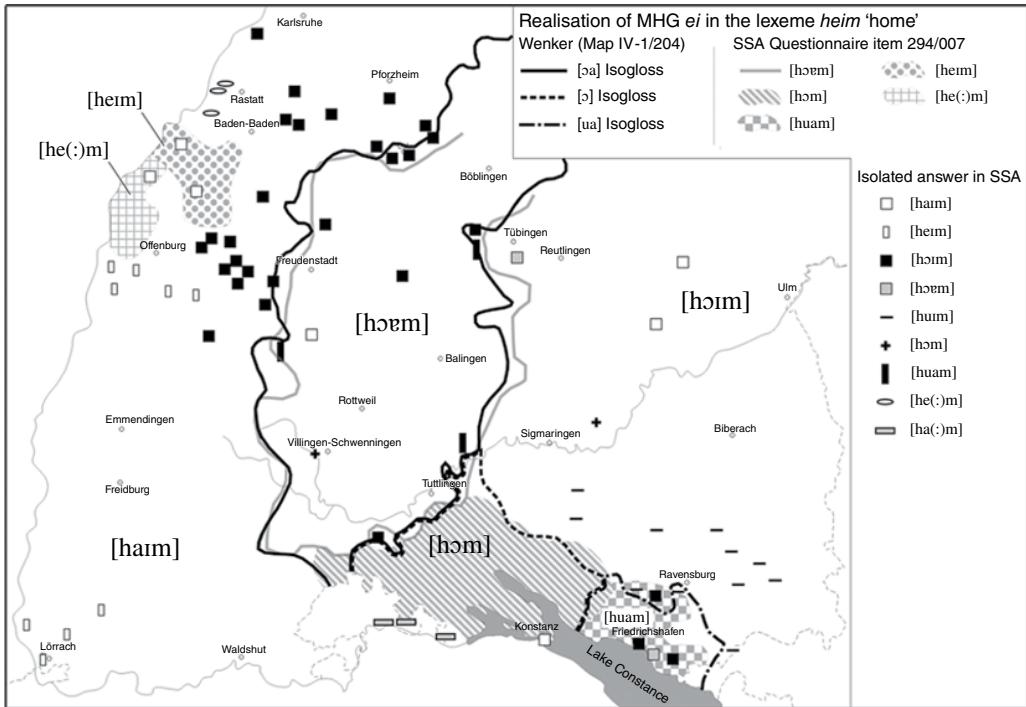


Figure 9.2 Reflexes of MHG /ei/ in the word *heim* 'home' in the traditional dialects of southwest Germany, according to the *Deutscher Sprachatlas* (DSA) (black isoglosses) and the *Südwestdeutsche Sprachatlas* (SSA) (gray isoglosses and shaded areas). The symbols represent additional SSA questionnaire results in individual locations outside the main areas. Adapted from Schwarz 2015.

- With regard to the larger dialect areas, we note that the /hom/ and /hoam/ areas, sandwiched between western /haim/ and eastern /hoim/, show a lot of variation in actual usage; several of the informants who mentioned /o/ and /oa/ as local forms in the questionnaire replaced these older forms in their spontaneous speech with forms from adjoining areas (horizontal convergence, cf. the /hom/ forms in the traditional /hoam/ area) or the standard forms (vertical convergence, cf. /haim/ forms in the traditional /hom/ area).

The old /oi/ forms did not undergo the same change. In this (eastern) area, we only very rarely find forms from adjoining areas (/hom, hoam/). On the other hand, the /oi/ forms spread into the north of the /haim/ and the /hoam/ area, as well as the area north of Lake Constance where they replaced /huam/. Clearly, it is the /oi/ realization that has become part of the (Swabian) regiolect, which is stable not only in its traditional area, but is even expanding geographically.

Finally, it can be noted that both the /hoam/ and /hoim/ areas, as well as (very pronouncedly) the old /hom/ area, show numerous examples of the standard /haim/ in spontaneous speech, testifying to the "intrusion" of standard forms into the dialect (Auer and Schwarz 2014).

The same pattern—stability in dialect knowledge but change in usage—can be observed repeatedly. In other cases, the change is already visible directly or indirectly in the traditional dialect studies based on elicited data. An example is the singular umlaut in *Bruder* "brother."

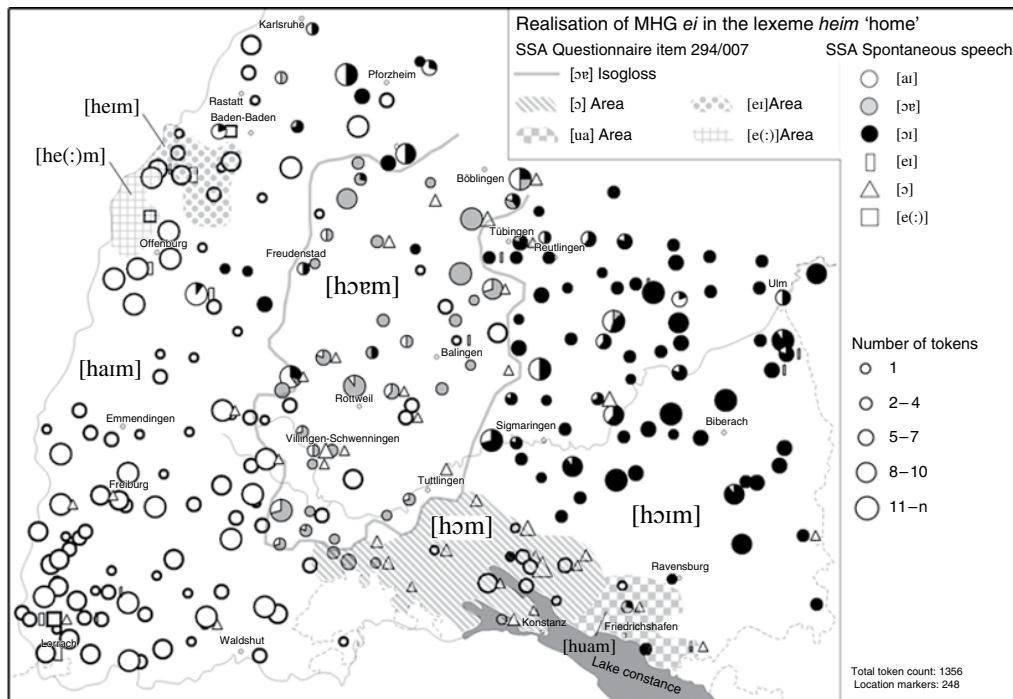


Figure 9.3 Reflexes of MHG /ei/ in the word *heim* 'home' in southwest German Alemannic in spontaneous speech (based on 1,356 tokens from 248 locations). For comparison, the map shows the old SSA isoglosses as well (gray isoglosses and shaded areas). The sizes of the symbols represent the number of tokens (adapted from Schwarz, 2015).

This is a particularly salient feature, as umlaut is one of the morphological means by which plural is marked in standard German (and in the dialect). Against this background, umlaut in the singular is highly noticeable and potentially leads to misunderstandings (Figure 9.4).

The traditional maps show a dominant form ([bruəðə]) in which the MHG diphthong is preserved. This covers most of the area. In the north, the Franconian (and standard German) monophthongal variant ([bru:də]) is dominant. In the south, toward Lake Constance, umlaut is observed. The isoglosses drawn in the various dialect atlases, all based on traditional speakers, more or less agree on the northern extent of this area, but disagree on its western and eastern extents: for Fischer's *Schwäbischer Dialektatlas*, the area is smaller in the west; for the SSA, it is smaller in the east. This is because all atlases had to deal with a considerable amount of variation in the area, which made it difficult to draw isoglosses. Particularly between Sigmaringen and the lake, the SSA (which tended to eliminate innovation responses through its interviewing strategy) found many non-umlauted forms in the 1970s. It is obvious that the umlaut area was already in a process of dissolution among these speakers, who often used the northern diphthongal, non-umlaut form instead. This points to an ongoing change.

The tokens for "brother" found in the spontaneous speech of (roughly) the same informants show that many of them still know about the older umlaut form, but no longer use it. Of the 23 locations in which tokens of the lemma occurred in the interview data, the traditional umlaut form could only be found in six. Clearly, the change already observed indirectly in the elicited data had continued. The winner is usually the non-umlaut, diphthongal form, but only occasionally the standard monophthong: the regional dialect favors

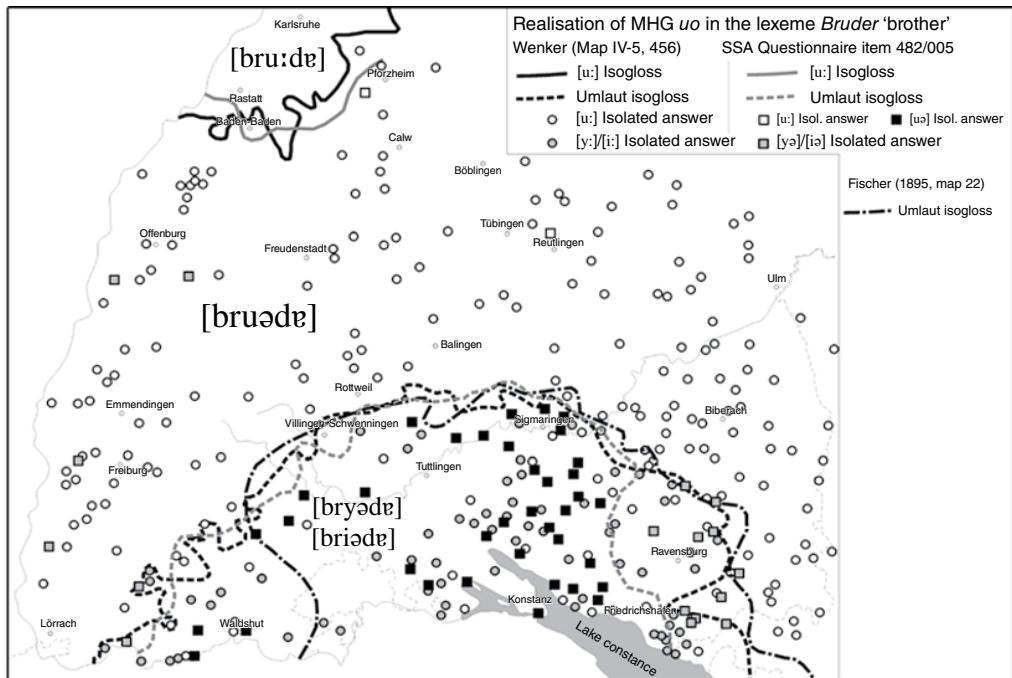


Figure 9.4 Realization of MHG /uo/ in std.G. *Bruder* “brother” in the traditional dialects (elicited data): *Deutsche Sprachatlas* (DSA) (solid black line, black dashed line), Fischer’s *Schwäbischer Dialektatlas* (1895) (black dashed line), and *Südwestdeutscher Sprachatlas* (SSA) (gray solid line, gray dashed line; adapted from Schwarz, 2015).

the diphthong, which is used over a very large area. The geographically more restricted umlaut is about to disappear.

Recent data among older and younger speakers in the same area (Kehrein 2012, Ch. 6) confirm that the process of dialect leveling is still going on: traditional dialectal features are being replaced by regiolectal ones, and these in turn by near-standard ones.

However, the process of dialect leveling has not reached completion in southwestern Germany. Regional dialects still exist, and speakers continue to express their local identity through them, at least on certain occasions. In other parts of Europe, the development has gone further. An example is Denmark, one of the countries in Europe in which dialect-to-standard convergence has progressed almost to the point of making all dialect features obsolete. However, there is evidence that this dialect loss also proceeded through a phase of regional dialect leveling. The processes involved have been documented with great empirical depth and precision by the Copenhagen LANCHART research center, using real-time comparison over the last 30–40 years (see Gregersen 2009 for an overview). Consider, for instance, the development of the dialectal forms of the participle of strong verbs in Jutland (Jensen and Maegaard 2012). In Jutland dialects, there is lexical variation between a past participle form ending in *-en* and one ending in *-et* (hence, *bliven* versus *blivet* “become”). In standard Danish, only the *-et* ending is used. Jensen and Maegaard compared recordings from two locations (Odder and Vinderup; see Figure 9.5) made in the 1980s and the 2000s, and found a sharp decline in the use of the *-en* form, particularly in Vinderup. In Odder, the *-en* form had already declined. Lexical text frequency was a significant predictor variable, with more frequent words following the dominant *-et* pattern first, contrary to what is usually assumed in usage-based approaches to language change (cf. Bybee 2010, among

others). Age also had a significant effect on the use of the traditional *-en*, testifying to the ongoing disappearance of this dialect feature, and indicating overall dialect loss.

Alongside variation between *-en* and *-et*, the final consonant in *-et* participles in Jutland showed geographical variation between [ð] (the standard form), [f] and [d]. There is also a “zero” form ending in schwa. The geographical distribution of these variants in the traditional Jutland dialects is shown in Figure 9.6. The authors point out that the tap and the zero realization were undistinguishable in their recordings; these were therefore conflated in the analysis.

This spatial distribution suggests that the central [əð] zone represents the regiolectal outcome of leveling and partial convergence to the standard, whereas [ə] is peripheral and receding. Vinderup and Odder are in the traditional [ə(r)] zone and traditional [əð] zone, respectively. If the regiolect were still an attractor for ongoing processes (as in the Alemannic-German data from the 1970s discussed above), we would expect a change [ə(r)]>[əð]; if the regiolect were irrelevant and the standard the only target of ongoing change, we would expect a change [ə(r)], [əð]>[əð]. The data show that the second scenario is the case. However, Jensen and Maegaard also found some [əð] realizations in Vinderup, where they are not attested in the traditional dialect. They hypothesize that these are remnants of an older process of regionalization toward the stop realization, which preceded the more recent phase in which all dialectal variants tend to be abandoned in favor of the Copenhagen pronunciation. We may conclude that the situation in Jutland some 50 years ago was highly similar to that in southwest Germany, but over the last 30 years, it has precipitated toward dialect loss.

However, even in a place like Jutland, with its strong pressure on dialect speakers to accommodate the standard, there are exceptions. Monka (2013) investigated linguistic

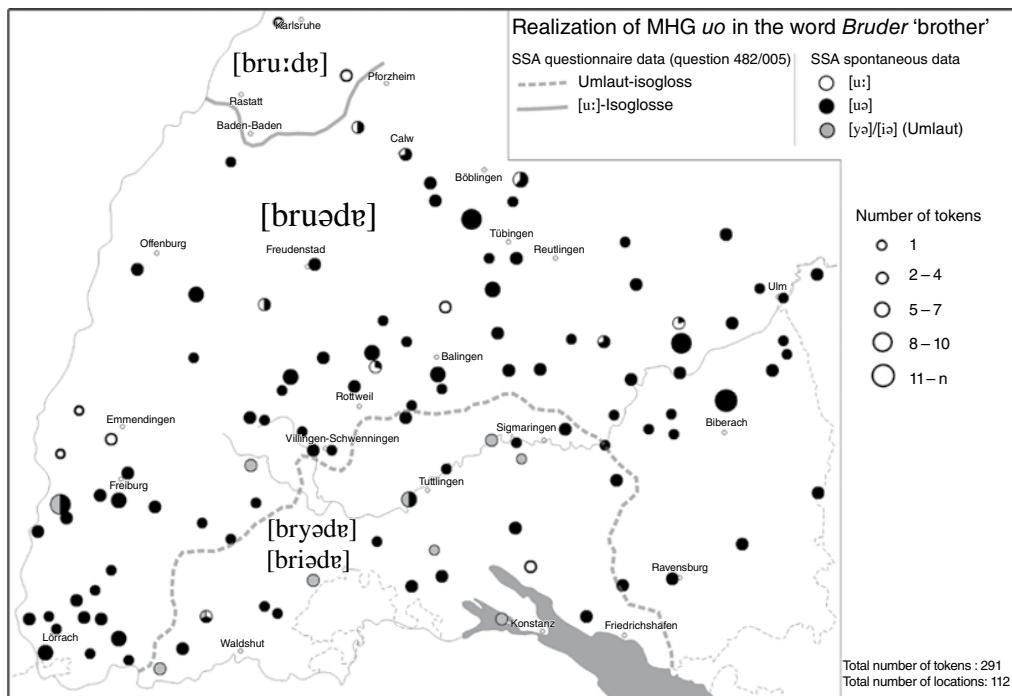


Figure 9.5 Realization of MHG /uo/ in std.G. *Bruder* “brother” in the traditional dialects (spontaneous data: 291 tokens from 112 locations). For comparison, the isoglosses given in the *Südwestdeutscher Sprachatlas* (elicited data) are added in gray (adapted from Schwarz, 2015).

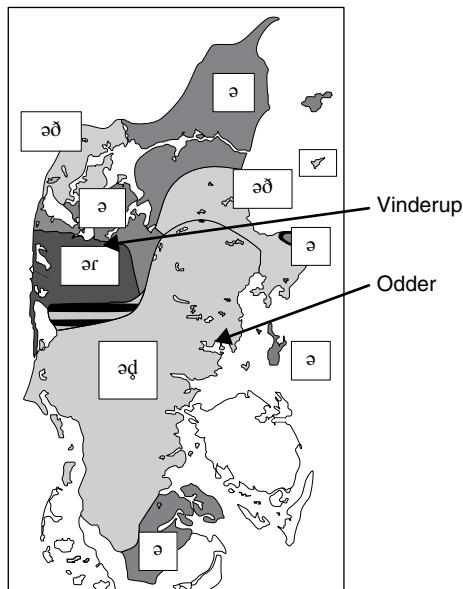


Figure 9.6 Pronunciation of the past participle (standard Danish) suffix *-et* in different Jutland dialect areas (Rasmussen *et al.* 2000; Map K 6.1, adapted from Jensen and Maegaard 2012, 172). Copyright Jysk Ordbog, reprinted with permission.

change in another location in Jutland, Tingslev, in the very south (closer to the German border, an area of the former Duchy of Schleswig that was under German rule until World War I). Comparing Tingslev with Vinderup, Monka found huge differences. Whereas the overall dialectality index in Vinderup sharply decreased between the two interview surveys, as expected, it remained level or even slightly increased in Tingslev, based on a comparison of recordings from 1986 and 2010. The reason seems to be the largely positive attitude of the Tingslev people toward their dialect, which they use in their place-making activities.

9.6 What About the Standard?

Present-day developments in the European standard languages have been much discussed in the recent literature (for a recent overview, see Kristiansen and Coupland 2011; Cerruti, Crocco & Marzo eds. 2016 for Italian). It is widely believed that the standard language gets “destandardized” by assuming regional features and hence converging downwards in the direction of the (regional) dialect. To justify this view, it is often observed that regional accents are acceptable in the public media, particularly on national TV, which was not the case 50 years ago. However, there are several problems with this view. First of all, it disregards the fundamental changes (national) TV and radio have undergone during this period. The standard language used today in the broadcast media is close to the demoticized standard language used in everyday life, whereas until the middle of the twentieth century it was an artificial variety only accessible to a handful of (trained) speakers. Second, and linked to this first point, the presumed “destandardization” is often based on a confounding of regional and informal linguistic features. Often, it is the second (“informalization”) which is taken as evidence for “destandardization.”

Spiekermann (2008; cf. the English summary in Auer and Spiekermann 2013) investigated real-time changes in the southwest German standard variety, as spoken in

Baden-Württemberg in the cities of Freiburg, Heidelberg, Karlsruhe, Mannheim, Stuttgart and Tübingen, by comparing two corpora of rather formal interview data:

1. a subset of the so-called Pfeffer corpus, which was recorded in 1961. Only those southwest German recordings were selected in which the speaker had a higher level of education (*Abitur* or higher), and which were classified by Pfeffer and his colleagues as “standard speakers”;
2. a corpus collected by the author between 2001 and 2003 from teachers and would-be teachers.

Spiekermann studied five highly frequent features of the regional dialects spoken in his six locations, viz. (1) the coronalization of /ç/ > [ç] in coda position after front vowels, as in *dich* “you” (accusative case), a salient feature of the Rheno-Franconian regional dialect of Heidelberg and Mannheim; (2) the lowering of /e:/ > /ɛ:/, as in *lesen* “to read,” a salient Swabian feature found in the vernacular of Stuttgart and Tübingen; (3) the palatalization of /s/ > [ʃ] before a tautosyllabic and tautomorphemic obstruent, as in *Fest* “feast,” a salient general Alemannic feature of the regional dialect spoken in Freiburg, Stuttgart, and Tübingen; (4) lenis realization (lenition) of intervocalic fortis consonants ([t]>[d]), as in *hatte* “had,” a non-salient feature of all Upper German dialects; and (5) the raising of the low vowel in the word *das* “that” > *des* ([a]>[e]), another non-salient Upper German feature. He found a radical reduction (50% to 70%) of regional forms among the younger speakers for variables (1) to (3), a moderate reduction for the non-salient feature (4), but no reduction in *des*, which I interpret to have lost its regional value.

On the other hand, Spiekermann (2008) also looked at changes in the use of informal “allegro” forms in the standard which are not regional, that is, they are used all over Germany. These features would have been considered sub-standard in the 1960s. These include (1) deletion of the first-person singular suffix, for example, *(ich) hol+e>hol* “I fetch”; (2) final /t/-deletion in the copula *ist>is* (“is”); (3) final /t/-deletion in the negative adverbial *nicht>nich* “not”; and (4) cliticization of the indefinite article *eine(n)>ne(n)*. For all four variables, a huge increase was observed. Thus, the frequency of schwa deletion in the first-person singular oscillated between 79% and 91% in the modern standard corpus, depending on location, and between 33% and 78% in the Pfeffer corpus. Final /t/-deletion in *ist* had tripled over 40 years in Mannheim, Stuttgart, Tübingen, and Freiburg. In *nicht*, /t/-deletion was only frequent (~30%) in Stuttgart in the 1960s, whereas it remained below 10% in Mannheim, Heidelberg, Freiburg, and Karlsruhe, where the regiolectal form *ned* was still fairly frequent. Around 2000, the percentages of /t/-deletion ranged between 36% and 84%. Reduction of the indefinite article varied between 47% and 94% in the 2000s, an increase of approximately 20% to nearly 70% on average.

The results make it clear that even in the standard variety, the loss of regional variants continued in the second half of the twentieth century. The impression of a destandardization process is due to the rapid and large-scale increase of reduced (informal) forms, which occurs everywhere in the German language area (in Germany), as well as in many other parts of Europe. For this new, informal standard, the term neo-standard (as used in Italian sociolinguistics) seems adequate (cf. Cerruti *et al.* 2016).

9.7 Divergent Changes

As pointed out, the overall tendency in the European dialects has for a long time been to recede and be replaced by regiolectal or near-standard ways of speaking. As a consequence, the national standard languages have, contrary to popular belief, been strengthened. There are, however, two recent countertrends. The first results from recent political, economic

and social processes of what geographers call “New Regionalism” (e.g., Amin 2004, Werlen 2007). New Regionalism consists of a positive re-evaluation of the traditional regions within the nation states. It is a reaction to state centrism, but also to globalization, which is seen as another threat to territorial unity and sovereignty. Supported by the European Union (“regional governance”), regions such as Scotland and Wales in the UK, Catalonia, the Basque country and Galicia in Spain (and France), and Northern Italy have argued for regional political control and autonomy. New Regionalism can establish a countertrend to the general tendency to give up regional ways of speaking in favor of more standard-like, or at least non-regionalized, ones. It sometimes receives support from the commodification of dialectal features in modern tourism, often based on notions of “authenticity.” Obviously, such new social constellations will lead speakers to deploy regionally indexed linguistic features in ways not systematically accounted for in traditional dialectology and variationism. Whereas traditional speakers of a dialect are predicted to accommodate toward more standard-like usage in more formal situations or in communication with outsiders, speakers for whom a possibly enregistered dialect has the primary function of gatekeeping for the regional community will use a comparatively small number of highly salient regional features. At the same time, they will be far less willing to give them up in communication with outsiders, and will opt out of the modern mainstream sociolinguistic consensus according to which the regional signal is weakened as we move from informal to formal situations, or from working-class to middle-class interactants.

An example of this is the province of Limburg in the Netherlands (Cornips 2013). Speaking dialect nowadays is not considered an expression of lower social class status, as it is in many parts of the Netherlands. The dialect is looked upon as a symbol of pride, particularly in the old city of Maastricht. (Note, however, that things are very different in working-class-dominated Limburg cities such as Heerlen, with its industrial mining tradition.)

The second countertrend that involves divergent changes is the emergence of urban youth languages, also called multiethnolects, *straatstaal*, *Kiezdeutsch*, and so on, which have their origin in ethnic groups of young speakers with immigrant backgrounds, but partly seem to be losing their (multi-)ethnic symbolic value (see *inter alia*, the contributions in Kern and Selting 2011, Quist and Svendsen 2010, and Cheshire *et al.* 2011). These ways of speaking imply divergence from the established (autochthonous) repertoires of linguistic forms; however, their emergence should be seen as entirely separate from the standard~dialect dynamics discussed in this chapter.

NOTES

- 1 For reasons of space, the present chapter only cites some exemplary recent studies, with a strong bias toward the Germanic languages (other than English). The reader is referred to Auer (2005, 2011), Auer and Hinskens (1996), and Hinskens, Auer, and Kerswill (2005) for systematic references to the older literature.
- 2 The dialects in the Slavic (particularly East Slavic) language areas are mostly younger and present a slightly different picture.
- 3 However, using various geographical-distributional measures, Auer, Baumann, and Schwarz (2011) show that even in a diaglossic repertoire (see below), horizontal convergence without concomitant vertical convergence toward the standard occurred for certain variables in the second half of the last century in the traditional dialects of German Alemannic.
- 4 See also the discussion in Vanderkerckhove (2010).
- 5 Wenker asked schoolteachers to answer his questionnaire. Whether they portrayed the language of their pupils (as requested) or rather the current dialect in the location is unknown. Note that if

- it was the former, the speakers in Wenker's atlas are almost the generation interviewed in the SSA; the difference would then be that between children and elderly people. More importantly, the direct interviewing policy of the SSA aimed at reconstructing the most archaic dialect with the help of a questionnaire and trained fieldworkers (cf. Auer 2010), whereas Wenker's teachers presumably wrote down the dialect of the location or their pupils as it was actually used.
- 6 Several individual respondents north of the /ua/ area even gave /ui/ as the local variant, most likely a remnant of an older compromise form between the more prevalent /oi/ and southern /ua/.

REFERENCES

- Amin, Ash. 2004. "Regions unbound: Towards a new politics of place." *Geografiska Annaler (Series B: Human Geography)*, 86: 33–44.
- Auer, Peter. 1988. "MHG ï and û in the city dialect of Constance." In *Variation and Convergence: Studies in Social Dialectology*, edited by Peter Auer, and Aldo di Luzio, 44–75. Berlin: de Gruyter.
- Auer, Peter. 2005. "Europe's sociolinguistic unity, or: A typology of European dialect/standard constellations." In *Perspectives on Variation*, edited by Nicole Delbecque, Johan van der Auwera, and Dirk Geeraerts, 7–42. Berlin: de Gruyter.
- Auer, Peter. 2010. "Der Grunddialekt als Konstrukt: Wie Gewährspersonen und Erheber in der direkten Befragung die Daten der Atlasdialektologie konstituieren." In *Parole(s) et Langue(s), Espaces et Temps – Mélanges Offerts à Arlette Bothorel-Witz*, edited by Dominique Huck, and Thiresia Choremis, 23–36. Strasbourg: Université de Strasbourg.
- Auer, Peter. 2011. "Dialect vs. standard: A typology of scenarios in Europe." In *The Languages and Linguistics of Europe: A Comprehensive Guide*, edited by Bernd Kortmann, and Johan van der Auwera, 485–500. Berlin: de Gruyter.
- Auer, Peter. 2013a. "Sociolinguistic change, indexical fields, and the *longue durée*: Examples from the urban sociolinguistics of German." In *L'Interface Langage-Cognition/The Language-Cognition Interface*, edited by Stephen Anderson, Jacques Moeschler, and Fabienne Reboul, 201–232. Geneva: Droz.
- Auer, Peter. 2013b. "State borders and language change: The (non-)effects of political border permeability on language." In *Theorizing Borders through Analyses of Power Relationships*, edited by Peter Gilles, Harlan Koff, Carmen Maganda, and Christian Schulz, 227–248. Brussels: Peter Lang.
- Auer, Peter, Peter Baumann, and Christian Schwarz. 2011. "Vertical vs. horizontal change in the traditional dialects of southwest Germany: A quantitative approach." *Taal en Tongval*, 63(1). DOI: <http://dx.doi.org/10.5117/TET2011.1.AUER>.
- Auer, Peter, and Frans Hinskens. 1996. "The convergence and divergence of dialects in Europe: New and not so new developments in an old area." *Sociolinguistica*, 10: 1–30.
- Auer, Peter, and Christian Schwarz. 2014. "Dialect/standard advergence: The relevance of compound borrowing." In *Linguistic Variation: Confronting Fact and Theory*, edited by Rena Torres Cacoullos, Nathalie Dion, and André Lapierre, 263–282. London: Routledge.
- Auer, Peter, Christian Schwarz, and Tobias Streck. 2008. "Phonologischer Dialektwandel in Südwestdeutschland: Erste Ergebnisse einer Sekundäranalyse von Dialektdateien des 19. und 20. Jahrhunderts." In *Dialektgeographie der Zukunft*, edited by Peter Ernst, and Franz Patocka, 115–130. Stuttgart: Steiner.
- Auer, Peter, and Helmut Spiekermann. 2011. "Demotisation of the standard variety or destandardisation? The changing status of German in late modernity (with special reference to south-western Germany)." In *Standard Languages and Language Standards in a Changing Europe*, edited by Tore Kristiansen, and Nikolas Coupland, 161–177. Oslo: Novus.
- Bellmann, Günter. 1997. "Between base dialect and standard language." In *Folia Linguistica* 32(1–2): 23–34.
- Berruto, Gaetano. 1989. "Tra italiano e dialetto." In *La Dialettologia Italiana Oggi: Studi Offerti a Manlio Cortelazzo*, edited by Günter Holtus, Michael Metzeltin, and Max Pfister, 107–122. Tübingen: Narr.
- Berruto, Gaetano. 1995. "Dialect/standard convergence, mixing, and models of language contact: The case of Italy." In *Dialect Change: Convergence and Divergence in European*

- Languages*, edited by Peter Auer, Frans Hinskens, and Paul Kerswill, 81–95. Cambridge: Cambridge University Press.
- Britain, David. 2009. "One foot in the grave? Dialect death, dialect contact, and dialect birth in England." *International Journal of the Sociology of Language*, 196/197: 121–155.
- Bybee, Joan. 2010. *Language, Usage, and Cognition*. Cambridge: Cambridge University Press.
- Cerruti, Massimo, Claudia Crocco and Stefania Marzo (eds.). 2016. *Towards a New Standard. Theoretical and Empirical Studies on the Restandardization of Italian*. Berlin, Boston: de Gruyter.
- Chambers, Jack, and Peter Trudgill. 1980. *Dialectology*. Cambridge: Cambridge University Press.
- Cheshire, Jenny, Paul Kerswill, Sue Fox, and Eivind Torgersen. 2011. "Contact, the feature pool and the speech community: The emergence of Multicultural London English." *Journal of Sociolinguistics*, 15(2): 1–46.
- Cornips, Leonie. 2013. "Recent developments in the Limburg area." In *Language and Space: Dutch*, edited by Frans Hinskens, and Johan Taeldeman, 378–399. Berlin: de Gruyter.
- Coseriu, Eugenio. 1980. "'Historische Sprache' und 'Dialekt'". In *Dialekt und Dialektologie*, edited by Joachim Göschel, Pavle Ivić, and Kurt Kehr, 106–116. Wiesbaden: Steiner.
- Deutscher Sprachatlas auf Grund des Sprachatlas des deutschen Reichs (DSA). 1927–1956. By Georg Wenker, started by Ferdinand Wrede, continued by Walther Mitzka, and Bernhard Martin. Marburg: Elwert. <http://www.regionalsprache.de/>, accessed 21st October 2015.
- Ferguson, Charles. 1959. "Diglossia." *Word*, 15: 325–340.
- Gilles, Peter. 1999. *Dialektausgleich im Lützbuergeschen: Zur phonetisch-phonologischen Fokussierung einer Nationalsprache*. Tübingen: Niemeyer.
- Gregersen, Frans. 2009. "The data and design of the LANCHART study." *Acta Linguistica Hafniensis*, 41: 3–29.
- Haas, Walter. 2010. "A study on areal diffusion." In *Language and Space, Vol. 1: Theories and Methods*, edited by Peter Auer, and Jürgen Schmidt, 649–667. Berlin: de Gruyter.
- Herrgen, Joachim. 1986. *Koronalisierung und Hyperkorrektion*. Stuttgart: Steiner.
- Hinskens, Frans, Peter Auer, and Paul Kerswill. 2005. "The study of dialect convergence and divergence: Conceptual and methodological considerations." In *Dialect Change: The Convergence and Divergence of Dialects in Contemporary Europe*, edited by Peter Auer, Frans Hinskens, and Paul Kerswill, 1–50. Cambridge: Cambridge University Press.
- Hinskens, Frans. 1996. *Dialect Levelling in Limburg: Structural and Sociolinguistic Aspects*. Tübingen: Niemeyer.
- Jensen, Torben J., and Marie Maegaard. 2012. "Past participles of strong verbs in Jutland Danish: A real-time study of regionalization and standardization." *Nordic Journal of Linguistics*, 35(2): 169–195.
- Kehrein, Roland. 2012. *Regionalsprachliche Spektren im Raum: Zur linguistischen Struktur der Vertikale*. Stuttgart: Steiner.
- Kern, Friederike, and Margret Selting, eds. 2011. *Ethnic Styles of Speaking in European Metropolitan Areas*. Amsterdam: Benjamins.
- Kloeke, Gesinus. 1927. *De Hollandsche Expansie in de Zestiende en Zeventiende Eeuw en haar Weerspiegeling in de Hedendaagsche Nederlandsche Dialecten*. The Hague: Nijhoff.
- Kristiansen, Tore, and Nikolas Coupland, eds. 2011. *Standard Languages and Language Standards in a Changing Europe*. Oslo: Novus.
- Leinonen, Therese. 2010. *An Acoustic Analysis of Vowel Pronunciation in Swedish Dialects*. PhD thesis, University of Groningen. <http://dissertations.ub.rug.nl/faculties/arts/2010/t.n.leinonen/> (accessed October 21, 2015).
- Mattheier, Klaus. 1997. "Über Destandardisierung, Umstandardisierung und Standardisierung in modernen europäischen Standardsprachen." In *Standardisierung und Destandardisierung Europäischer Nationalsprachen*, edited by Klaus Mattheier, and Edgar Radtke, 1–9. Frankfurt: Peter Lang.
- Mihm, Arend. 2000. "Die Rolle der Umgangssprachen seit der Mitte des 20. Jahrhunderts." In *Sprachgeschichte - Ein Handbuch zur Geschichte der Deutschen Sprache und ihrer Erforschung*, 2nd ed., edited by Werner Besch, Anne Betten, Oskar Reichmann, and Stefan Sonderegger, 2107–2137. Berlin: de Gruyter.
- Lenz, Alexandra. 2003. *Struktur und Dynamik des Substandards: Eine Studie zum Westmitteldeutschen (Wittlich/Eifel)*. Stuttgart: Steiner.
- Lenz, Alexandra. 2010. "Emergence of varieties through restructuring and reevaluation." In *Language and Space, Vol. 1: Theories and Methods*, edited by Peter Auer, and Jürgen Schmidt, 295–314. Berlin: de Gruyter.
- Milroy, James, Lesley Milroy, Sue Hartley, and David Walshaw. 1994. "Glottal stops and Tyneside glottalization: Competing patterns of variation and change in British English." *Language Variation and Change*, 6: 327–358.

- Monka, Malene. 2013. "Sted og sprogforandring – En undersøgelse af sprogforandring i virkelig tid hos mobile og bofaste informanter fra Odder, Vinderup og Tinglev." *Danske Talesprog*, 13: 1–336.
- Quist, Pia, and Bente Svendsen, eds. 2010. *Multilingual Urban Scandinavia: New Linguistic Practices*. Clevedon: Multilingual Matters.
- Rasmussen, Ove, Viggo Sørensen, Torben Arboe, Inger Schoonderbeek Hansen, and Nina Grøftehauge. 2000. *Jysk Ordbog*. Aarhus: University of Aarhus. www.jyskordbog.dk (accessed October 21, 2015).
- Schmidt, Jürgen. 2009. "Die modernen Regionalsprachen als Varietätenverbund." In *Variatio Delectat: Empirisch Evidenzen und Theoretische Passungen Sprachlicher Variation*, edited by Peter Gilles, Joachim Scharloth, and Evelyn Ziegler, 125–144. Frankfurt: Peter Lang.
- Spiekermann, Helmut. 2008. *Sprache in Baden-Württemberg: Merkmale des Regionalen Standards*. Tübingen: Niemeyer.
- Schwarz, Christian. 2015. *Phonologischer Dialektwandel in den Alemannischen Basisdialekten Südwesdeutschlands im 20. Jahrhundert: Eine Empirische Untersuchung zum Vokalismus*. Stuttgart: Steiner.
- Stöckle, Philipp. 2014. *Subjektive Dialekträume im Alemannischen Dreiländereck*. Hildesheim: Olms.
- Südwestdeutscher Sprachatlas (SSA)* (1989–2011), edited by Volker Schupp, and Hugo Steger. Marburg: Elwert.
- Svahn, Margareta, and Jenny Nilsson. 2014. *Dialektutjämning i Västsverige*. Gothenburg: Institutet för Språk och Folkminnen.
- Taeldeman, Johan. 2005. "The influence of urban centres on the spatial diffusion of dialect phenomena." In *Dialect Change. The Convergence and Divergence of Dialects in Contemporary Europe*, edited by Peter Auer, Frans Hinskens, and Paul Kerswill, 263–287. Cambridge: Cambridge University Press.
- Van Coetsem, Frans. 1988. *Loan Phonology and the Two Transfer Types in Language Contact*. Dordrecht: Foris.
- Vanderkerckhove, Reinhild. 2010. "Urban and rural language." In *Language and Space, Vol. 1: Theories and Methods*, edited by Peter Auer, and Jürgen Schmidt, 315–331. Berlin: de Gruyter.
- Werlen, Benno. 2007. *Globalisierung, Region und Regionalisierung*, 2nd ed. Stuttgart: Steiner.

10 Perceptual Dialectology

DENNIS R. PRESTON

10.1 Introduction

Perceptual dialectology (PD) is the branch of folk linguistics that deals with regional distribution of linguistic features from the point of view of nonspecialists (the “folk”), but has, almost from the beginnings, attended to both social and attitudinal factors.

Interest in PD dates back to at least the nineteenth century (Willem 1886) but was extensively developed in the mid-twentieth, especially in the Netherlands and Japan (e.g., Daan 1969; Grootaers 1959; Mase 1964a,b; Sibata 1959; Weijnen 1946). A late-twentieth century revival has established it as a research technique often accompanying general studies of variation, or carried out independently for its own ethnographic value.

In this chapter the goals, methods, and findings of PD are summarized and evaluated, focusing on the following questions:

1. In what places, geographically speaking, do people believe speech differs?
2. Do PD boundaries differ from those offered by professionals?
3. What linguistic cues do people use to identify varieties?
4. In what ways do people believe speech differs?
5. Which variant linguistic facts influence comprehension?
6. What attitudinal factors trigger, accompany, and influence any of the above?

10.2 PD Boundaries: The Netherlands and Japan

The first folk maps of language difference were probably those of Willem 1886; see Goeman 1989 [1999]), who devised the “little arrow” method, an extension of which has come to be known as the “degree-of-difference” method (Preston 1999b, xxxiv). In such approaches, respondents are asked where people speak similarly and/or differently. In the first uses of the method, an arrow was drawn from the respondent’s site to each surrounding site identified as “the same.” Figure 10.1 shows a map of part of the North Brabant (a southern province of the Netherlands) with thick dark lines indicating the professionally determined dialect regions of the province.

Areas in the upper left of the figure illustrate the method. The respondent from “W” (Willemstad) rates no nearby community as similar, and, therefore, no arrow is drawn from W, and no surrounding communities identify W as similar, so no arrows are drawn toward it.

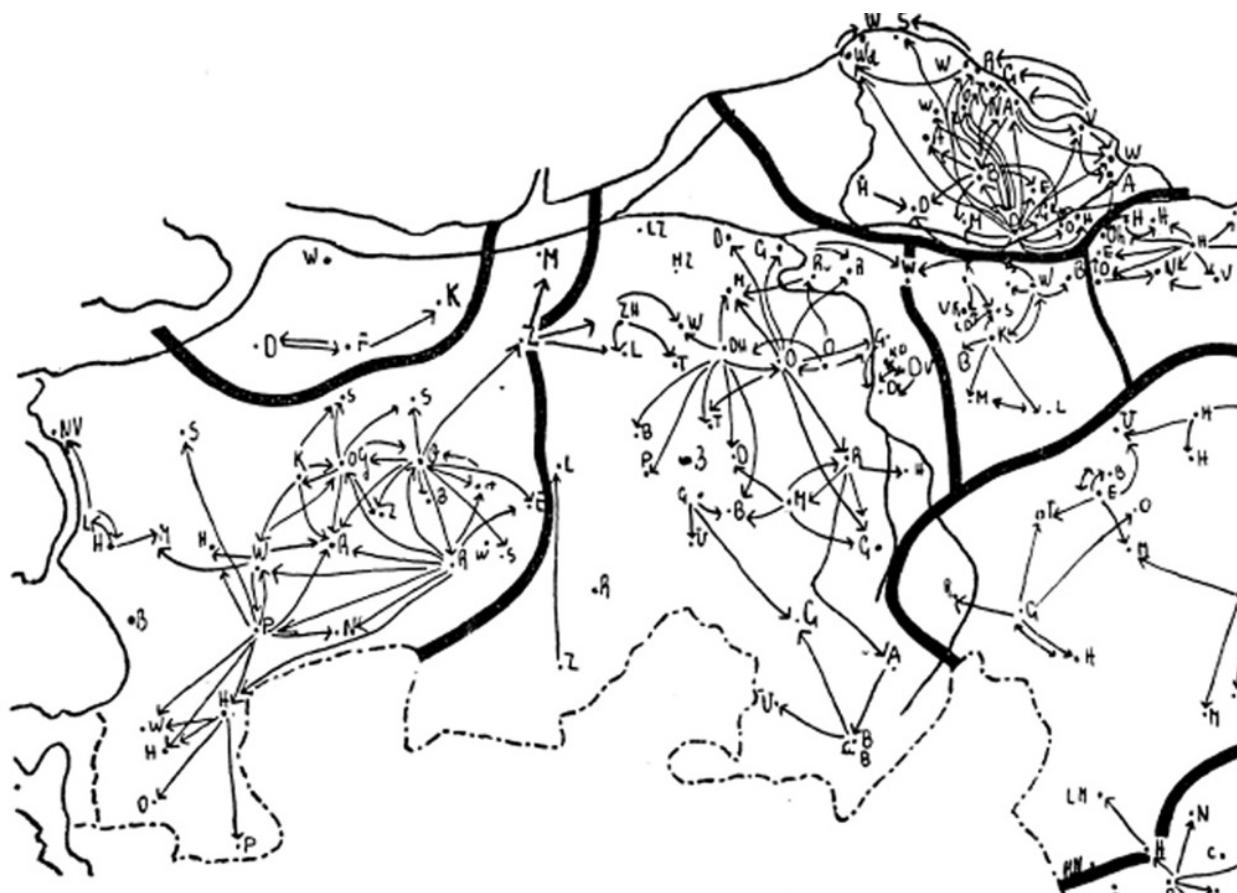


Figure 10.1 The westernmost section of the North Brabant, showing dialectologists' boundaries (thick lines) and the "little arrows" of respondent similarity perceptions (enlarged from Weijnen 1946).

The respondent from "D" (Dinteloord), however, believes that "F" (Fijnaart) is the same, and the respondent from F returns the favor, so arrows are drawn from D to F and F to D. The F respondent also identifies "K" (Klundert) as the same, but the perception is not reciprocal.

If there were a perfect match between perception and production, each pair in the set {W, D, F, K} would be connected with two arrows. That is not the case, but there is a good match, for, although not all the sites are connected to one another, none of them identifies as similar a site outside the production boundary, nor is any identified as similar by a respondent from outside the boundary; moreover, in Figure 10.1 in general, the interconnected bundles of arrows seldom cross the professional boundaries. These findings in Dutch-speaking areas have been incorporated into more general maps of both perceptual and production data. Goeman (1989 [1999, 139]), for example, believes that Van Genniken's map of Dutch dialects (1913 [1928]) used some of Willems' data, and Daan in a general map (1969 [1999]) incorporated the same little arrow data that Weijnen determined and Rensink (1955) used, although the latter of these studies was exclusively based on perception. Other maps using the little arrow method include Kremer (1984 [1999], on the German-Netherlands border), Pearce (2009, in northeast England), and Twilfer (2010, in Westphalia).

In the late 1950s a Dutch-Japanese controversy arose. In western Japan (Sibata 1959), respondents indicated which nearby villages were (1) not different, (2) a little different, (3) quite different, or (4) mostly incomprehensible, but (1) and (2), the bases for the Dutch perceptual studies, were found to be of little value. Grootaers (1959) called them "superfluous" (p. 356), and the results of question (1) were ignored; questions (2) and (3) were combined into one map (as in Figure 10.2), and question (4) was treated separately. The Dutch maps were, therefore, ones of similarity, and the Japanese ones of difference.

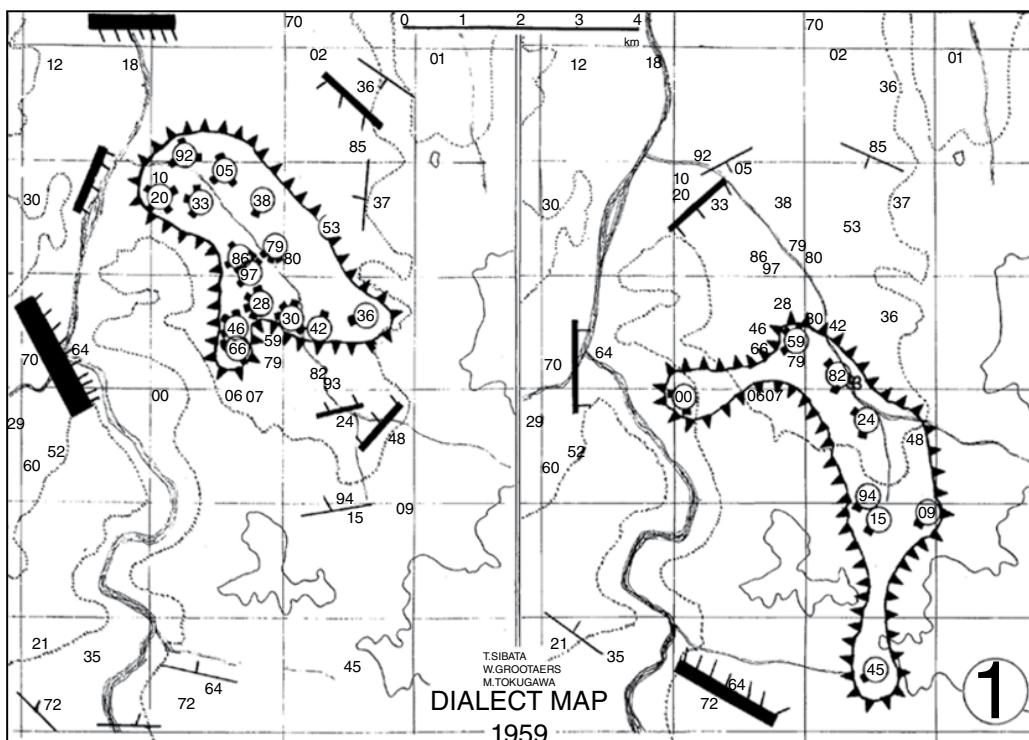


Figure 10.2 The determination of two perceptual dialectology areas in Itoigawa, Japan (Sibata 1959 [1999, 42]).

In Figure 10.2, the Japanese researchers indicate by increasingly thick lines (with small bristle-like ones pointing back to the site of the response) those areas that formed the difference boundaries. When respondents performed similarly in stating where the differences were, they were grouped into subjective speech communities, outlined by the saw-toothed lines. Both Sibata (1959) and Grootaers (1959, 1964) state that these boundaries were of little interest, since they did not correspond to professionally determined ones. The sites inside the two saw-toothed outlines (one on the left, a second of the same area on the right) are different perceptually due to agreement about which surrounding areas sound different. Weijnen (1968 [1999]) suggested that the failure to discover parallels to production boundaries was the result of the Japanese reliance on differences, which he claimed always existed to some degree.

Mase (1964a,b) asked respondents to indicate surrounding areas that sounded the same or different, but used both categorizations in devising his maps of Alpine Japan (mountainous regions of prefectures in central Honshū), finding a good match between production and perception. His were the first to include a mathematical calculation. Figure 10.3 shows his technique for areas #11 through #26. He counted a point for each site at which any respondent mentioned a “little difference.” He counted a half-point if the respondent modified that a degree downward (e.g., a “very slight difference”). He then calculated the number of points for all respondents in the region. If they equaled two-thirds or more of the respondents, he considered the boundary major; if they equaled more than one-third

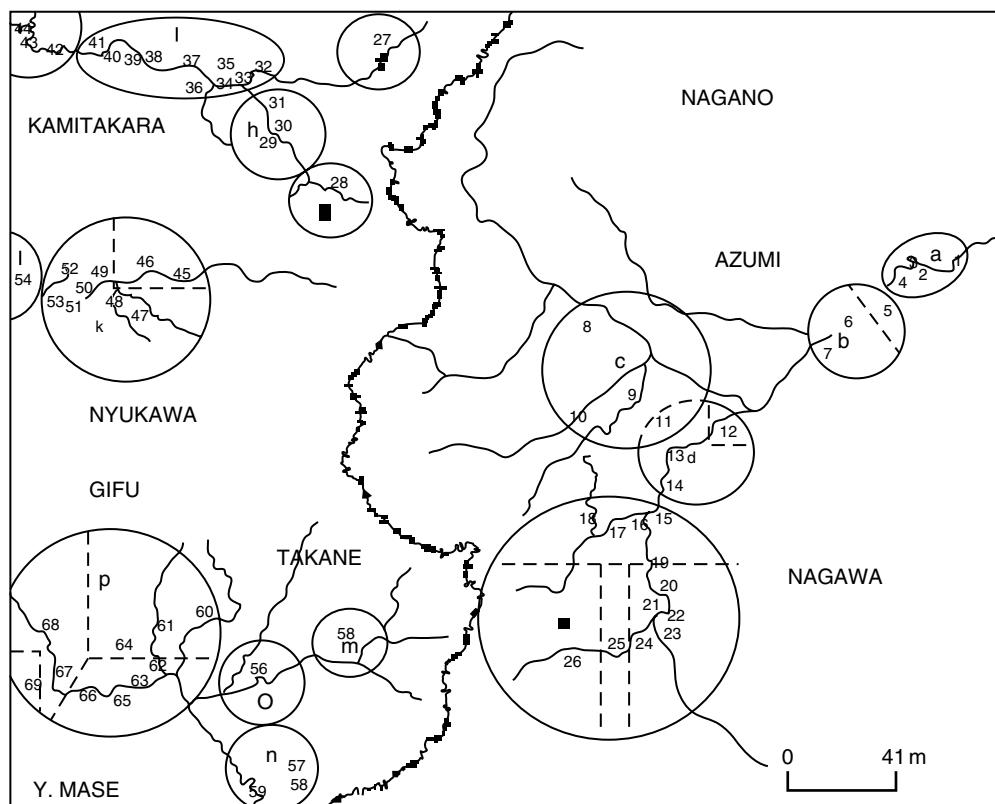


Figure 10.3 Mase's perceptual dialect areas for a section of Alpine Japan (Mase 1964a [1999, 80]).

(but less than two-thirds), he considered it minor. In Figure 10.3, 11.5 points were calculated for the boundary between #14 and #15. Since 11.5 is greater than two-thirds of 16 (the total number of sites, i.e., #11 through #26), #11 through #14 are grouped into one major perceptual region, labeled (d), and #15 through #26 are grouped into a second, (e). Within those regions, however, seven points were given between #24 and #25, six between #25 and #26, and 5.5 between both #12 and #13 and #19 and #20. Figure 10.3 shows these minor divisions using dashed lines, since their point totals equal more than one-third but less than two-thirds of all judgments.

Although Mase's calculation reduces the ability to distinguish regions identified on the basis of similarity and those on the basis of difference (and their relative importance), his treatment is more quantitatively sophisticated than those of his predecessors. The Itoigawa team drew thicker lines to indicate areas that were agreed on as different by a larger number of respondents, but a numeric standard was apparently not used. There is also no quantitative approach in the little arrow technique, since only one connection causes a site to be included in a perceptual area (but see Pearce 2009 for a quantitative use of the little arrow method).

It is odd in these traditions to find value assigned to PD only if the folk agree with professionals. Misao Tōjō, a leading figure in modern Japanese dialectology, said that work on regional speech should go forward *only* after local folk ideas about language were determined (1953, 11), and Daan (1969, 27–29) suggested that cultural practices (e.g., religious ones) could not only cause the perception of differences but also trigger actual differences, implying the importance of PD to the actuation problem as well as the problems of (social) embedding and evaluation identified by Weinreich *et al.* (1968).

10.3 Degree-of-Difference

In a newer PD task known as "degree-of-difference," the scale was considerably expanded over the local area approach of the Dutch and Japanese researchers. Preston (1993, 1996), for example, asked respondents to rank US states as 1 = same, 2 = a little different, 3 = different, and 4 = unintelligibly different. Figure 10.4 shows the responses of southeastern Michigan respondents to this task, in which the mean scores were divided into four groups: 1.00–1.75, 1.76–2.50, 2.51–3.25, and 3.26–4.00. Note that Figure 10.4 shows that when Michigan raters evaluate degree of difference they perceive a large local area of similarity (contrary to Weijnen's prediction). The ratings of the South are also of interest; a large South emerges as a "3" (the same rating given to the Northeast). Texas, Arkansas, Oklahoma, and Missouri are rated along with obviously Southern states (e.g., Georgia and South Carolina). But a "core" South (Alabama, Mississippi, and Louisiana) earns a "4." These ratings suggest that the Michigan raters are aware of a wide area of influence of Southern speech, emanating from an unintelligibly different core. In the similarly rated Northeast, however, there is no such "unintelligible" core.

In later degree-of-difference work, statistical procedures such as factor analyses and multidimensional scaling produced alternative visual representations. Figure 10.5 shows the results for Madrid respondents (with the same 1 to 4 assessment values) for 17 regions of Spain. The two dimensions scaled here offer an opportunity for further interpretation beyond the similarities and dissimilarities discovered in rankings. The authors interpret Dimension #1 (the horizontal) as a multilingual one, in which "non-Spanish" areas—1 (Galicia), 4 (Basque Country), 7 (Catalonia), 13 (Valencia), and 14 (Balearic Islands)—form a cluster on the right. Dimension #2 (the vertical) appears to be one of dialect distinctiveness; one set of the most distinctive dialects is at the top (5 [Navarra], 10 [Extremadura], 16 [Murcia], 17 [Canary Islands]), another at the bottom (9 [Rioja] and 15 [Andalusia]), although these latter two are widely separated on the first dimension, suggesting there is something more native-like about Rioja. The norms are the local area (11 [Madrid], closely linked to 12 [Castille-La Mancha]),

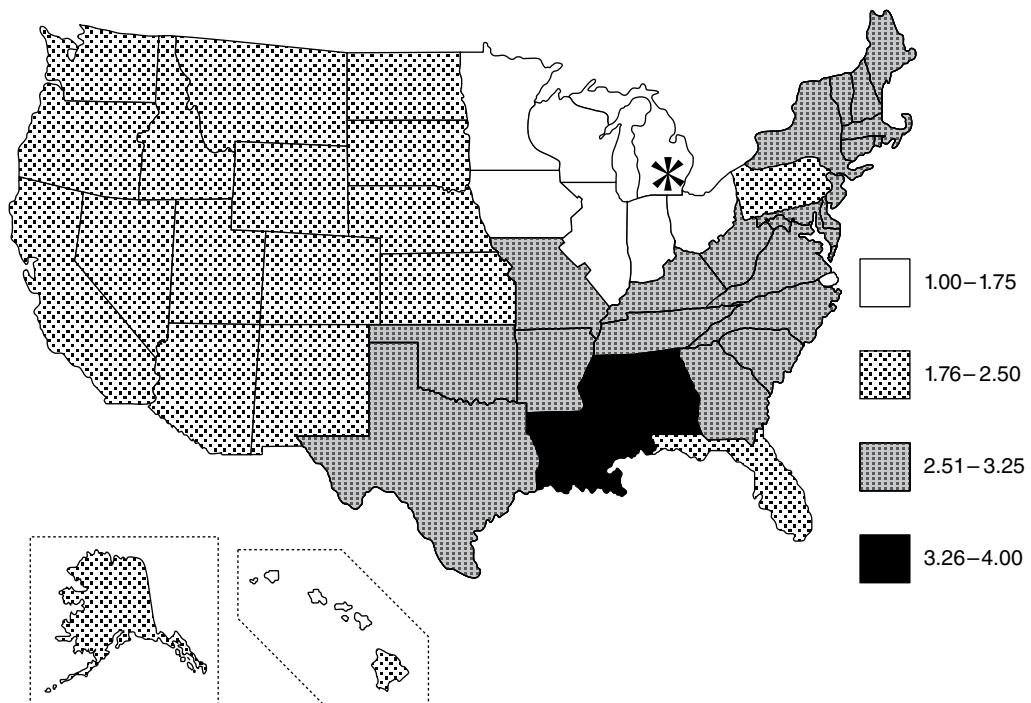


Figure 10.4 Southeastern Michigan (marked by an asterisk) respondents' rating of degree-of-difference for the 50 US states (Preston 1996, 318).

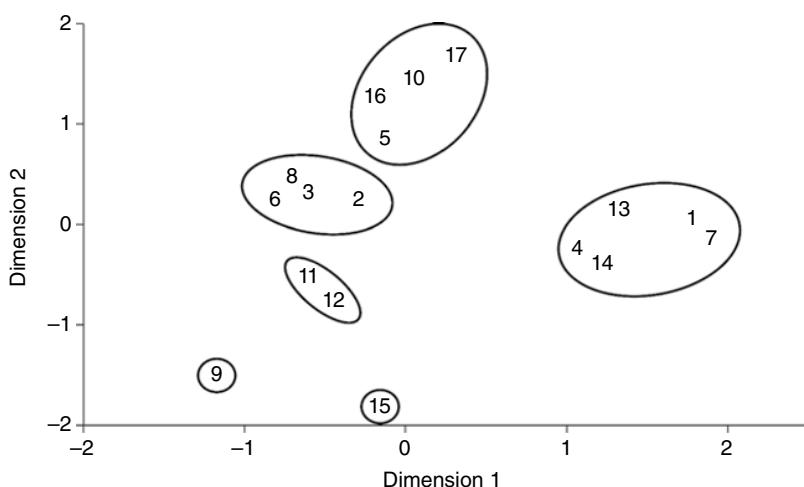


Figure 10.5 A multidimensional scaling of Madrid respondents' evaluations of degree-of-difference for 17 areas of Spain (1 = Galicia, 2 = Asturias, 3 = Cantabria, 4 = Basque Country, 5 = Navarra, 6 = Aragon, 7 = Catalonia, 8 = Castile-Leon, 9 = Rioja, 10 = Extremadura, 11 = Madrid, 12 = Castile-La Mancha, 13 = Valencia, 14 = Balearic Islands, 15 = Andalusia, 16 = Murcia, 17 = Canary Islands) (Moreno and Moreno 2002, 304).

both not far from another group (2, 3, 6, 8) which, since it is above 11 and 12 on Dimension 2, we must assume is slightly more marked dialectally, perhaps, in the direction of the topmost group (Moreno and Moreno 2002, 303).

Such statistical treatment offers other opportunities that help realize the sociolinguistic dimensions of PD. In this study, for example, the authors go on to compare men and women, three age groups, and three educational levels. They note that Dimension #1 (language) is more important in the classifications offered by male, middle-aged, and university-educated respondents, while Dimension #2 (dialect) is more significant for women and youth. In some studies (e.g., Hartley 1999), the groups within multidimensional scales were combined on the basis of such further statistical tests as K-means clustering.

Another technique for uncovering the distinctiveness of regional varieties was borrowed from cultural anthropology (Tamasi 2003), again focusing on US states. She provided respondents with 50 state-named cards, and asked them to sort them into piles of dialect similarity. The piles were subjected to hierarchical cluster analyses, revealing the states most frequently grouped together. Tamasi then derives maps from them, showing degrees of similarity for the clusters at 25%, 50%, and 70% levels. She also considers the match between these groups and traditional dialect boundaries, but since her work, like the work in the degree-of-difference task, used predetermined nonlinguistic areas (states), the comparisons are not easy to draw. On the other hand, the clear advantage to Tamasi's method is that it allows an overall comparison of differences, not one based on the respondent's reckoning of difference from the home site. In Figure 10.4, for example, northeastern states are given the same degree of difference as those in the South, but they are distinct from one another in the cluster analysis derived from Tamasi's pile-sorts shown in Figure 10.6.

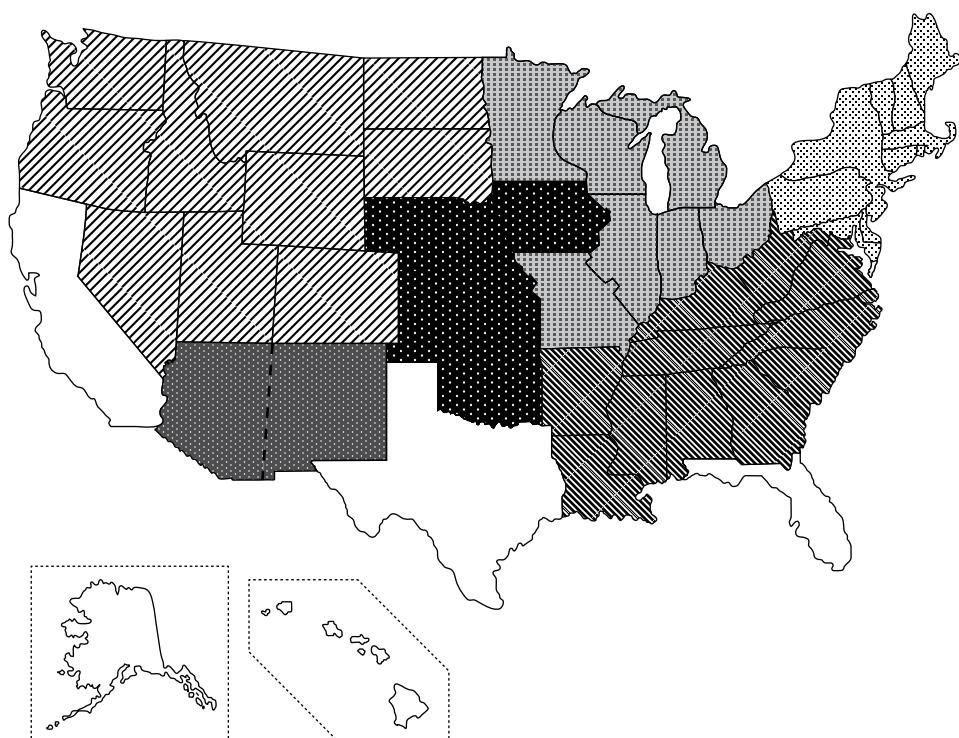


Figure 10.6 Cluster analysis (at .25) of Georgia respondents' completion of a similarity pile-sort task (derived from Tamasi 2003, 66).

10.4 Listen for Differences

In the above techniques, the respondents are given no voice samples on which to base their judgments; more recent work has used such samples. In Preston (1996) a scrambled but relatively evenly-spaced north-south continuum of nine middle-aged, college-educated male voices was played to respondents from southeastern Michigan who were asked to associate each with a site on the map shown in Figure 10.7. The samples contained no lexical or grammatical features that were regionally diagnostic.

A cluster analysis (Figure 10.8) might suggest considerable success. The northernmost voices (Coldwater and Saginaw) are linked first (i.e., joined together furthest to the left). They are the only two areas dialectologists would label “Inland North” (Labov *et al.* 2006). This pair is then linked to South Bend, the next voice south, perhaps the only voice in the professionally-determined “North Midland”; this group of three is then linked to Muncie, the next voice to the south and solidly “Midland,” but then these northern and midland four are linked to New Albany. In a professional dialect geography, New Albany should first be linked to sites south of it (Bowling Green and Nashville), all “South Midland” areas.

There is also a southern grouping, but the distance of its linkages from the left shows that it is not as strong as the northern one. Nashville and Florence are first linked, then tied to Bowling Green, although, as suggested above, dialectologists would probably have first linked New Albany, Bowling Green, and Nashville, and then those three to Florence. The most striking fact for professionals, however, is that Dothan, the southernmost voice, is not linked to the southern cluster of Bowling Green-Nashville-Florence. That cluster is linked first to the large northern group before all are finally linked to Dothan. Perhaps Dothan is phonetically so southern (it is the only /r/-less voice, although variably) that all other southern varieties are linked to everything north of them before Dothan is included. Professional dialectologists could identify many Southern features (e.g., /aɪ/ monophthongization, /ɪ/ ~ /ɛ/ conflation before nasals) in all the voices from New Albany to Dothan, so the perceptual grouping tantalizingly suggests which features are salient and how very distinct the southernmost variety of US English is

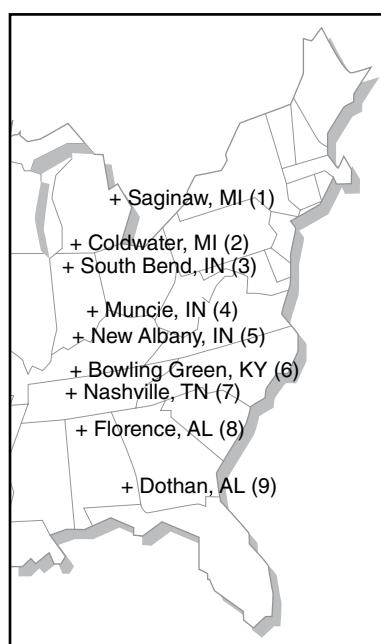


Figure 10.7 The nine home sites of the male voices (Preston 1996, 322).

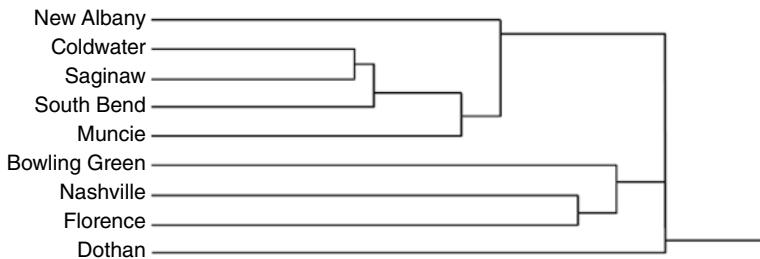


Figure 10.8 Cluster analysis of southern Michigan placement of nine voices on the map in Figure 10.7. (Source: Plichta and Preston 2005, 118.)

among nonlinguists; it also displays in this case a much greater sensitivity among these Michigan respondents to more nearby (Northern and Midland) areas than to more southern (South Midland and Southern) ones.

Degree-of-difference has also been indirectly measured in a voice-stimulus technique called the “starburst” method, introduced in Montgomery (2007). He asked respondents from various sites in the north of England to identify voice samples from around the country by marking on a map where they thought the voice was from. He then showed, in a “starburst” diagram, the relationship of each folk placement to the actual site of the sample voice. This technique does away with the forced-choice linearity used in Preston (1996; e.g., Figure 10.4 above) although it continues the focus on differences from the point of view of a single area.

10.5 Draw-a-Map

Another PD practice was borrowed from cultural geographers’ interests in respondents’ hand-drawn maps (e.g., Gould and White 1974). The technique (called ‘draw-a-map’) was introduced in Preston (1982) and was followed by increasingly sophisticated means of combining individual respondent maps into general ones. Figure 10.9 shows an individual

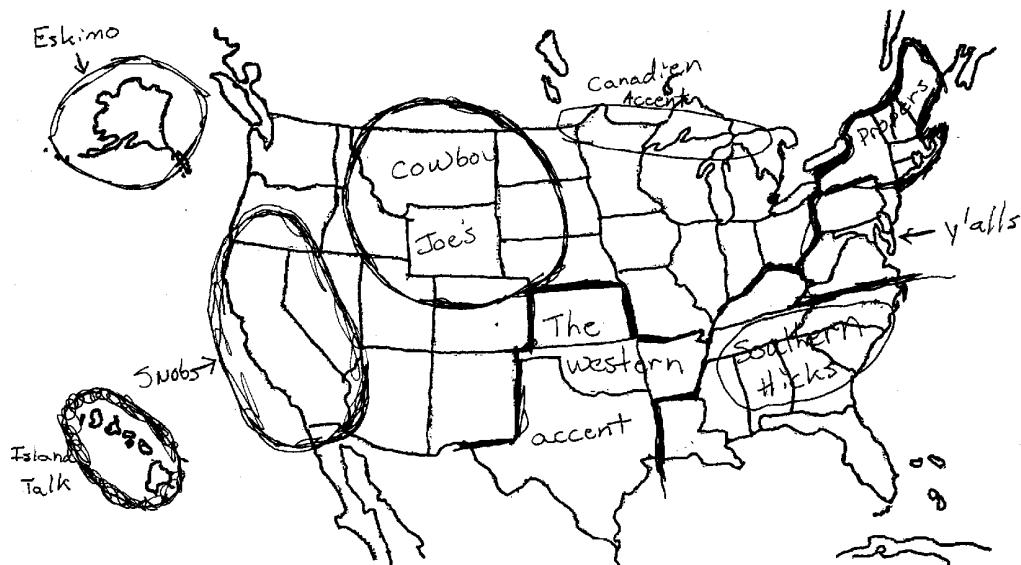


Figure 10.9 A hand-drawn map of US dialect areas by a southeastern Michigan European-American female, aged 18 in 1984.

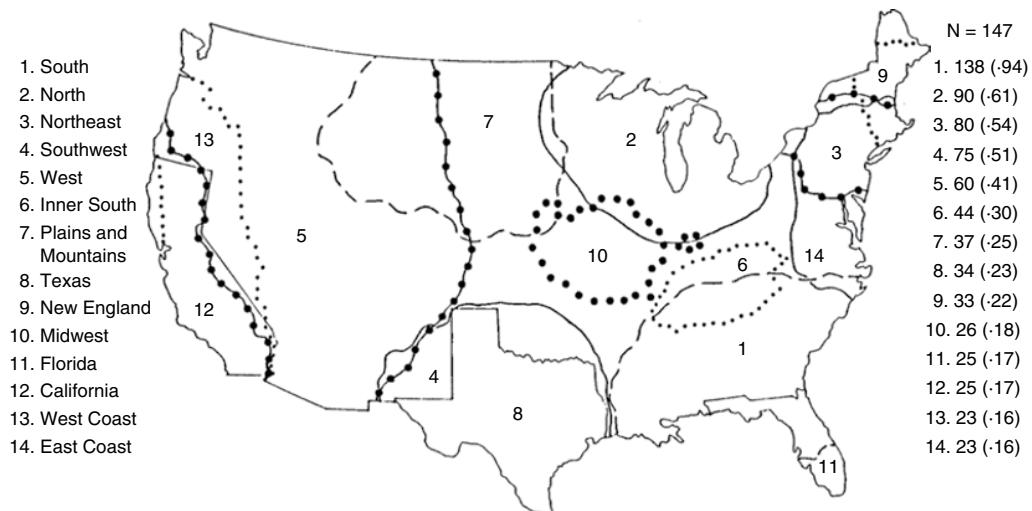


Figure 10.10 Computer-generalized regions from 147 southeastern Michigan hand-drawn maps of US dialect areas (Preston 1996, 305).

map and Figure 10.10 a map generalized from ones drawn by the same southeastern Michigan respondents whose home site is indicated by the asterisk in Figure 10.4.

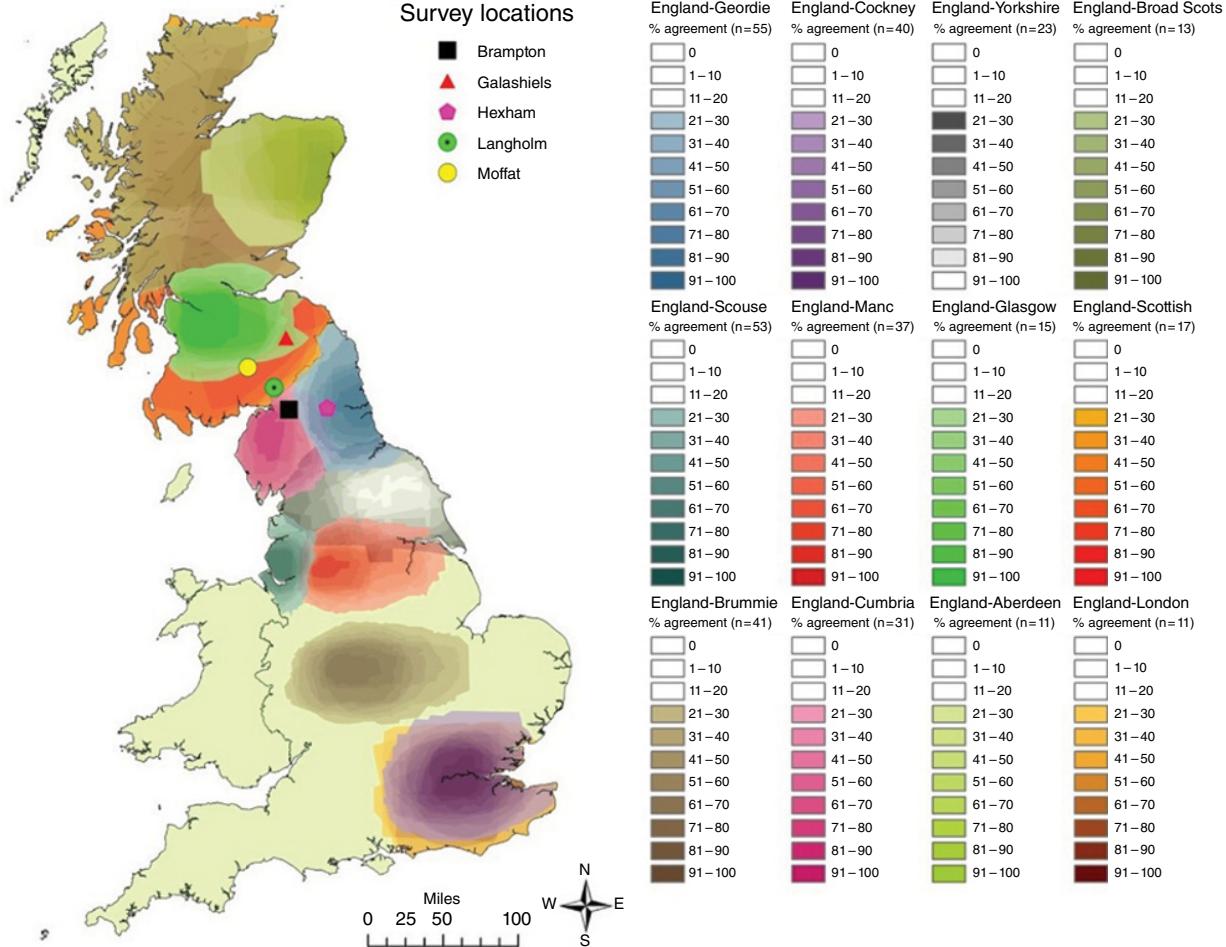
The map in Figure 10.10 was realized by arbitrarily cutting off regions not drawn by 15% or more of the respondents. The remainder were outlined with a light-pen onto a light-sensitive pad, and the aggregated areas were built by asking the computer to identify those pixels that were enclosed by the outlines at various levels of intensity. The map in Figure 10.10 is based on 50% agreement among the respondents.

This computational procedure was improved upon by Long (1990) and can now be realized in a variety of Geographic Information Systems (GIS) mapping software that allows for full-color, quantitatively precise representations of the hand-drawn data as well as maps that contrast social subgroups of respondents. Individual maps, often studied for their ethnographic content, and pre-GIS generalizations have been obtained from many areas, and selections representing the British Isles, Canada, France, French-speaking Switzerland, Germany, Japan, North and South Korea, Quebec, Turkey, the United States, and Wales can be found in Preston (1999a) and Long and Preston (2002), although there are many other examples covering an even wider range of areas.

A how-to for the construction of GIS maps of perceptual areas is available in Montgomery and Stoeckle (2013), and Figure 10.11 shows the perceptual area potential for such maps.

Each of 12 dialect areas is outlined in a “heat map” showing the intensity of respondent agreement for the extent of the area. The procedure also allows comparison of maps drawn by different social groups and for the comparison of perceptual maps with such other facts as population density or, as shown in Figure 10.12, the correlation between the perceptual mapping of an area in southwestern Germany and the Catholic/Protestant areas of the same region.

Because the “attribute tables” of GIS mapping software can contain any information about respondent or area identity, and because GIS software systems contain a wealth of information about areas that can be overlaid on perceptual maps, the potential for more



This work is based on data provided with the support of the ESRC and JISC and uses boundary material which is copyright of the Crown and the ED-Line Consortium.
Location information is ©Crown Copyright/database right 2011. An Ordnance Survey/EDINA supplied service.

Figure 10.11 A generalized perceptual map of English and Scottish dialects from the point of view of two north of England sites: Brampton and Hexham (Montgomery and Stoeckle 2013, Map 25). (See insert for colour representation of the figure.)

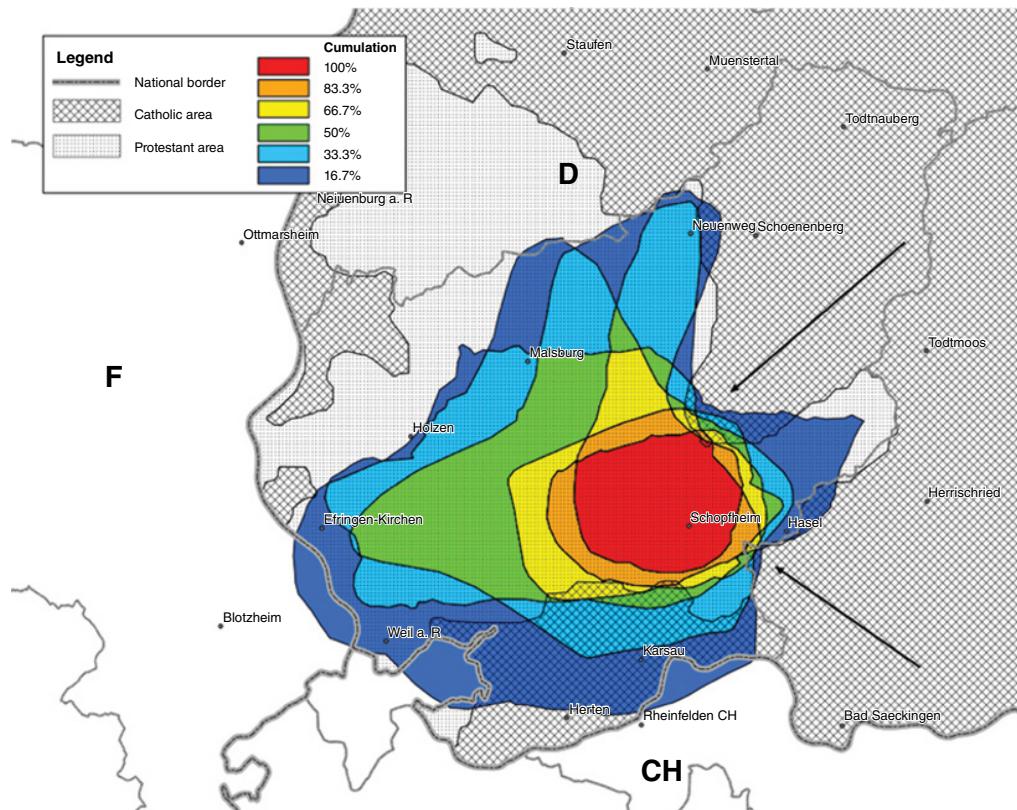


Figure 10.12 A generalized perceptual map of Schopfheim respondent identification of the local dialect area compared to Catholic and Protestant areas in the same region (F = France, D = Germany, CH = Switzerland). Schopfheim is marked with a white arrow (Montgomery and Stoeckle 2013, Map 16). (See insert for colour representation of the figure.)

sophisticated investigation of PD is greatly enhanced. For example, Figures 10.13 and 10.14 compare intensity maps for the areas labeled “twang” and “drawl” by respondents from across Texas.

Although “drawl” and “twang” overlap in northern and northeast Texas, “drawl” is much stronger there and in the entire state than “twang” and very seldom perceived as characteristic of the border area with Mexico. “Twang,” however, is heaviest in the north-eastern part of the state, an area perhaps regarded as more South Midlands in character. Other recent examples of this technique include Evans (2011) in Washington state; Jeon (2012) in Korea; Montgomery (2007, 2012) in northern England, and Stoeckle (2012) in southwestern Germany. Many more studies are in progress.

10.6 PD with an Attitude

The relationship of language attitudes to the perception of region was an early consideration in PD, and the first map of regional attitudes appears to be that of Inoue (1977/8, 1978/9, and see Inoue 1999, 149 [Figure 11.1]), based on the semantic differential and matched-guise techniques used in attitude studies carried out by social psychologists of language. Preston

Texans' Perceptions of Language Variation
Frequency of Respondents Who Labeled an Area *Drawl*

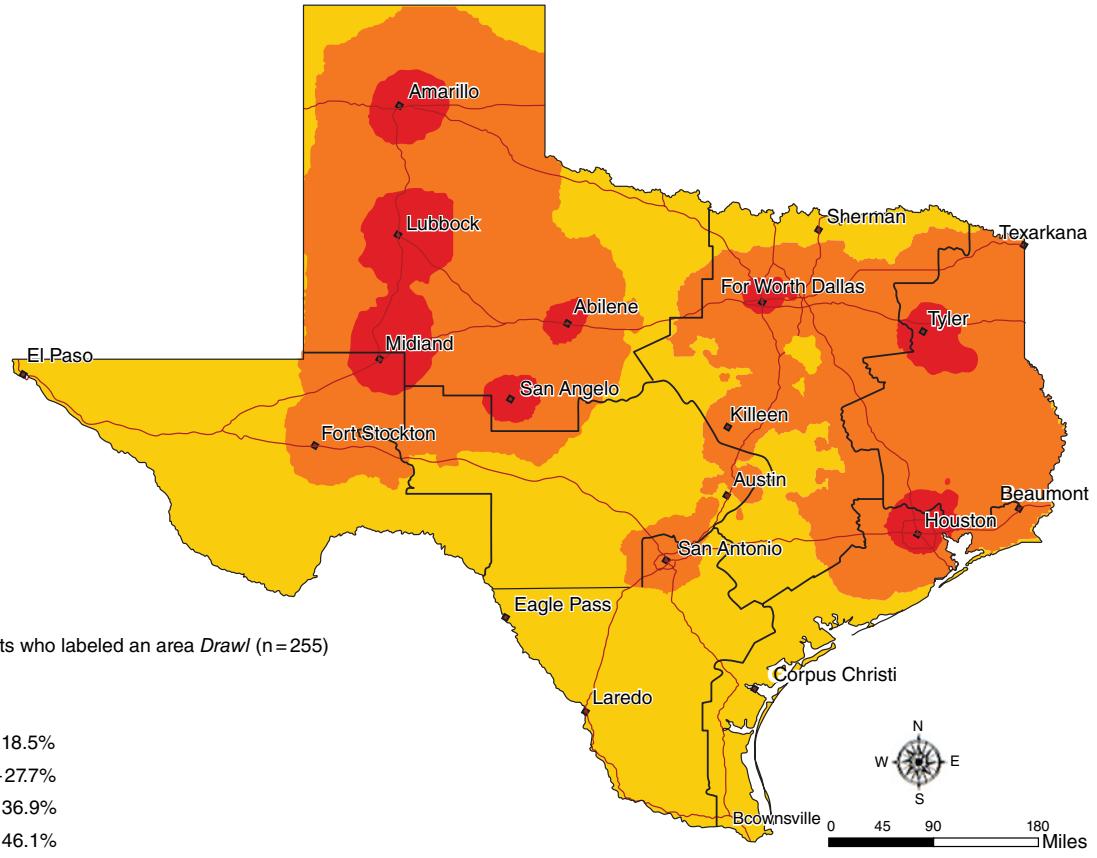


Figure 10.13 Areas of Texas identified as having a "drawl" (Cukor-Avila, forthcoming).

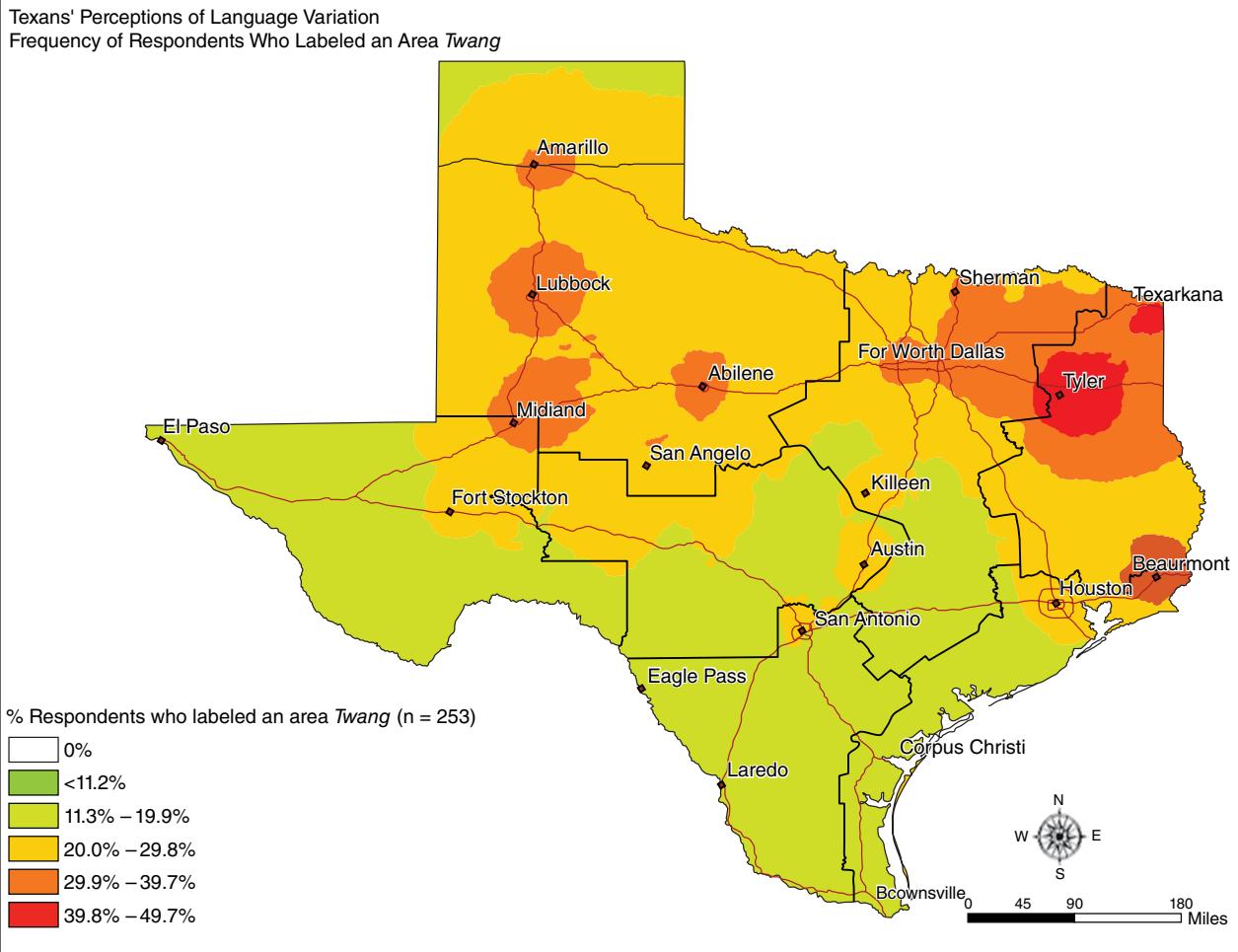


Figure 10.14 Areas of Texas identified as having a “twang” (Cukor-Avila, forthcoming).

(1985), again borrowing from the cultural geographical tradition outlined in Gould and White (1974), established a second method of mapping evaluative judgments in PD, although Preston (1982) comments extensively on the evaluative commentary Hawai'ian respondents wrote on their hand-drawn maps (a technique more extensively made use of in Hartley and Preston, 1999). Such comments were early indicators that hand-drawn maps contained much more than the perception of linguistic differences alone (as many of the respondent labels in Figure 10.9 clearly show).

Social psychological studies of attitude were, in fact, criticized in PD work. Preston (1989: 3) suggested that ratings of voices from various sites were interpreted as responses to voices from those sites, but, in fact, most social psychological studies did not determine whether respondents could identify the home sites of the voices presented for evaluation, and the very few of those that did ask found that many identifications were incorrect (e.g., Tucker and Lambert (1969) or Milroy and McClenaghan (1977)).

A combination of several methods and the incorporation of the results from hand-drawn map studies are illustrated by the following studies. Preston (1996) asked respondents to rate the US states for language "correctness" and "pleasantness," attempting to short-cut the usual factor-analytic approach taken in social psychological studies by instead directly accessing the commonly-discovered constructs of "status" and "solidarity" (e.g., numerous chapters in Ryan and Giles 1982). Figure 10.15 shows the results for "correctness" and Figure 10.16 that for "pleasantness" among the same southeastern Michigan raters indicated by the asterisk in Figure 10.4, above.

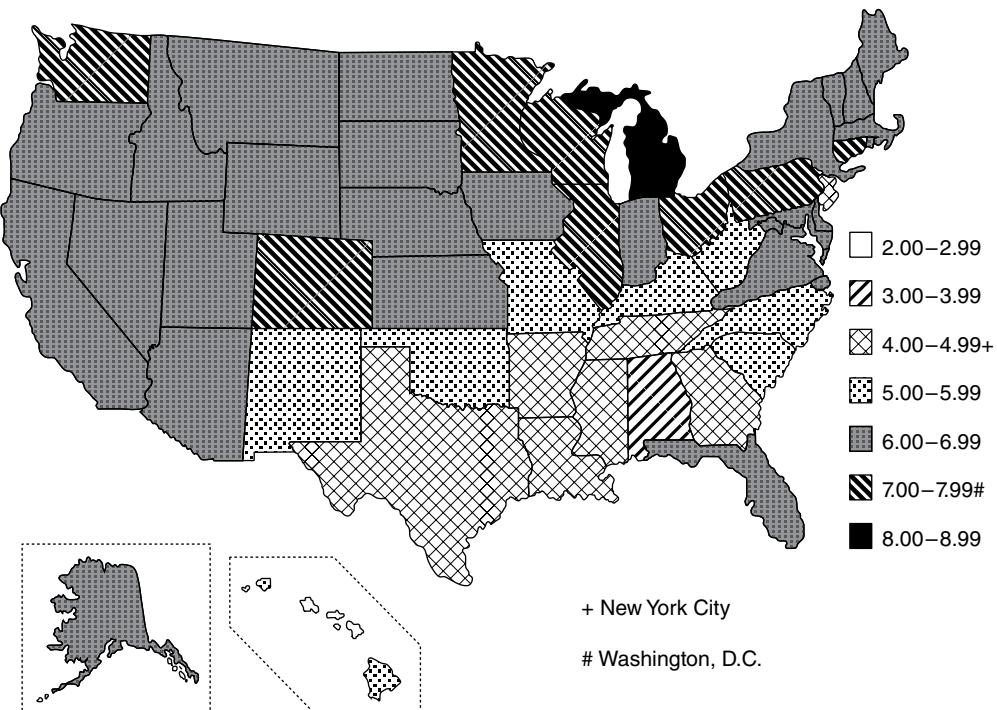


Figure 10.15 Southeastern Michigan ratings of the 50 US states, New York City, and Washington DC on a scale of 1 (least) to 10 (most) for language "correctness" (Preston 1996, 312).

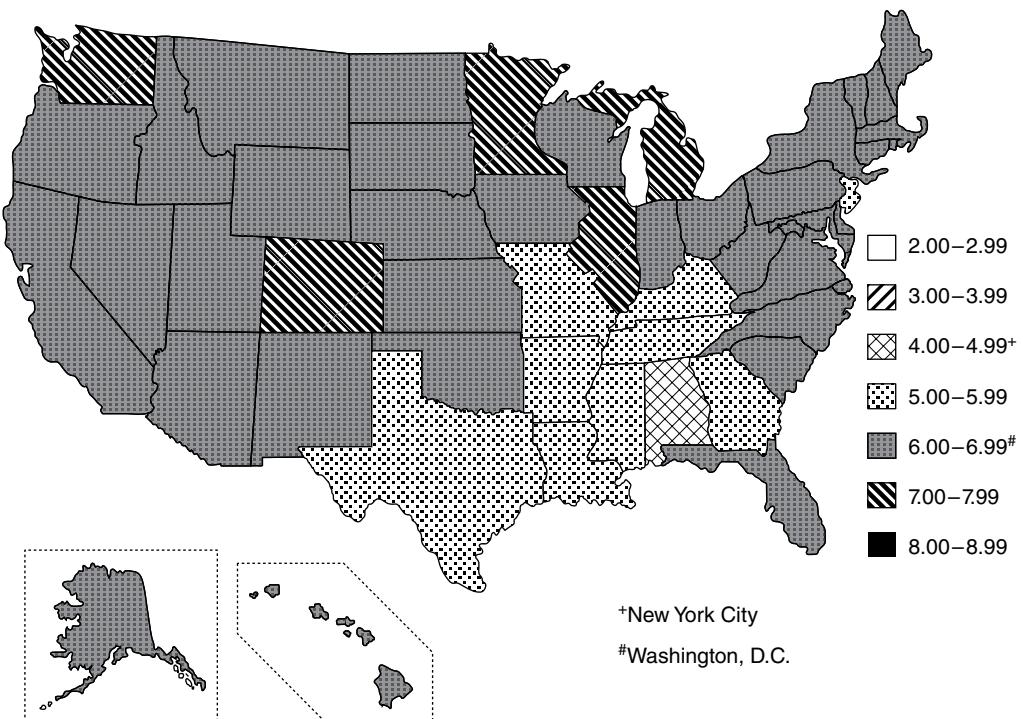


Figure 10.16 Southeastern Michigan ratings of the 50 US states, New York City, and Washington DC on scale of 1 (least) to 10 (most) for language "pleasantness" (Preston 1996, 316).

These Michigan raters think very highly of their own speech for status, and rate Michigan best; they also think of themselves as the most pleasant-sounding, although they share this honor with four other (noncontiguous) states. The entire South and the New York City/New Jersey area fare worst for both "correctness" and "pleasantness." This ranking sheds further light on the quantitative results shown in Figure 10.10. Why would Michigan respondents most frequently draw a US South (94%), then their home area (61%), and in third place, an area focused on New York City/New Jersey (54%)? Although the intensity of these hand-drawn representations might at first seem to confirm the degree-of-difference ratings seen in Figure 10.4, there are more than subtle differences. Why is the New York/New Jersey area represented so much more frequently than most of nearby New England, when both areas have the same degree-of-difference rating (Figure 10.4)? The most correct (and secondarily pleasant) site is Michigan, and the least correct (and least pleasant) areas are the New York City/New Jersey focal region and the entire US South. Are these regions most salient because they are most similar and most different linguistically, or does their linguistic salience emerge at least in part because of nonlinguistic stereotypes held about the people and language of the regions themselves?

These simple ratings of "correct" and "pleasant," although they exposed broad patterns of preference, did not make good use of available methodologies. Preston (1999c) combined techniques associated with the matched-guise technique of language attitude studies and the results of previous hand-drawn map investigations. Southeastern Michigan raters were presented with a simplified version of Figure 10.10, which displayed the major US perceptual regions previously determined by similar respondents, a technique that did away with the arbitrary use of states and two major urban areas. Participants were asked to write down

as many descriptors of the way people talked in these different regions as they could think of, and the following most frequently offered ones were used in the next step of the investigation.

slow – fast	smart – dumb	speaks with – without a drawl
polite – rude	formal – casual	speaks with – without a twang (Preston 1999c, 363)
snobbish – down-to-earth	bad English – good English	
educated – uneducated	friendly – unfriendly	
normal – abnormal	nasal – not nasal	

The map was then shown to another group of respondents from southeastern Michigan who were asked to rate each of the regions shown in Figure 10.10 on six-point Likert scales for the 12 locally provided attributes. Table 10.1 shows the results for areas 1 and 2 of Figure 10.10 (the home area of the respondents and the US South), the areas most frequently drawn by the respondents who carried out the hand-drawn map task.

As shown in Table 10.1, these respondents rate the 12 attributes at 4.00 and higher for their home area ("North"), with the exceptions of only "not nasal" and "casual." Except for "no drawl" and "no twang," the highest ratings are for the "status" dimension characteristics (e.g., "normal," "smart," "good English"). Those status attribute ratings are reversed in their ratings for the South, and are lowest rated. More importantly, however, this more detailed study shows that these northern raters actually find southern speech superior on the solidarity scales of "casual," "friendly," "down-to-earth," and "polite." This reveals a linguistic insecurity that the simple state-ranking studies of "pleasant" and "correct" did not: Michiganders do not just have a less intense feeling about the "pleasantness" of their speech, they actually find their speech lacking in the solidarity function when compared to southern US English in terms of the respondent-elicited and more detailed categories used in this study, as opposed to the researcher-imposed dichotomous notions "pleasant" and "correct."

Table 10.1 Ratings for speech in the North (area 2 in Figure 10.10) and South (area 1 in Figure 10.10) for 12 attributes (* indicates the only two adjacent scores that are significantly different, and ‡ indicates negative ratings; Preston 1999c, 366).

South		North			
Rank	Attribute	Mean	Rank	Attribute	Mean
1	Casual	4.66	1	No drawl	5.11
2	Friendly	4.58	2	No twang	5.07
3	Down-to-earth	4.54	3	Normal	4.94
4	Polite	4.20	4	Smart	4.53
5	Not nasal	4.09	5	Good English	4.41
		*	6	Down-to-earth	4.19
6	Normal [Abnormal]	‡3.22	7	Fast	4.12
7	Smart [Dumb]	‡3.04	8	Educated	4.09
8	No twang [Twang]	‡2.96	9.5	Friendly	4.00
9	Good English [Bad Eng.]	‡2.86	9.5	Polite	4.00
10	Educated [Uneducated]	‡2.72	11	Not nasal	3.94
11	Fast [Slow]	‡2.42	12	Casual	3.53
12	No drawl [Drawl]	‡2.22			

10.7 Talk About Language Variety

Even more complex experimental methods in attitude studies have arisen, but a brief survey of them will be given in the final section. It will not do, however, to leave this more general discussion of attitudes to language variety without mention of discourse. A number of discourse, conversational, speech act, and other pragmatic tools have been used to investigate what people say about language, but the trick has always been to convert the structural-interactional interests of those analytic procedures into ones that will be revealing with regard to the *content* of the discourse rather than its structure.¹ One problem has been that folk interaction on PD matters appears to be limited to a listing of assertions, for example, "People in Kinki speak funny Japanese." Although such assertions may make up a substantial portion of the metalinguistic discussion, topic handling, presuppositions, and other discoursal, pragmatic facts have been shown to be valuable in determining respondents' attitudes to language variety (e.g., Preston 1994).

Since not all of these methods can be demonstrated, this potential for revealing PD in discourse can be illustrated by pointing out the possibility of extracting *pragmatic presuppositions*, those related to lexical and structural triggers (e.g., Levinson 1983, 181–185). For example, "started" in "Bill started smoking" presupposes that there was a time in the past when Bill did not smoke (e.g., Levinson 1983, 182). Although "Bill didn't flunk Algebra" doesn't presuppose that Bill flunked anything, "What Bill didn't flunk was Algebra" suggests that he did (e.g., Levinson 1983, 182–3). When discourses turn to language, the search for such presuppositions may be rewarding.

In the following exchange, a Taiwanese fieldworker (C) discusses African American English with an African American friend (D).

1 C: We uh – linguistics, in this field, uh – from the book I s- I mean, I saw from the book that – many linguists quite interest in black English. So could you tell me - a little bit about – your dialect?

2 D: Dialects.

3 C: Heh yeah

4 All: ((laugh))

[

5 D: Well, uh: – well – see the world's getting smaller. There's =

[]

6 C: ((laughs)) I- I mea- do you have-

7 D: =not – even among all the ethnic groups we're- we're getting- getting less and less of dialectual in- inFLUence. (.hhh) Uh I'm- happen - not to be - from the South,(Preston 1994, 286–287)

Without an account of presuppositions, this discourse is difficult to interpret, particularly 5–7 D. The first clue lies in the presupposition(s) of "So could you tell me a little bit about your dialect?" (1 C). "Your dialect" presupposes the existence of "dialect(s)" and that "you" are the speaker of one. D's perception of these presuppositions leads to the odd assertions in 5–7 D:

The world's getting smaller.

We're getting less and less of dialectual influence (i.e., there are fewer and fewer dialects)
I happen not to be from the South.

"The world's getting smaller" explains why there are fewer dialects (education, media, mobility, etc.), but the assertion that there are fewer dialects responds to C's presupposition that they exist (a *definite description*; e.g., Levinson 1983, 181). More subtly, D confirms C's

presupposition that dialects exist, but, for D, they exist only in such areas as “the South.” D appears to suggest that if C had only been lucky enough to interview a speaker from the South, he might have had his query about “your dialect” answered.

How can D’s observation that he is not from the South be taken unless it related to his response to C’s query about D’s dialect? Recall that Michiganders, D included, find the South very salient as a regional speech area and that its salience is undoubtedly related to its perceived incorrectness (see Figures 10.10 and 10.15); that is, it is “a dialect.”

Presuppositions may also explain why D “happens” not to be from the South. Why does he not just say “I am not from the South”? “Happen” is an *implicative verb* (Levinson 1983, 181) and presupposes “inadvertence,” “lack of planning,” or “by chance.” D “happens” not to be from the South because it is only happenstance that C picked on a respondent who was not from the South (and could therefore not respond to his request for personal “dialect” information).

A great deal more on this conversation and various pragmatic approaches to its content is provided in Preston (1994). Work on discourse, then—from many perspectives, but surely from both formal and informal pragmatic ones—reveals not only what speakers have said or asserted (the conscious) but also what they have associated, entailed, and presupposed (the subconscious). The growing interest in subconscious attitudinal reactions is explored below.

10.8 The Linguistic Content of PD

In more recent approaches to PD (and the attitudes so intimately connected to it) linguistic detail rather than the more global properties of speech used in most social psychological work has surfaced. Respondents are sensitive to specific features in varieties (e.g., Graff, Labov, and Harris 1986; Purnell, Idsardi, and Baugh 1999), and in experimental efforts to test this sensitivity a variety of sophisticated techniques borrowed from the speech sciences and acoustic phonetics have played an important role, since phonological features are the most frequently investigated.

Plichta and Preston (2005) selected a well-known southern US speech stereotype (/ay/ monophthongization) and resynthesized a sample of the word *guide* so that it increased in monophthongization in seven regular steps, from a fully diphthongal form ([aɪ]) to a fully monophthongal one ([a:]). The 14 voice samples (seven each for one male and one female speaker) representing each of the seven steps were played to a group of 96 listeners three times, yielding from each listener a total of forty-two judgments. In each case, the respondent was to assign the word to one of the nine sites shown in Figure 10.7 above. The assigned site numbers were averaged to ascertain if degree of monophthongization was perceived by the respondents, who came from all over the US, as an increasingly Southern feature. Table 10.2 shows the results.

A post-hoc ANOVA test shows that each of these mean scores is significantly different from every other one, revealing considerable sensitivity to very minor phonetic changes and very clearly showing an association between monophthongization and the respondents’ regional perception of it.

Other studies of specific features have focused on local sensitivity to regional norms. Labov (2001) reports on a study in which high school (HS) and college (Col) students who were local Inland Northern speakers from Chicago, Illinois (Chi) and non-locals of the same age groups from Philadelphia, Pennsylvania (Phi) and Birmingham, Alabama (Bir) listened to the word *socks*, the phrase *wear socks*, and the sentence *You had to wear socks, no sandals*.

The Chicagoans are involved in a change in which the vowel of *socks* (i.e., the American English *LOT* vowel) is pronounced farther forward along the F2 dimension (in the direction of TRAP).² As Figure 10.17 shows, the younger (HS) locals outstrip all other groups (even slightly older locals) in understanding the word in isolation and the short phrase, but it also shows that even the young native speakers of this system fall below

Table 10.2 Mean scores based on regional values assigned each step of the increasingly monophthongized versions of /ay/ (Plichta and Preston 2005, 121).

Step	Mean	Region
1	2.85	1. Saginaw
2	3.17	2. Coldwater
3	3.87	3. South Bend
4	4.89	4. Muncie
5	5.99	5. New Albany
6	6.58	6. Bowling Green
7	7.02	7. Nashville
		8. Florence
		9. Dothan

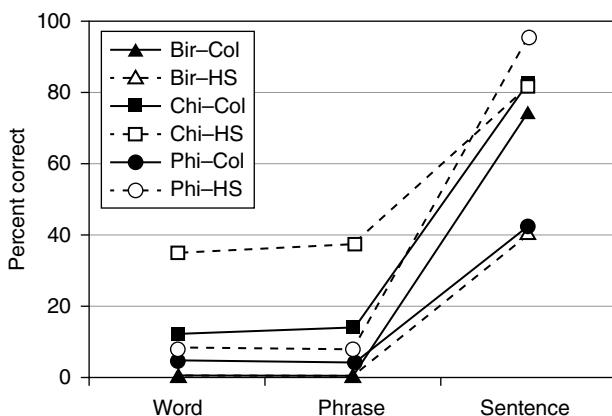


Figure 10.17 Local and non-local respondent groups' correct understandings of the item *socks* as an isolated word, in a phrase, and in a sentence (Labov 2001, 69).

40% correct on the isolated word test, an important fact for a dialectology that involves perception as well as production.³

Herold (1990) records another interesting mismatch between production and perception in a study in Pottsville, Pennsylvania, an area where the THOUGHT and LOT vowels are merging. In Figure 10.18 we see that the production change is mirrored in perception for the girls over an 11-year period, but in the same time period the boys have significantly changed their perception in keeping with the emerging local norm, but not their production.

More complex studies of local versus nonlocal detection and comprehension of individual linguistic items have been carried out. In Rakerd and Plichta (2003), for example, seven-step resynthesized versions of the LOT vowel fronted along the F2 dimension (the items *hot* and *sock*) were played to southeastern Michigan respondents. In some cases, carrier phrases with the same or other vowels from the local system (i.e., fronted LOT, raised and fronted TRAP, and lowered DRESS) preceded the items to be judged; in other cases, carrier phrases with unshifted vowels were used. When the local system carrier phrases appeared (regardless of the specific items they included) the respondents continued to recognize the test item as *hot/sock* in a much more fronted position than when the carrier phrases were not local, under

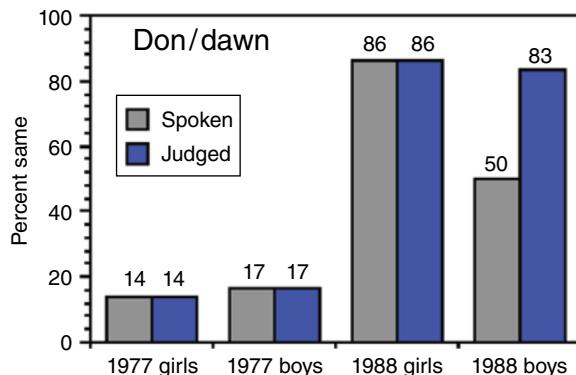


Figure 10.18 Change over time in girls' and boys' production and perception of the merger of the THOUGHT/LOT vowels (Herold 1990).

which condition they changed their interpretation to *hat/sack* at an earlier point in the F2 fronting). It is a long-standing idea in the study of variety perception that hearers adjust their classificatory strategies to the perception of system provided in input (Ladefoged and Broadbent 1957), and that is surely an important consideration in cataloging the general facts about variety, at least if such a catalog includes perception.

10.9 Putting It All Together: PD, Attitude, and the Linguistic Facts

Experimental PD took its most important turn in the work of Niedzielski (1999). She asked 42 southeastern Michigan respondents to listen to a recorded voice, the local identity of which was indicated; they were told to concentrate on the vowel they heard in particular words and to compare that vowel to a set of three resynthesized vowels (from the same speaker's data). They were then asked to choose the one that best matched the original.

The speaker was influenced by the Northern Cities Shift (described above), and the F1 of her /æ/ (TRAP) is at about 700Hz; the F1 norm for female speakers of American English (according to Peterson and Barney 1952, 183) should be considerably higher, at around 860Hz. Niedzielski examined the respondents' classification of the word *last*. The formant frequencies for the three resynthesized tokens that the respondents were given to choose from in the matching task are shown in Table 10.3.

The results of this matching experiment are shown in Table 10.4.

Not one of the respondents chose token #3, the variant that matched the one first produced by the speaker. Instead, they overwhelmingly chose the lower, more central token, #2. A few respondents even chose the hyper-standard token.

This work shows a considerable mismatch between perception and acoustic reality. The respondents reported that they heard a Michigan speaker (who, importantly, had been identified as one) use the canonical forms of the vowel rather than the shifted ones. Why are these respondents so inaccurate in this task?

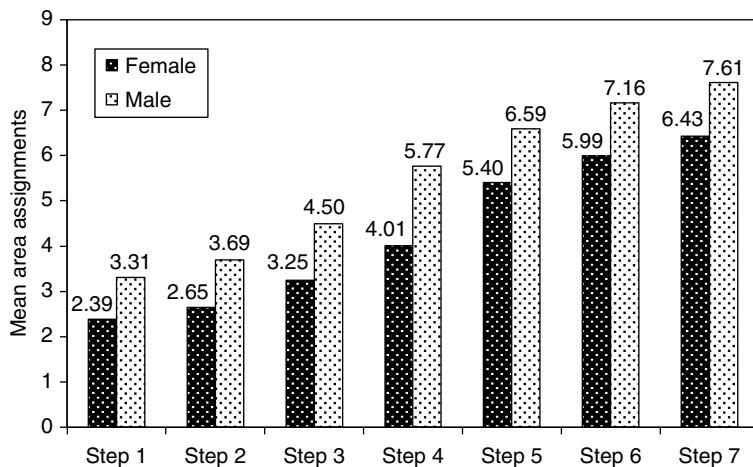
When these respondents are presented with data from a speaker who they think is a fellow Michigander, the stereotype of Michigan English as standard emerges (see above). As a result of this folk stereotype, the respondent selects the "standard" vowel. The linguistically secure can alter their perceptual and even production systems easily since they cannot conceive that their own performance would stray from the standard (i.e., their norms).

Table 10.3 Formant values of tokens offered to respondents to match with the vowel in the speaker's pronunciation of *last* (Niedzielski 1999, 74).

<i>Token</i>	#F1	F2	<i>Label</i>
1	900	1530	hyper-standard
2	775	1700	canonical
3	700	1900	actual token

Table 10.4 Respondent matching results for the vowel in *last* (adapted from Niedzielski 1999, 72).

<i>Token</i>	1	2	3	
	<i>hyper</i>	<i>canonical</i>	<i>actual</i>	
	<i>standard</i>	/æ/	<i>token</i>	<i>Total</i>
n =	10%	90%	0%	
	4	38	0	42

**Figure 10.19** Assignment of seven-step monophthongized male and female samples of *guide* to the nine sites in Figure 10.7 (Plichta and Preston 2005, 121).

Michiganders are so linguistically secure that they acoustically recalibrate the vowels of those around them and avoid noticing change or difference.

Attitudinal factors, however, can also be shown to interact with regional and social features simultaneously. Table 10.2 showed the relative success of US respondents in identifying the North-South location (Figure 10.7) of seven degrees of /aɪ/ monophthongization (a southern US speech caricature). The more monophthongal the vowel, the more southern the identification. That study (Plichta and Preston 2005), however, provided both male and female (resynthesized) samples of the word *guide*. Figure 10.19 shows the mean score assignments separated by sex of speaker.

Since the degree of monophthongization for the male and female voices was exactly equal through resynthesis, why would women's voices be consistently identified as "more northern" (or men's as "more southern"), as determined by independent t-tests? The full answer lies not just in the perception of region (/aɪ/ monophthongization is southern), perception of degree (more monophthongization is more southern), but also social (women are more northern/men more southern with equal degrees of monophthongization). Attitudes to region very clearly cut across the regional/perceptual/social nature of this task. It is a sociolinguistic commonplace that women tend to be more standard speakers than men (e.g., Trudgill 1972). It is an equally strong stereotype that English in the South of the US is (perhaps along with the New York City/New Jersey area) the least correct English in the country.

10.10 Conscious and Nonconscious

The experimental work outlined above will surely lead the up-to-date reader to ask whether newer techniques in the exploration of attitudes that purport to elicit participants' nonconscious attitudes toward variety are being undertaken. Indeed they are, and they involve not only the time-honored matched guise mode but also reaction-timed techniques (including so-called implicit association tests, "IATs"), eye-tracking measures, and even neurological responses. One example will have to suffice. Experiments performed by Koops, Gentry, and Pantos (2008) reveal implicit knowledge of the correlation between variation and age, using photographic priming and eye-tracking. In Houston, Texas, older Anglo speakers merge high front lax vowels before nasals; however, these vowels are not merged by younger Anglos. Direct measures of language attitudes do not reveal knowledge of this variation. However, Koops *et al.* show results that suggest that respondents are in fact implicitly aware of this variation. When primed with a photograph of an older speaker, respondents fixate longer on words that are homophonous (e.g., *rinse* versus *rents*) in the merged (but not the unmerged) dialect.⁴

This conscious-nonconscious split in PD studies is an important one. In a recent proposal, Kristiansen (2009) finds that Danes from all over Denmark say that they like their home variety best, but also that, when a carefully constructed matched-guise test is taken, they seem instead to prefer the emerging "New Copenhagen" standard, the features of which are influencing Danish across the entire country. If matched guise is an actually nonconscious (or implicit) method of collection (but see Preston 2009), and if the generalization reached about this dichotomy for Denmark is found in other areas, these different methods of investigation will prove essential to PD and dialectology in general, perhaps particularly in those places where standardized or more widespread forms are replacing local ones.

10.11 Conclusion

Readers will by now have realized that the term *perception* in this chapter has referred to two different things. On the one hand, it refers to the ideas that respondents have about the facts around them that surface in such tasks as drawing dialect boundaries on a blank map or assigning attributes to a variety's speech. On the other, it refers to the perceptual abilities respondents possess that allow them not only to recognize variety differences but also to detect subtle differences in specific linguistic markers of variety. The term *variety* has also appeared with greater frequency after the almost exclusively regional considerations of the first section, but the first sentence of the introduction notes that sociolinguistic factors were from early on a consideration of most work in PD, in keeping with Chambers and Trudgill,

who declare that “[d]ialectology without sociolinguistics at its core is a relic” (1998, 188). This chapter has illustrated the importance of social groups in PD, in both senses of “perception.”

Finally, however, the role of attitude has been shown to cut across both these concerns. Respondents delineate areas as distinct or different on the basis of their likes and dislikes with respect to speakers and the stereotypes that respondents hold of them, giving concrete expression to Silverstein’s notion of higher-order *indexicality*, in which the attributes of people (slow, smart, fun-loving, etc.) are assigned to their language variety and, in fact, become intrinsic parts of that variety’s description (2003). Respondents hear (and refuse to hear) the linguistic details of variety based on those same attitudes, adding another dimension to the second definition of perception.

Just as Chambers and Trudgill claim that dialectology without sociolinguistics is a relic, I believe that the work in PD over the years has shown that dialectology without PD is only half the story. The study of what people identify (in both regional and social senses), hear, think they hear, process, comprehend, and hold attitudes towards is a necessary part of the scientific investigation of linguistic variation.

NOTES

- 1 Critical discourse analysts have found it easy to make this jump, though not everyone agrees (e.g., Widdowson 1998).
- 2 This is the repositioning of vowels in the Inland North of the US that is known as the Northern Cities Shift (Labov *et al.* 2006).
- 3 Preston (2010) shows similarly bad performances by southeastern Michiganders, who are involved in the same Northern Cities Shift, in the comprehension of single-word tokens heard in isolation.
- 4 Preston and Niedzielski (2013) review a number of these recent studies. Two anthologies, Prikhodkine and Preston (2015) and Babel (2015), focus on the nonconscious-conscious split in attitude studies.

REFERENCES

- Babel, Anna, ed. 2016. *Awareness and Control in Sociolinguistic Research*. Cambridge: Cambridge University Press.
- Chambers, Jack, and Peter Trudgill. 1998. *Dialectology*, 2nd ed. Cambridge: Cambridge University Press.
- Cukor-Avila, Patricia. Forthcoming. A variationist approach to studies of language regard. In Erica Benson, Betsy Evans, and James Stanford (eds), *Language regard: Methods, variation, and change*. Cambridge: Cambridge University Press.
- Daan, Jo. 1969 [1999]. “Dialekten”. In *Van Randstad tot Landrand*, edited by Jo Daan and Dirk Blok, 7–43. Amsterdam: Nord-Hollandsche U.M. (Translated as “Dialects.” In *Handbook of Perceptual Dialectology*, Vol. 1, edited by Dennis R. Preston, 9–30. Amsterdam: Benjamins.)
- Evans, Betsy. 2011. “Seattletonian to faux hick: Perceptions of English in Washington State.” *American Speech*, 86(4): 383–413.
- Ginneken, Jacobus van. 1913 [1928]. *Handboek der Nederlandsche Taal*. 's Hertogenbosch: Malmberg.
- Goeman, Ton. 1989 [1999]. “Dialectes et jugements subjectifs des locuteurs: Quelques remarques de méthode à propos d'une controverse.” In *Espaces Romans: Études de Dialectologie et de Géolinguistique Offertes à Gaston Tuaillet*, Vol. II (various editors), 532–544. Grenoble: Éditions Littéraires et Linguistiques de l’Université de Grenoble (ELLUG). (Translated as “Dialects and the Subjective Judgments of Speakers: Remarks on Controversial Methods” In *Handbook of Perceptual Dialectology*, Vol. 1, edited by Dennis R. Preston, 135–144. Amsterdam: Benjamins.)
- Graff, David, William Labov, and Wendell Harris. 1986. “Testing listeners’ reactions to

- phonological markers of ethnic identity: A new method for sociolinguistic research." In *Diversity and Diachrony*, edited by David Sankoff, 45–58. Amsterdam: Benjamins.
- Grootaers, Willem. 1959. "Origin and nature of the subjective boundaries of dialects." *Orbis* 8: 355–384.
- Grootaers, Willem. 1964 [1999]. "La discussion autour des frontières dialectales subjectives." *Orbis* 13: 380–398. (Translated as "The Discussion Surrounding the Subjective Boundaries of Dialects" In *Handbook of Perceptual Dialectology*, Vol. 1, edited by Dennis R. Preston, 115–129. Amsterdam: Benjamins.)
- Gould, Peter, and Rodney White. 1974. *Mental Maps*. New York: Penguin.
- Hartley, Laura. 1999. "A view from the west: Perceptions of U.S. dialects by Oregon residents." In *Handbook of Perceptual Dialectology*, Vol. 1, edited by Dennis R. Preston, 315–332. Amsterdam: Benjamins.
- Hartley, Laura, and Dennis R. Preston. 1999. "The names of US English: Valley girl, cowboy, Yankee, normal, nasal, and ignorant." In *Standard English*, edited by Tony Bex, and Richard Watts, 207–238. London: Routledge.
- Herold, Ruth. 1990. Mechanisms Of Merger: The Implementation And Distribution Of The Low Back Merger In Eastern Pennsylvania. Unpublished doctoral dissertation, University of Pennsylvania.
- Inoue, Fumio. 1977/8. "Hōgen imēji no tahlenryō kaiseki (part 1)." *Gengo Seikatsu* 311: 82–91.
- Inoue, Fumio. 1978/9. "Hōgen imēji no tahlenryō kaiseki (part 1)." *Gengo Seikatsu* 312: 82–88.
- Inoue, Fumio. 1999. "Classification of dialects by image: English and Japanese." In *Handbook of Perceptual Dialectology*, Vol. 1, edited by Dennis R. Preston, 147–159. Amsterdam: Benjamins.
- Jeon, Lisa. 2012. *Drawing Boundaries and Revealing Language Attitudes: Mapping Perceptions of Dialects in Korea*. Unpublished MA thesis, University of North Texas.
- Kremer, Ludger. 1984 [1999]. "Die niederländisch-deutsche Staatsgrenze als subjektive Dialektgrenze." In *Grenzen en Grensproblemen: Een Bundel Studies uitgegeven door het Nedersaksisch Instituut va.n de R. U. Groningen ter Gelegenheid van zijn 30-Jarig Bestaan (Driemaandelijkse Bladen)* 7: 76–83. (Translated as "The Netherlands-German border as a subjective dialect boundary." In *Handbook of Perceptual Dialectology*, Vol. 1, edited by Dennis R. Preston, 31–36. Amsterdam: Benjamins.)
- Kristiansen, Tore. 2009. "The macro-level meanings of late-modern Danish accents." *Acta Linguistica Hafniensia* 41: 167–192.
- Labov, William. 2001. *Principles of Linguistic Change*, Vol. 3: *Cognitive and Cultural Factors*. Oxford: Wiley Blackwell.
- Labov, William, Sharon Ash, and Charles Boberg. 2006. *The Atlas of North American English*. Berlin: de Gruyter.
- Ladefoged, Peter, and Donald Broadbent. 1957. "Information conveyed by vowels." *Journal of the Acoustical Society of America*, 29(1), 99–104.
- Levinson, Stephen. 1983. *Pragmatics*. Cambridge: Cambridge University Press.
- Long, Daniel. 1990. "Hōgen ninchi chizu no kakikata to yomikata." *Proceedings of the Dialectological Circle of Japan* 50: 7–16.
- Long, Daniel, and Dennis R. Preston. 2002. *Handbook of Perceptual Dialectology*, Vol. 2. Amsterdam: Benjamins.
- Mase, Yoshio. 1964a [1999]. "Hōgen ishiki to hōgen kukaku." In *Nihon Hōgen Kenkyūkai*, edited by Misao Tōjō, 270–302. Tokyo: Tokyodo. (Translated as "Dialect consciousness and dialect divisions." In *Handbook of Perceptual Dialectology*, Vol. 1, edited by Dennis R. Preston, 71–99. Amsterdam: Benjamins.)
- Mase, Yoshio. 1964b (1999). "Hōgen ishiki ni tsuite: Washa no genkyūshita hōpenteki. tokuchō." *Nagano-ken Tanki Daigaku Kiyō*, 18: 1–12. (Translated as "On dialect consciousness: Dialect characteristics given by speakers." In *Handbook of Perceptual Dialectology*, Vol. 1, edited by Dennis R. Preston, 101–113. Amsterdam: Benjamins.)
- Milroy, Lesley, and Paul McClenaghan. 1977. "Stereotyped reactions to four educated accents in Ulster." *Belfast Working Papers in Language and Linguistics* 2(4): 1–11.
- Montgomery, Christopher. 2007. *Northern English Dialects: A Perceptual Approach*. Unpublished PhD dissertation, University of Sheffield.
- Montgomery, Christopher. 2012. "Mapping the perception of non-linguists in Northern England." In *Dialectological and Folk Dialectological Concepts of Space: Current Methods and Perspectives in Sociolinguistic Research on Dialect Change*, edited by Sandra Hansen, Christian Schwarz, Phillip Stoeckle, and Tobias Streck, 164–178. Berlin: de Gruyter.
- Montgomery, Christopher, and Phillip Stoeckle. 2013. "Geographic information systems and perceptual dialectology: A method for processing draw-a-map data." *Journal of Linguistic Geography* 1(1): 52–85.

- Moreno Fernández, Juliana and Francisco Moreno Fernández. 2002. "Madrid perceptions of regional varieties in Spain". In *Handbook of Perceptual Dialectology*, Vol. 2, edited by Daniel Long and Dennis R. Preston, 295–320. Amsterdam: Benjamins.
- Niedzielski, Nancy. 1999. "The effect of social information on the perception of sociolinguistic variables." *Journal of Language and Social Psychology*, 18(1): 62–85.
- Pearce, Michael 2009. A perceptual dialect map of North East England. *Journal of English Linguistics* 37:162–192.
- Peterson, Gordon, and Harold Barney. 1952. "Control methods used in a study of the vowels." *Journal of the Acoustical Society of America* 24(2): 175–184.
- Plichta, Bartłomiej, and Dennis R. Preston. 2005. "The /ay/ s have it: The perception of /ay/ as a north-south stereotype in United States English." In *Subjective Processes in Language Variation and Change (Acta Linguistica Hafniensia)* 37, edited by Tore Kristiansen, Peter Garrett, and Nikolas Coupland, 107–130. Copenhagen: Reitzel.
- Preston, Dennis R. 1982. "Perceptual dialectology: Mental maps of United States dialects from a Hawaiian perspective." *University of Hawaii Working Papers in Linguistics*, 14(2): 5–49.
- Preston, Dennis R. 1985. "Southern Indiana perceptions of 'correct' and 'pleasant' speech." In *Methods/Méthodes V (Papers from the Fifth International Conference on Methods in Dialectology)*, edited by Henry Warkentyne, 387–411. Victoria, BC: University of Victoria.
- Preston, Dennis R. 1989. *Perceptual dialectology: Nonlinguists' view of areal linguistics*. (Topics in Sociolinguistics 7). Dordrecht/Providence: Foris.
- Preston, Dennis R. 1993. "Folk dialectology." In *American Dialect Research*, edited by Dennis R. Preston, 333–377. Amsterdam: Benjamins.
- Preston, Dennis R. 1994. "Content-oriented discourse analysis and folk linguistics." *Language Sciences* 16(2): 285–330.
- Preston, Dennis R. 1996. "Where the worst English is spoken." In *Focus on the USA*, edited by Edgar Schneider, 297–360. Amsterdam: Benjamins.
- Preston, Dennis R., ed. 1999a. *Handbook of Perceptual Dialectology: Vol. 1*. Amsterdam: Benjamins.
- Preston, Dennis R. 1999b. "Introduction." In *Handbook of Perceptual Dialectology: Vol. 1*, edited by Dennis R. Preston, xxiii–xl. Amsterdam: Benjamins.
- Preston, Dennis R. 1999c. "A language attitude approach to the perception of regional variety." In *Handbook of Perceptual Dialectology: Vol. 1*, edited by Dennis R. Preston, 359–373. Amsterdam: Benjamins.
- Preston, Dennis R. 2009. "Are you really smart (or stupid, or cute, or ugly, or cool)? Or do you just talk that way?" In *Language Attitudes, Standardization and Language Change: Perspectives on Themes Raised by Tore Kristiansen on the Occasion of his 60th Birthday*, edited by Marie Maegaard, Frans Gregersen, Pia Quist, and Jens Jørgensen, 105–129. Oslo: Novus.
- Preston, Dennis R. 2010. "Belle's body just caught the fit gnat." In *A Reader in Sociophonetics*, edited by Dennis R. Preston, and Nancy Niedzielski, 241–252. Berlin: de Gruyter.
- Preston, Dennis R., and Nancy Niedzielski. 2013. "Approaches to the study of language regard." In *Language (De)standardisation in Late Modern Europe: Experimental Studies*, edited by Tore Kristiansen, and Stefan Grondelaers (eds), 287–307. Oslo: Novus.
- Prikhodkine, Alexei, and Dennis R. Preston, eds. 2015. *Language Attitudes: Variation, Processes, and Outcomes*. Amsterdam: Benjamins.
- Purnell, T., William Idsardi, and John Baugh, J. 1999. Perceptual and phonetic experiments on American English dialect identification. *Journal of Language and Social Psychology* 18,1:10–30.
- Rakerd, Brad, and Bartłomiej Plichta. 2003. "More on perceptions of /a/ fronting." Paper presented at NWA 32, University of Pennsylvania.
- Rensink, Wim. 1955. "Dialektindeling naar opgaven van medewerkers." *Amsterdam Dialectbureau Bulletin*, 7: 20–23. (Translated as "Informant classification of dialects." In *Handbook of Perceptual Dialectology: Vol. 1*, edited by Dennis R. Preston, 3–7. Amsterdam: Benjamins.)
- Ryan, Ellen Bouchard, and Howard Giles, eds. 1982. *Attitudes towards Language Variation: Social and Applied Contexts*. London: Arnold.
- Sibata, Takeshi. 1959. "Hōgen kyōkai no ishiki." *Gengo Kenkyū* 36: 1–30. (Translated as "Consciousness of dialect boundaries." In *Handbook of Perceptual Dialectology: Vol. 1*, edited by Dennis R. Preston, 39–62. Amsterdam: Benjamins.)
- Silverstein, Michael. 2003. "Indexical order and the dialectics of sociolinguistic life." *Language and Communication*, 23: 193–229.

- Stoeckle, Phillip. 2012. "The folk linguistic construction of local dialect areas – linguistic and extra-linguistic factors." In *Dialectological and Folk Dialectological Concepts of Space: Current Methods and Perspectives in Sociolinguistic Research on Dialect Change*, edited by Sandra Hansen, Christian Schwarz, Phillip Stoeckle, and Tobias Streck, 142–163. Amsterdam: Benjamins.
- Tamasi, Susan. 2003. *Cognitive Patterns of Linguistic Perceptions*. Unpublished PhD dissertation, University of Georgia.
- Tōjō, Misao. 1953. *Nihon Hōgengaku*. Tokyo: Yoshikawakobunkan.
- Trudgill, Peter. 1972. "Sex, covert prestige and linguistic change in the urban British English of Norwich." *Language in Society*, 1(2): 179–195.
- Tucker, G. Richard, and Wallace Lambert. 1969. "White and Negro listeners' reactions to various American-English dialects." *Social Forces*, 47: 463–468.
- Twilfer, Daniela. 2010. *Dialektgrenzen im Kopf: Der Westfälische Sprachraum aus Volkslinguistischer Perspektive*. Gütersloh: Verlag für Regionalgeschichte.
- Weijnen, Antonius. 1946. "De grenzen tussen de Oost-Noordbrabantse dialecten onderling." In *Oost-Noordbrabantse Dialectproblemen*, edited by Antonius Weijnen, J. Renders, and Jacobus van Ginneken, 1–15. Amsterdam: Noord Hollandsche U.M.
- Weijnen, Antonius A. 1968 (1999). "Zum Wert subjektiver Dialektgrenzen." *Lingua*, 21: 594–96. (Translated as "On the value of subjective dialect boundaries." In *Handbook of Perceptual Dialectology: Vol. 1*, edited by Dennis R. Preston, 131–133. Amsterdam: Benjamins.)
- Weinreich, Uriel, William Labov, and Marvin Herzog. 1968. "Empirical foundations for a theory of language change." In *Directions for Historical Linguistics: A Symposium*, edited by Winfred Lehmann, and Yakov Malkiel, 95–188. Austin, TX: University of Texas Press.
- Widdowson, Henry. 1998. "The theory and practice of critical discourse analysis (review article)." *Applied Linguistics*, 19(1): 136–151.
- Willems, Pieter. 1886. *De enquête werd gehouden in 1886 de antwoorden zijn het eigendom van de Koninklijke Vlaamsche Academie voor Taal- en Letterkunde, en worden daar bewaard*. Institutes of Dialectology and Phonetics in Leuven, the Catholic University Nijmegen, and the Meertens Institute, Amsterdam.

11 Dialect Intelligibility

CHARLOTTE GOOSKENS

11.1 Introduction

The present chapter focuses on the communicative consequences of dialectal variation. One of the main functions of language is to enable communication, not only between speakers of the same variety but also between people using different accents, dialects or closely related languages. Most research on dialect intelligibility is relatively recent, especially when it comes to actual testing. The methods for testing and measuring are getting more and more sophisticated, including web-based experiments that allow for collecting large amounts of data, opening up new approaches to the subject.

Some of the first to develop a methodology to test dialect intelligibility were American structuralists (e.g., Hickerson, Turner, and Hickerson 1952; Pierce 1952; Voegelin and Harris 1951), who tried to establish mutual intelligibility among related indigenous American languages around the middle of the previous century. They used the so-called recorded text testing (RTT) method. This methodology has been standardized and is still being used, for example in the context of literacy programs where a single orthography has to be developed that serves multiple closely-related language varieties (Casad 1974; Nahhas 2006). Since then, numerous intelligibility investigations have been carried out with various aims, for instance to resolve issues that concern language planning and policies, second-language learning, and language contact. Data about distances between varieties and detailed knowledge about intelligibility can also be important in sociolinguistic studies. Varieties that have strong social stigma attached to them could unfairly be deemed hard to understand (Giles and Niedzielski 1998; Wolff 1959). The relationship between attitudes and intelligibility is not a straightforward one, but advances in the field of intelligibility testing provide sociolinguists with objective data to help to resolve conflicts that arise concerning non-standard varieties. Knowledge about mutual intelligibility is also needed for standardization and development of new orthographies in communities where no standardized orthography exists.

In this chapter we will first deal at some length with two questions concerning dialect intelligibility, which we consider of major importance to linguistics in general and to dialectology in particular. We will look into the role of intelligibility in the definition of “dialect” and “language,” and examine how intelligibility can be used to validate research on distances between dialects. An overview of factors that determine intelligibility is provided, and finally, some gaps in our knowledge of processes and phenomena in the area of dialect intelligibility are signalled, and desiderata for future research are formulated.

11.2 Definition of “Dialect” and “Language”

The question of how to define a “language” as opposed to a “dialect” is one of the oldest and most central questions that linguists have asked themselves. Reasons for wanting to distinguish between the two concepts have sometimes been theoretical but more often practical or political in nature. For example, people might want to know how many languages there are in the world, and to be able to answer this question it is necessary to be able to define when a variety (or a group of varieties) can be considered a language in its own right. Also, it has been important for language planning and policies at both the national and the more global levels to find criteria that define a language variety as a language. A language often represents a community and is tightly connected to standardization processes and development of new orthographies. For people fighting for the rights of a language variety it is of great importance that the variety is recognized as a language rather than a dialect. Official recognition will give the variety a stronger position. This becomes clear from part 1 of the European Charter for Regional Minority Languages (1992). The right to use one’s variety in public life (e.g., in educational, juridical, administrative, or media contexts) is dependent upon the status of this variety as an official, regional or minority language. Speakers of varieties that are classified as “dialects” do not have these rights under the Charter. On the other hand, there are also areas in which languages that are clearly different are characterized as dialects of a single language because it is desirable to preserve the unity of the area.

Discourse on language rights and standardization of language varieties rests on an underlying assumption that one can somehow objectively identify which varieties are languages and which are dialects. Kloss (1967) introduced the terms *AusbauSprache* (language by development) and *Abstandssprache* (language by distance) for analyzing and categorizing language varieties that are closely related and often are used within the same society. An *AusbauSprache* typically has its own standardized form used autonomously with respect to other languages, it is often taught in schools, and it is used as a written language for a wide variety of social and political functions, possibly including that of an official national language. In an *Ausbau* definition, “languages” and “dialects” are social constructs definable only in terms of their socio-political and cultural status and breadth of use, and they have little to do with independently identifiable structural entities. An often-mentioned example of *Ausbau* languages are the Scandinavian languages (Danish, Norwegian, and Swedish) that are so closely related that the speakers of the three languages can, with some effort, understand each other. Still, they are regarded as different languages because they have distinct, codified, standardized forms, with their own orthographies, grammar books, and literatures, and are spoken in three separate nation states. Another example is Serbian and Croatian. These languages may be even closer to each other than the Scandinavian languages are, but speakers of “Serbo-Croatian” insist that they are two different languages. In fact, the differences are mainly lexical, with very few differences in pronunciation and grammar. The perceived differences are reinforced by other factors such as different alphabets, religions and ethnicity. Tamburelli (2014) points out that the *Ausbau* definition may lead to a circularity effect: the use of a linguistic variety in educational, juridical, administrative, or media contexts is a right reserved for varieties with language status, but to achieve the “language” label a variety has to have a certain socio-political status. This means that the language varieties that language legislation is meant to protect may in fact be excluded *a priori* from this protection.

Linguists in general prefer to define languages as *Abstandssprachen*. In this view, one language variety is called an *Abstandssprache* with respect to another language variety if the two are so different from each other that they are in fact different languages. Kloss (1967) left unspecified exactly how the differences between two language varieties are to be measured objectively, presumably because he lacked the tools to do so. Methods to measure linguistic

distances objectively have now been developed by dialectometrists (see Chapter 7 of this volume). The problem remains, however, that languages do not differ along just one dimension, but may differ to different extents in their lexicon, phonetics and phonology, morphology and syntax. It is not clear how much weight should be given to these separate linguistic dimensions when assessing overall distance. We will return to this point in Section 11.4.

Maybe to circumvent the problem of how to weight different linguistic dimensions, Trudgill (2000) introduced the intelligibility criterion, and this has become the standard—or at least the primary—criterion among many linguists. According to this definition, dialects are mutually intelligible varieties, whereas languages are so linguistically different that their speakers are unable to understand each other. From this it follows that a language is a collection of mutually intelligible dialects. Intelligibility was used as the main criterion by the compilers of *Ethnologue*, an online database of all the world's known languages, to decide what should count as a language (Lewis *et al.* 2013). In the seventeenth edition (2013), *Ethnologue* contains a list of 7,105 languages. It has been criticized for splitting language varieties into too many languages, but Hammarström (2005) asserts that *Ethnologue* is consistent with specialist views most of the time.

The intelligibility criterion for defining languages is not without its problems, which has caused some linguists to reject it (see Hammarström (2008) for an overview of citations). We will look further into some of these objections below.

First, intelligibility is not easy to measure. It can be tested by means of opinion testing, whereby subjects are asked to indicate (with or without speech fragments) how well they think they understand the language at hand. However, these intuitive and impressionistic judgments may be distorted by extra-linguistic factors (see below). Intelligibility may also be determined experimentally by means of functional tests, which typically express the degree of intelligibility as the percentage of input that is correctly recognized by the subject. This approach comes with its own problems. It is generally difficult to abstract away from individual speakers and choice of test. In addition, an effort must be made to avoid priming effects, ceiling effects, excessive memory load, and other unwanted issues. These considerations often make it rather time-consuming both to develop suitable tests and to carry out the tests themselves. Gooskens (2013) lists a number of methods for measuring dialect intelligibility, and discusses their advantages and disadvantages.

Both opinion testing and functional testing will result in numbers that express how well subjects can understand a language (variety). This means that intelligibility is a matter of degree. It is difficult to decide when the mutual intelligibility of two varieties is so high that they should be considered dialects of the same language rather than separate languages. Thus far, it has not been possible to define an intelligibility threshold. Hudson (1996, 35) states that "this is clearly a question which is best avoided, rather than answered, since any answer must be arbitrary." Still, a number of studies have made an attempt to establish a reference point, below which it is difficult to achieve successful communication (see Tamburelli (2014) for an overview).

The fact that intelligibility scores are a matter of degree reflects the situation in many dialect areas. Traditional dialectologists present the geographical spread of dialectal features, for instance a particular word form or pronunciation, by drawing isoglosses (lines separating features on a map; for example, Weijnen (1941) and several of the other chapters in this volume). A dialect division is said to be major if several isoglosses coincide (isogloss bundles). However, isoglosses usually only coincide approximately, resulting in different dialect areas with a transition zone in between (so-called "dialect continua"). A well-known example of a dialect continuum is found in the Dutch-German dialect area. In the early nineteenth century, one could start from the far south of the German-speaking area and travel to the far west of the Dutch-speaking area without encountering any sharp boundary across which mutual intelligibility is broken, but the two end points of this chain are speech

varieties so different from one another that they are not mutually intelligible. In Europe there are many other dialect continua, for example, the Romance continuum stretching across the Iberian peninsula through France and parts of Belgium down to the southern tip of Italy and as far east as the Black Sea, including Portuguese, Spanish, Catalan, French, Italian, and Romanian. Outside of Europe also dialect continua are found, for example in the Chinese, Arabic, Indic, Turkic, and Algonquian language areas.

Since we often have to deal with dialect continua and the resulting gradient intelligibility, it is inevitably quite difficult to calculate how many, and which, languages are spoken in a certain area, or indeed worldwide. Hammarström (2008), however, adopted an abstract perspective in order to show that it may in fact be possible to state such numbers. Say, for example, that we are dealing with a dialect chain of three dialects A, B, and C in a language area where the neighboring dialects (A and B, or B and C) are mutually intelligible, whereas the non-neighboring dialects (A and C) are not. Applying the intelligibility principle for defining languages we must be dealing here with two languages (A/B and C, or A and B/C). However, although Hammarström shows how to count the number of languages in a continuum, he fails to define languages uniquely by means of this line of reasoning. Tamburelli (2014) suggests that a choice between the two possible options in the example above can be made by measuring objective linguistic distances or testing intelligibility. The two language varieties that are linguistically closest or show the highest level of mutual intelligibility should be considered dialects of a single language.

As an objection to using intelligibility measurements for the definition of languages, some linguists have contended that intelligibility scores may be influenced by extra-linguistic factors. Subjects may be influenced by their positive or negative biases and attitudes toward the country and its speakers, interest in or familiarity with other cultures, political borders, or the geographical distance to the place where the language is spoken. Also, the personal characteristics of the subjects, such as age, amount of schooling, psycho-cognitive traits, metalinguistic awareness, previous experience, knowledge of various registers and vocabulary in their own language, learning style, fatigue, and motivation may influence their intelligibility rates. We will deal in greater detail with such extra-linguistic factors in Section 11.4.

As a result of the extra-linguistic and personal characteristics of (groups of) subjects, mutual intelligibility is often asymmetric, such that one group of speakers has more difficulty understanding the other variety than the other way round. Asymmetric intelligibility has been described in the literature for many language pairs, including Swedish and Danish (Delsing and Lundin Åkesson 2005; Schüppert 2011), Spanish and Portuguese (Jensen 1989), and the indigenous Californian Indian languages Achumawi and Atsugewi (Merriam 1926; Voegelin and Voegelin 1946). Asymmetric intelligibility is mostly attributed to extra-linguistic factors such as attitude and contact, but there is also evidence that linguistic characteristics of the language varieties may cause asymmetry. For example, Gooskens and Van Bezooijen (2006) showed that asymmetries in the number of non-cognates and the opacity of the relatedness of cognates between Dutch and Afrikaans result in asymmetric intelligibility. These asymmetries are caused by divergent historical developments in Dutch and Afrikaans with respect to lexicon, grammar, and spelling. German-Dutch mutual intelligibility has also been found to be asymmetrical, with German being easier to understand for speakers of Dutch than Dutch is for speakers of German. This finding has been attributed to the fact that German is an obligatory subject at school in the Netherlands and that many Dutch people watch German television. This, however, appears not to be the whole story. Gooskens, Van Bezooijen, and Van Heuven (2015) presented Dutch and German cognate nouns to Dutch and German children between 9 and 12 years of age who did not know the other language or a related dialect, and who all expressed positive attitudes toward the other language, its speakers and the country. The Dutch subjects proved to be significantly better at understanding the German

cognates than the German subjects were at understanding the Dutch cognates. However, since extra-linguistic factors had been ruled out, this asymmetry must have a linguistic basis. A closer look at the data revealed that asymmetries between the two languages are found at the phonetic level and in the presence or absence of neighbors, that is, competing word forms that are very similar to the stimulus word.

Reflecting on the objections to intelligibility as the most important criterion for distinguishing between dialects and languages (the *Abstand* criterion), we see that it is not unproblematic to use this criterion. More research is needed before we will be able to establish when two varieties are so different that they are no longer mutually intelligible, and which linguistic factors play a role. Since there is no universally accepted criterion for distinguishing a language from a dialect, the examples provided in the rest of this chapter are from language varieties that are traditionally referred to as dialects as well as closely related varieties that are mostly referred to as languages.

11.3 Intelligibility as a Measure of Distance

As mentioned in the previous section, it is not a straightforward task to quantify distance. The problem is that languages do not differ along just one dimension, but may differ at all linguistic levels. At each of the linguistic levels, languages may furthermore vary on many different parameters. For example, vowel distances may differ from consonant distances, and the percentages of common loanwords may differ from the percentages of common inherited words (see Section 11.4). Ideally, we would like to express the linguistic distance between language varieties using a single number on a one-dimensional scale. However, there is no *a priori* way of weighing the different linguistic dimensions. As objective techniques have become more sophisticated, resulting in more methods by which language varieties might be distinguished, many researchers have felt an increasing need to “validate” objective methods by means of subjective, behavioral tests (Heeringa *et al.* 2006).

Intelligibility testing is an adequate way of determining how different two languages or language varieties are. If two language varieties have a high degree of mutual intelligibility the linguistic distance must be small, and if they have a low degree of mutual intelligibility the distances are likely to be larger, unless some extra-linguistic factor interferes (see Section 11.4.2). A few investigations have been carried out to validate objectively-measured linguistic distances by means of functional intelligibility tests. Gooskens, Heeringa, and Beijering (2008) assessed the intelligibility of seventeen Scandinavian language varieties and standard Danish among young Danes from Copenhagen by means of a translation task. In addition, distances between standard Danish and each of the seventeen varieties were measured at the lexical level, expressed as the percentage of cognates, and at the phonetic level, by means of Levenshtein distances, a dialectometric technique that calculates distances on the basis of matched segment strings (for an explanation of this algorithm, see Nerbonne and Heeringa 2010). They correlated the intelligibility scores with the linguistic distances and found fairly high, significant correlations. Phonetic distance was a better predictor of intelligibility ($r = -.86$) than lexical distance ($r = -.64$). Similar results are reported by Tang and Van Heuven (2008), who tested mutual intelligibility among 15 Chinese dialects by means of a word-intelligibility and a sentence-intelligibility task. They correlated the scores with measures of lexical similarity and phonological correspondences, and found significant correlations of between .75 and .79. These results show that objective distance measures reflect experimental intelligibility results to a large extent, yet not perfectly. We will suggest explanations for the discrepancy in Section 11.4.

As explained in Section 11.2, it is time consuming both to develop and to carry out suitable functional tests. An easy and efficient alternative to get a quick impression of the intelligibility of a language is to ask subjects to rate on scale(s) how well they think they understand the language at hand. Such opinion testing may provide a shortcut to functional intelligibility tests.

In the investigation by Tang and Van Heuven (2008) described above, intelligibility rates gained by opinion testing were correlated with functional intelligibility rates, yielding correlations of between .70 and .80. These imperfect correlations suggest that tests of impressionistic intelligibility and functional intelligibility tests are sensitive to different factors. In order to make a choice between the two, the authors generated hierarchical cluster trees from their data matrices and compared the results to traditional taxonomies of Chinese dialects proposed by dialectologists. Functional intelligibility measures correspond better to traditional dialect taxonomies than opinion scores do. The authors therefore advocate that whenever the resources are available, mutual intelligibility should be tested functionally. The results also show that mutual intelligibility can to some extent be used as a criterion to illustrate the genetic relationship between speech varieties.

It seems likely that linguistic distance judgments are based on how difficult a listener thinks it would be for him or her to understand speakers of the other language variety. Another shortcut to functional intelligibility testing could therefore be to ask listeners to judge linguistic distances. Tang and Van Heuven (2009) compared the results from their intelligibility tests to perceived linguistic distances, which are gathered by having subjects listen to speech recordings and asking them to judge how deviant the varieties are from their own variety. They found significant correlations of .74 for word intelligibility and .78 for sentence intelligibility. Again, this shows that although there is a relatively large overlap between the two measurements they are still sensitive to different phenomena. The extent to which perceived distance is a reflection of intelligibility, and why the two differ, remains uncertain.

11.4 The Role of Linguistic and Extra-Linguistic Factors for Intelligibility

In the previous section we stated that intelligibility measurements can be used as a way of expressing linguistic distances between language varieties in a single number on a one-dimensional scale. Language varieties may differ at all linguistic levels, and when testing intelligibility extra-linguistic factors such as attitude and linguistic experience may also play an important role. In this section, an overview will be given of investigations that have dealt with the role of various linguistic and extra-linguistic factors in the intelligibility of dialects and closely related languages.

11.4.1 The Role of Linguistic Factors

11.4.1.1 Lexicon

At the lexical level the linguistic distance is often expressed as the percentage of non-cognates between two language varieties. The larger the proportion of non-cognates, the lower the intelligibility will be. The Scandinavian investigations discussed earlier revealed that lexical distances can only predict intelligibility to a limited extent. There are a number of explanations for this finding.

First, it is difficult to predict the effect of individual lexical differences. One single non-cognate word in a sentence or text can lower intelligibility considerably if the non-cognate word is a central concept. For example, one of the texts that were used in an investigation of mutual intelligibility between Swedish and Danish by Delsing and Lundin Åkesson (2005) was about frogs. Since the word for "frog" is a non-cognate noun (Danish *fro*, Swedish *groda*), the whole text was very difficult to understand. On the other hand, if the non-cognate words in a text have little semantic content or can easily be interpreted from the context, lexical differences will have less influence on intelligibility.

Furthermore, it is possible that listeners understand some non-cognate words because they are familiar with the words from previous experience with the test language, or because they are loanwords from a language that they are familiar with. For example, Swedish has many French loanwords that are not found in Danish. Knowledge of French might therefore enable a Dane to understand some Swedish non-cognates.

Whereas non-cognates will in principle hinder intelligibility, so-called “false friends” may cause even larger problems because they may actually mislead the listener. False friends are pairs of words in two language varieties that sound similar, but differ in meaning. They may arise because words with shared etymology shifted in meaning, or acquired additional meanings in at least one of these languages. For example, the meanings of German *Meer* “sea” and its Dutch cognate *meer* “lake” have changed over time. In certain cases, false friends evolved from words with different etymological roots. Words usually change by small shifts in pronunciation accumulated over long periods of time, and sometimes converge by chance towards the same pronunciation or spelling. For example, the English word *bra* has a different etymology from the Swedish word *bra* “good.”

11.4.1.2 Phonetics/Phonology

The results presented earlier on the relationship between phonetic distances and intelligibility show that at an aggregate level, that is, summed over larger stretches of speech, phonetic distances are a good predictor of the intelligibility of whole texts, and in the Scandinavian case they are better predictors than lexical distances.

A number of investigations have focussed on the role of specific phonetic characteristics in the intelligibility of words. In their investigation of the intelligibility of 17 Scandinavian language varieties among Danes (see Section 11.3), Gooskens, Heeringa, and Beijering (2008) investigated the role of different consonant and vowel operations (insertions, deletions, substitutions, lengthenings, and shortenings). The correlations for the consonants were significantly stronger than those for the vowels ($r = -.74$ versus $r = -.29$). Consonant substitutions play a particularly important role in intelligibility, probably because the “framework” of the word is changed when consonants in a word are substituted. By contrast with consonant substitutions, vowel substitutions play a negligible role in intelligibility.

To be able to determine the role of specific phonetic factors in detail, researchers have often chosen to test word intelligibility rather than the intelligibility of sentences or whole texts. The underlying assumption here is that word recognition is the key to speech understanding: if the listener correctly recognizes a minimal proportion of words, he or she will be able to piece the speaker’s message together. Van Bezooijen and Van den Berg (1999) looked at the basis of intelligibility ratings given by speakers of Standard Dutch, and tried to explain why three Dutch dialects and the closely related language Frisian yielded widely diverging results. They made a linguistic profile for each variety, distinguishing six categories of relationships between the target noun in the dialect and the semantically equivalent noun in Standard Dutch (no difference, difference in one vowel, difference in one consonant, differences in several phonemes, non-cognate). One of the results was that the intelligibility of Frisian was equal to that of West Flemish, so one would expect the two to have similar linguistic profiles. However, Frisian has considerably more instances of words that were identical to the Standard Dutch equivalent and considerably fewer non-cognates than West Flemish. Compared to West Flemish, vowel differences between Standard Dutch and Frisian were considerably less transparent. This would mean that Frisian is relatively difficult to understand not only for quantitative reasons, that is, because of the number of nouns showing the various relationships, but also for qualitative reasons, because of the types of deviations within particular categories. But since vowels

played a smaller role in the Scandinavian context (see above) it also means that the role of deviating vowels may be dependent upon the variety and the listeners.

Phonetic details may play an important role in the intelligibility of cognates in related languages and language varieties, in as yet unpredictable ways. Broad transcriptions are therefore sometimes unfit to be used as a basis for the calculation of the phonetic distance between pairs of words with a view to predicting intelligibility. As discussed in Section 11.2, Gooskens, Van Bezooijen, and Van Heuven (2015) tested the mutual intelligibility of German and Dutch among children with no previous knowledge of the test language. They found several cases in which the word in the stimulus language and the corresponding cognate in the response language were represented by the same sequence of phonetic symbols in the Levenshtein algorithm, but in which many subjects nevertheless did not succeed in recognizing the stimulus word. This holds, for example, for Dutch *zoon* /zo:n/ "son," which is phonemically transcribed with the same symbols as its German cognate *Sohn* /zo:n/, but which was nevertheless correctly identified by no more than 20.6% of the German subjects. The high proportion of incorrect responses suggest that there are subtle differences in the phonetic realizations of Dutch and German /z/, which are not expressed in the broad transcription the authors used, and which is commonly used in other intelligibility studies as well. On the other hand, there were cases in which different transcriptions of words nevertheless yielded high intelligibility. For example, half of the transcription symbols in Dutch *stad* /stat/ "city" differ from those in German *Stadt* /stat/, resulting in a Levenshtein distance of 50%, but the mutual intelligibility was high nonetheless (92.9% for the Dutch subjects and 94.1% for the German subjects).

In order to find out with which sound in the listener's native language a non-native sound from a closely-related language is identified, we may turn to the Perceptual Assimilation Model (PAM) developed by Catherine Best and her co-workers (e.g., Best 1995; Best, McRoberts, and Goodell 2001). PAM was developed to predict and explain the behavior of learners of a second language when first confronted with the sounds of the target language. The results of perceptual assimilation experiments reveal which categories in the listener's native language are likely to be matched with a non-native sound (Van Heuven 2008). Such knowledge might be used to weight phonetic differences differentially, for example, depending on the intuitions of listeners about the differences between the two segments involved in a substitution (see Wieling, Margaretha, and Nerbonne (2012) and references therein).

It should be noted that the effect of phonetic similarity between the stimulus and the intended response may be overruled by the presence of neighbors. However similar a stimulus and the intended response may be, if there is another word in the subject's language that is even closer to the stimulus, the latter has a high chance of being preferred. This will lead to (severely) reduced intelligibility for that word, especially in the absence of linguistic or extra-linguistic context.

Listeners are in general better at translating loanwords correctly than inherited words. Part of the explanation may be that they know the loanwords from the source language, but it is also possible that particular characteristics of loanwords make them easier to recognize. Loanwords may have specific segmental and/or prosodic properties that make them resistant to the linguistic changes affecting inherited words. They are often longer than inherited words because the word length of the loan-giving languages is generally longer, and we know from the literature that longer words are better recognized than shorter words (Wiener and Miller 1946; Scharpf and Van Heuven 1988). This is explained in terms of the relationship between word length and the number of "neighbors" competing to be recognized. Longer words have fewer neighbors than shorter words (Vitevitch and Rodriguez 2005). Furthermore, redundancy increases with word length, which is assumed to enhance intelligibility as well. Furthermore, inherited words have been part of the lexicon for a much longer time than loanwords, so that certain historical sound changes, which affected the inherited vocabulary were no longer active at the time the loans entered the language. As a consequence,

loan words in the neighboring language often have more transparent phonetic correspondences with their counterparts in the mother tongue than inherited words have.

11.4.1.3 Morphosyntax

Previous studies of intelligibility have focussed primarily on the role of lexical and phonetic factors. Although there is reason to believe that differences in morphology and syntax might degrade the ability to comprehend a closely related linguistic variety, this claim has hardly been tested. An exception is Hilton, Gooskens, and Schüppert (2013), who carried out an experimental investigation to see whether Danes' comprehension of the closely related language Norwegian is impeded by certain Norwegian grammatical constructions. They tested sentence comprehension in four different conditions to assess the relative effect on intelligibility of non-native morphosyntactic features as opposed to non-native phonology. The results indicated that word-order differences cause larger problems for listeners than morphological differences. However, the non-native phonology featured in the experiment impedes comprehension to a larger degree than the morphosyntactic differences do. Just as in the case of other linguistic factors, the role of morphosyntax may be language-dependent. In language areas with larger morphological and syntactic variability, morphosyntax may play a more important role in intelligibility than in areas with less variability.

11.4.2 The Role of Extra-Linguistic Factors

11.4.2.1 Attitude

The existence of negative attitudes or social stigmas attached to languages is often seen as a potential obstruction for successful intergroup communication. The fact that Danes understand Swedish better than Swedes understand Danish, for example, is often explained by less positive attitudes among Swedes toward the Danish language, culture, and people than vice versa (Delsing and Lundin Åkesson 2005). Wolff (1959) investigated mutual intelligibility between the closely related Nigerian Ijo languages Kalabari and Nembe, and reports that Nembe speakers claim to understand Kalabari, whereas speakers of Kalabari judge Nembe to be unintelligible to them. Wolff suggests that this asymmetry in intelligibility is linked to an asymmetry in language attitudes. He states that when his study was conducted, the Kalabari were the most prosperous group in the Eastern Niger Delta and that they regarded other Ijo-speaking groups as inferior to them.

Boets and De Schutter (1977) found low intelligibility to correlate with low appreciation. According to Boets and De Schutter, the (subjective) appreciation scores are determined by the (objective) intelligibility scores. In the literature (e.g., Wolff 1959; Van Bezooijen and Gooskens 2007) the opposite is often contended, namely that low (high) intelligibility is caused by low (high) appreciation. It is assumed that the reported or measured comprehension problems are not so much due to a lack of transparency of the meaning of the language at hand, but rather to a lack of motivation on the part of the listeners.

It has not been possible, so far, to establish the direction of the causality, that is, whether negative attitudes are a result of poor intelligibility, or poor intelligibility is a result of negative attitudes caused by some other factor. Language attitude research shows that people have stereotypical associations with languages. An intriguing question is how such stereotypes arise. Giles, Bourhis, and Davies (1975) suggested two possible answers, termed the imposed-norm hypothesis and the inherent-value hypothesis. The imposed-norm hypothesis stresses the importance of extra-linguistic factors such as social connotations and cultural norms. A language is considered attractive when its speakers are

socially privileged. The inherent-value hypothesis, on the other hand, is linguistically based, and argues that some languages are intrinsically more aesthetically pleasing due to their sound characteristics.

Most of the older language-attitude studies seem to support the imposed-norm hypothesis (Trudgill and Giles 1978). More recent studies, however, found evidence for the inherent-value hypothesis. Van Bezooijen (1996) had Dutch subjects evaluate a number of languages aesthetically. Phoneticians rated the same languages on phonetic scales. The rank order of aesthetic evaluations could almost completely be predicted by a combination of melodiousness and softness. Also, fast tempo and precise and fronted articulation were positively correlated with the aesthetic evaluations. These outcomes suggest that aesthetic evaluations may have a phonetic basis. Similar results were obtained by Gooskens, Schüppert, and Hilton (2016) via a matched-guise experiment. They made recordings of a perfect Swedish/Danish bilingual speaker and presented them, together with a number of filler languages, to Chinese students. They were asked to judge how beautiful the languages sounded. The subjects found Swedish significantly more beautiful than Danish. Since the subjects were unfamiliar with the test languages and the speaker of both languages was the same, imposed norms and speaker characteristics cannot have influenced the judgments. The differences in judgments must therefore have been caused by characteristics of the languages themselves. As with Van Bezooijen's study, this investigation provides clear evidence that inherent language characteristics play a role in aesthetic evaluations. However, it still leaves open the question of whether these attitudes also influence intelligibility, and what the direction is of the causality.

11.4.2.2 Contact and Experience

Of course, the level of intelligibility also depends on the amount of experience and contact, including formal instruction, that the listener has had with the other language. However, it has often been difficult to find a direct link. It has been assumed, for example, that the asymmetric intelligibility between Swedish and Danish, for example, in the investigation by Maurud (1976), could at least partly be explained by the fact that the listeners came from the capitals of Sweden and Denmark. As Copenhagen is located only 30 kilometres (20 miles) from the Swedish border, whereas Stockholm is located about 570 kilometres (350 miles) from the Danish border, there is a substantial geographical asymmetry in the origin of the subjects. The Danes in Maurud's investigation had more opportunities to hear and read the neighboring language than did the Swedes. Bø (1978) therefore tested the intelligibility of the neighboring language among two groups, one living inside and one living outside the border regions of Sweden and Denmark. The border region group not only had more opportunities to visit the neighboring country, but also had access to television programs in the neighboring language. The results showed that this group of subjects had fewer difficulties decoding the neighboring variety than did subjects living outside the border region, thereby indicating that a high degree of contact indeed enhances intelligibility.

11.4.2.3 Orthography

Another explanation for the asymmetric intelligibility of Swedish and Danish might be found in the relationship between the written and the spoken forms of the languages. Spoken Swedish is close to both written Swedish and written Danish, whereas spoken Danish has undergone a number of reduction processes, which are not reflected in the orthographic system. Danish pronunciation has changed more rapidly during the last century than Swedish pronunciation has. As a consequence, spoken Danish has developed

away from its written form and is therefore rather distant from both Swedish and Danish in their written forms. Danes can understand spoken Swedish better because of its close similarity to written Danish, while Swedes get less help from written Swedish when listening to spoken Danish. For example, it is likely that literate Danes confronted with the Swedish word /land/ “country” can use their orthographic knowledge to match this word to their native correspondent *land*, whereas this is not the case for Swedish listeners confronted with Danish /lan?/ because of the absence of the phoneme /d/, which is present in Swedish pronunciation as well as orthography.

Doetjes and Gooskens (2009) quantified the relationship between the spoken and written representations of Swedish and Danish in a corpus of 86 frequent cognate words, first by measuring phonetic and orthographic distances between the languages by means of the Levenshtein algorithm. As expected, the phonetic distance was larger (53%) than the orthographic distance (24%). Next, they calculated the distances again, but this time corrected the phonetic distance values for the advantage that Danes and Swedes gain from their native orthography when listening to the neighboring language. This was done by setting the segment distance to zero in cases where a phoneme could be understood from its orthographic equivalent in the native language. After correcting for orthography, the distances turned out to be smaller for the Danes (30%) than for the Swedes (46%). This indicates that Danes obtain more potential help from the orthography than Swedes do. Doetjes and Gooskens (2009) tested this hypothesis by correlating the distances with the results of a word-intelligibility experiment run using Danish participants. Distance values corrected for the influence of orthography showed higher correlations with the intelligibility scores than pure phonetic distances. The authors conclude that Danish listeners indeed seem to make use of the additional information that the orthography can provide.

This claim was made even stronger by an investigation by Schüppert (2011). She played spoken Swedish words to Danish speakers in a translation task. The words were cognates in which the pronunciation differed in one phonetic segment only (e.g., the word *mild* “mild” is pronounced /mild/ in Swedish but /mil?/ in Danish). Half of the Swedish cognates were pronounced in a way that would be consistent with the spelling of the Danish word (i.e., orthographically consistent cognates), whereas the other half were pronounced in a way that would not be consistent with the spelling of the Danish word (i.e., orthographically inconsistent cognates). Event-related brain potentials (ERPs) in the translation task were measured for these consistent and inconsistent cognates to study the participants’ online brain responses during decoding operations over the first 1000 milliseconds (ms). The data showed that ERPs in response to inconsistent words were significantly more negative than ERPs for consistent words between 750 and 900 ms after stimulus onset. Together with higher word-recognition scores for consistent items, the data provide strong evidence that online activation of L1 orthography enhances word recognition among literate speakers of Danish who are exposed to samples of spoken Swedish.

11.4.2.4 *Gestures*

Co-speech gestures (movements of the hands, face, or other parts of the body that people spontaneously produce when speaking) and mouthing (movements of the mouth) reflect important aspects of oral communication (Kita and Özyürek 2003). Experimental studies have shown that subjects who see accompanying gestures while hearing native speech pick up significantly more relevant information than do subjects who only listen to speech (Graham and Argyle 1975; Riseborough 1981). The role of co-speech gestures and mouthing in the intelligibility of closely related languages has, however, hardly been studied. Voigt and Gooskens (in preparation) tested the influence of gestures and mouthing on the

intelligibility of spoken Spanish for Italian listeners. The experiment incorporated four conditions. In the first condition, the subjects saw a video recording of the full upper body of the speaker while she retold a story. In the second condition, the subjects could only see the head of the speaker while listening to what she said. In the third condition, the head of the Spanish speaker was obscured and the subjects could only hear the story and see the gestures. The fourth condition consisted of only the audio file and a blank screen. The mean percentages of correct answers were 66.8% for condition 1 (full body and audio), 50.3% for condition 2 (head and audio), 52.7% for condition 3 (gestures and audio) and 47.2% for condition 4 (audio only). The differences between conditions 1 and 2, and between conditions 1 and 3, are statistically significant. These results confirm the hypothesis that co-gestures and mouthing facilitate the intelligibility of an unknown, yet related, language.

11.5 Desiderata for Future Research

In this chapter the role of intelligibility in the definition of dialects as opposed to languages was discussed. It was shown that the intelligibility criterion brings with it a number of problems, but it was also stated that modern research methodologies can open up useful ways of measuring intelligibility. It is difficult to determine when two language varieties are mutually intelligible to such an extent that they can be considered dialects of the same language, and when communication is so difficult that they should be considered different languages. It would be useful to set up a standard for testing the intelligibility of dialects and closely related languages. This would make it possible to compare the results of different investigations, and perhaps even to define a level that can be considered the threshold at which intelligibility is sufficient for communication.

An interesting related question that still remains unanswered is whether intelligibility is gradual, or whether it is possible to define some breakdown point at which language varieties become unintelligible to listeners. Such a critical breakdown threshold would depend upon the linguistic differences between the language varieties in question. If the number of non-cognates varies around the breakdown threshold, lexical distance will be more important than phonological/phonetic distance within the cognates. Similarly, if a language has more (or fewer) consonants relative to vowels in its phoneme inventory, the importance of vowel and consonant distance will be different from what obtains in another language. Typically, the relationship between the number and magnitude of deviations and the intelligibility of a linguistic unit is non-linear. Identification of a sound or recognition of a word remains very good for small discrepancies from the norm, but abruptly breaks down when these discrepancies become larger. Future work with different language varieties and more controlled representations of various linguistic units can hopefully yield more insight into the relative contributions of linguistic phenomena to intelligibility, and show when the limits of intelligibility have been reached.

New methods for testing the intelligibility of closely related language varieties have so far mainly generated global results. In future research we should aim to gain more detailed knowledge about the mechanisms behind the intelligibility of language varieties. Methods that have been developed by experimental linguists and psycholinguists should be exploited when setting up controlled experiments that will give us more insight into the relative importance of various linguistic and extra-linguistic factors that impact upon the intelligibility of language varieties.

By testing the intelligibility of a large number of languages differing along many dimensions we may establish the relative importance of the various dimensions. This will allow us to provide a more solid, experimentally grounded, foundation for traditional claims made by linguists about genealogical relatedness among languages.

Intelligibility between languages may also serve as the ultimate criterion to decide how structural dimensions should be weighed against each other in the computation of linguistic distance.

REFERENCES

- Best, Catherine. 1995. "A direct realist perspective on cross-language speech perception." In *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues in Cross-Language Research*, edited by Winifred Strange, 167–200. Timonium, MD: York Press.
- Best, Catherine, Gerald McRoberts, and Elizabeth Goodell. 2001. "Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system." *Journal of the Acoustical Society of America*, 109: 775–794.
- Bø, Inge. 1978. *Ungdom og Naboland: En Undersøkelse av Skolens og Fjernsynets Betydning for Nabospråksførståelsen*. Stavanger: Rogalandsforskning.
- Boets, Herman, and Georges de Schutter. 1977. "Verstaanbaarheid en appreciatie: Nederlandse dialechten uit België zoals inwoners van Duffel die ervaren." *Taal en Tongval*, 29: 156–177.
- Casad, Eugene. 1974. *Dialect Intelligibility Testing*. Norman, OK: Summer Institute of Linguistics of the University of Oklahoma.
- Delsing, Lars-Olof, and Katarina Lundin Åkesson. 2005. *Håller Språket ihop Norden? En Forskningsrapport om Ungdomars Förståelse av Danska, Svenska och Norska*. Copenhagen: Nordiska Ministerrådet.
- Doetjes, Gerard, and Charlotte Gooskens. 2009. "Skriftsprøgets rolle i den dansk-svenske talesprogsforståelse." *Språk og Stil*, 19: 105–123.
- European Charter for Regional or Minority Languages. 1992. <http://conventions.coe.int/treaty/en/Treaties/Html/148.htm> (accessed September 19, 2015).
- Giles, Howard, Richard Bourhis, and Ann Davies. 1975. "Prestige speech styles: the imposed norm and inherent value hypotheses." In *Language in Anthropology IV: Language in Many Ways*, edited by William McCormack, and Stephen Wurm, 589–596. The Hague: Mouton.
- Giles, Howard, and Nancy Niedzielski. 1998. "Italian is beautiful, German is ugly." In *Language Myths*, edited by Laurie Bauer, and Peter Trudgill, 85–93. London: Penguin.
- Gooskens, Charlotte. 2013. "Methods for measuring intelligibility of closely related language varieties." In *The Oxford Handbook of Sociolinguistics*, edited by Robert Bayley, Richard Cameron, and Ceil Lucas, 195–213. Oxford: Oxford University Press.
- Gooskens, Charlotte, Anja Schüppert, and Nanna Haug Hilton. 2016. "Is Swedish more beautiful than Danish? – A matched-guise investigation." In *Nooit het Noorden kwijt*, edited by Sara Van den Bossche and Martje Wijers, 165–182. Ghent: Academia Press.
- Gooskens, Charlotte, Wilbert Heeringa, and Karin Beijering. 2008. "Phonetic and lexical predictors of intelligibility." *International Journal of Humanities and Arts Computing*, 2(1–2): 63–81.
- Gooskens, Charlotte, and Renée van Bezooijen. 2006. "Mutual comprehensibility of written Afrikaans and Dutch: Symmetrical or asymmetrical?" *Literary and Linguistic Computing*, 23: 543–557.
- Gooskens, Charlotte, Renée van Bezooijen, and Vincent van Heuven. 2015. "Mutual intelligibility of Dutch-German cognates by children: The devil is in the detail." *Linguistics*, 53(2): 255–283.
- Graham, Jean, and Michael Argyle. 1975. "A cross-cultural study of the communication of extra-verbal meaning by gestures." *Journal of Human Movement Study*, 1: 33–39.
- Hammarström, Harald. 2005. "Review of Ethnologue: Languages of the World, 15th Edition." <http://linguistlist.org/issues/16/16-2637.html> (accessed November 10, 2015).
- Hammarström, Harald. 2008. "Counting languages in dialect continua using the criterion of mutual intelligibility." *Journal of Quantitative Linguistics*, 15(1): 34–45.
- Heeringa, Wilbert, Peter Kleiweg, Charlotte Gooskens, and John Nerbonne. 2006. "Evaluation of string distance algorithms for dialectology." In *Linguistic Distances Workshop at the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics, Sydney, July*, 2006, edited by John Nerbonne,

- and Erhard Hinrichs, 51–62. Stroudsburg, PA: The Association for Computational Linguistics.
- Hickerson, Harold, Glen Turner, and Nancy Hickerson. 1952. "Testing procedures for estimation transfer of information among Iroquois dialects and languages." *International Journal of American Linguistics*, 18: 1–8.
- Hilton, Nanna Haug, Charlotte Gooskens, and Anja Schüppert. 2013. "The influence of non-native morphosyntax on the intelligibility of a closely related language." *Lingua*, 137: 1–18.
- Hudson, Richard. 1996. *Sociolinguistics*. Cambridge: Cambridge University Press.
- Jensen, John. 1989. "On the mutual intelligibility of Spanish and Portuguese." *Hispania*, 72 (4): 848–852.
- Kita, Sotaro, and Asli Özyürek. 2003. "What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking." *Journal of Memory and Language*, 48(1): 16–32.
- Kloss, Heinz. 1967. "Abstand languages and Ausbau languages." *Anthropological Linguistics*, 9(7): 29–41.
- Lewis, M. Paul, Gary Simons, and Charles Fennig, eds. 2013. *Ethnologue: Languages of the World*, 17th ed. Dallas, TX: SIL International.
- Maurud, Øivind. 1976. *Nabospråksforståelse i Skandinavia: En Undersøkelse om Gjensidig Forståelse av Tale- og Skriftspråk i Danmark, Norge og Sverige*. Stockholm: Nordiska Rådet.
- Merriam, Clinton. 1926. "The classification and distribution of the Pit River Indian tribes of California." *The Smithsonian Institution Miscellaneous Collection*, 78 (3): 1–52.
- Nahhas, Ramzi. 2006. *The Steps of Recorded Text Testing: A Practical Guide*. Chiang Mai: Payap University.
- Nerbonne, John, and Wilbert Heeringa. 2010. "Measuring dialect differences." In *Language and Space: An International Handbook of Linguistic Variation, Theories and Methods*, edited by Peter Auer, and Jürgen Schmidt, 550–567. Berlin: de Gruyter.
- Pierce, Joe. 1952. "Dialect distance testing in Algonquian." *International Journal of American Linguistics*, 18: 208–218.
- Riseborough, Margaret. 1981. "Physiographic gestures as decoding facilitators: Three experiments exploring a neglected facet of communication." *Journal of Nonverbal Behavior*, 5: 172–183.
- Scharpff, Peter, and Vincent van Heuven. 1988. "Effects of pause insertion on the intelligibility of low quality speech." In *Proceedings 7th FASE Symposium/Proceedings Speech '88*, edited by William Ainsworth, and John Holmes, 261–8. Edinburgh: The Institute of Acoustics.
- Schüppert, Anja. 2011. *Origin of Asymmetry: Mutual Intelligibility of Spoken Danish and Swedish*. PhD dissertation, University of Groningen.
- Tamburelli, Marco. 2014. "Uncovering the 'hidden' multilingualism of Europe: An Italian case study." *Journal of Multilingual and Multicultural Development*, 35(3): 252–270.
- Tang, Chaoju, and Vincent van Heuven. 2008. "Mutual intelligibility of Chinese dialects tested functionally." In *Linguistics in the Netherlands*, edited by Marjo van Koppen, and Bert Botma, 145–156. Amsterdam: Benjamins.
- Tang, Chaoju, and Vincent van Heuven. 2009. "Mutual intelligibility of Chinese dialects experimentally tested." *Lingua*, 119(5): 709–732.
- Trudgill, Peter. 2000. *Sociolinguistics: An Introduction to Language and Society*. London: Penguin.
- Trudgill, Peter, and Howard Giles. 1978. "Sociolinguistics and linguistic value judgments: Correctness, adequacy, and aesthetics." In *Functional Studies in Language and Literature*, edited by Frank Coppieters, and Didier Goyvaerts, 167–190. *Functional Studies in Language and Literature*. Gent: Story-Scientia.
- Van Bezooijen, Renée. 1996. "Aesthetic evaluation of Dutch: Comparison across dialects, accents, and languages." In *Handbook of Perceptual Dialectology*, Vol. 2, edited by Daniel Long, and Dennis Preston, 13–31. Amsterdam: Benjamins.
- Van Bezooijen, Renée, and Charlotte Gooskens. 2007. "Interlingual text comprehension: Linguistic and extralinguistic determinants." In *Receptive Multilingualism and Intercultural Communication: Linguistic Analyses, Language Policies and Didactic Concepts*, edited by Jan ten Thije, and Ludger Zeveaert, 249–264. Amsterdam: Benjamins.
- Van Bezooijen, Renée, and Rob van den Berg. 1999. "Word intelligibility of language varieties in the Netherlands and Flanders under minimal conditions." In *Linguistics in the Netherlands*, 16, edited by Renée van Bezooijen, and René Kager, 1–12. Amsterdam: Benjamins.
- Van Heuven, Vincent. 2008. "Making sense of strange sounds: (Mutual) intelligibility of related language varieties – A review." *International Journal of Humanities and Arts Computing*, 2(1–2): 39–62.
- Vitevitch, Michael, and Eva Rodríguez. 2005. "Neighborhood density effects in spoken word

- recognition in Spanish." *Journal of Multilingual Communication Disorders*, 3: 64–73.
- Voegelin, Carl, and Erminie Voegelin. 1946. "Linguistic considerations of Northeastern North America." In *Man in North Eastern North America*, edited by Frederick Johnson, 178–94. Andover, MA: Peabody Foundation.
- Voegelin, Carl, and Zellig Harris. 1951. "Methods for determining intelligibility among dialects of natural languages." *Proceedings of the American Philosophical Society* 95: 322–329.
- Voigt, Stefanie, and Charlotte Gooskens. In preparation. "Co-speech gestures and their influence on the intelligibility of Spanish for Italian listeners."
- Weijnen, Antonius. 1941. *De Nederlandse Dialecten*. Groningen: Noordhoff.
- Wieling, Martijn, Eliza Margaretha, and John Nerbonne. 2012. "Inducing a measure of phonetic similarity from pronunciation variation." *Journal of Phonetics*, 40(2): 307–314.
- Wiener, F., and G. Miller. 1946. "Some characteristics of human speech." *Transmission and Reception of Sounds under Combat Conditions: Summary Technical Report of Division 17, National Defence Research Committee*, 58–68. Washington, DC: National Defence Research Committee.
- Wolff, Hans. 1959. "Intelligibility and inter-ethnic attitudes." *Anthropological Linguistics* 1: 34–41.

12 Applied Dialectology: Dialect Coaching, Dialect Reduction, and Forensic Phonetics

DOMINIC WATT

12.1 Introduction

This chapter focuses on three areas of applied dialectology for which detailed information about dialect variation is crucial. They are broad areas, and it will only be possible to touch on selected aspects of each. I focus principally on the British context, and on matters of pronunciation. The term “dialect” will be used rather loosely to refer to accent (i.e., phonetics/phonology only) as well as in the strict sense, that is, a language subvariety defined solely by grammatical and lexical properties. In North American dialectology and sociolinguistics “dialect” and “accent” are often used interchangeably, but in the British tradition a more rigid distinction is generally drawn.

The applications discussed below are the preoccupation of people working in domains ranging from the arts and entertainment industries, through education and business, to law enforcement, the judicial system, and national security. We will consider aspects of these activities in turn, bearing in mind that in each case the availability of up-to-date descriptive dialect resources and training materials is vital. Another theme running through this chapter is *imitation*. The goal of dialect coaching is an adequately accurate rendition of a dialect other than the talker’s own, through imitation of native speakers of that dialect. Dialect reduction shares with dialect coaching, of which it is a subtype, the aim of helping a speaker to acquire a new variety. In forensic phonetic casework, practitioners will regularly encounter recordings of speech in which it is suspected that the speaker has used dialect imitation as a disguise, possibly so as to intimidate or coerce the hearer more effectively. Occasionally, cases arise of imitation in the form of impersonation of another individual, as a way to obtain money deceptively or to create a false alibi.

We look first, however, at more benign forms of vocal imitation that are used for artistic expression or entertainment.

12.2 Dialect Coaching¹

Dialect coaching is a form of voice training geared toward helping trainees to develop the ability to talk in dialects other than their own. It will be assumed that the approach taken by professional dialect coaches will prioritize mastery of the pronunciation of a new variety, with less of a focus on its other properties. Acquiring the relevant grammatical or lexical

features may be important, but grammatical constructions that are characteristic of particular dialects are liable to occur much less often than variety-specific pronunciations. The latter are unavoidable if the trainee has to use the freshly acquired variety even briefly. Learning dialect words, moreover, is a matter of rote memorization that is arguably quite different from mastering a new accent.

Dialect coaches—who sometimes also advertise themselves as “dialogue coaches”—work closely with actors, providing explicit instruction in how to use a new dialect such that the performance is both natural-sounding and authentic. These two qualities are not interchangeable: making too much of an effort to exactly reproduce a non-native dialect may divert the audience’s attention away from what is being said to *how* it is being said. Thus, it might be in the actor’s interests to focus on maintaining the illusion of “naturalness” at the expense of complete accuracy. As we will see below, listeners may be very tolerant of deviations from the expected pronunciation patterns of a dialect if they have no reason to doubt the speaker’s authenticity, or are unfamiliar with that individual’s normal dialect.

Where face-to-face coaching is impractical or unaffordable, actors may opt for self-study materials in the form of printed manuals, sound recordings (e.g., Strong and Dyer 2007), video recordings, and most recently smartphone apps (e.g., www.theaccentkit.com). Pronunciation and elocution manuals for the socially aspirant have been available in Britain for several centuries, although only relatively recently have these guides offered explicit advice on how to imitate social and regional accents for dramatic effect, rather than suggesting techniques to purge one’s speech of provincialisms and other habits perceived to be vulgar or defective (e.g., Milroy and Milroy 2012). Some of the influential early elocutionists, notably Thomas Sheridan and John “Elocution” Walker, had also been actors, helping to cement in the public mind an association between the theatre and vocal accomplishment (Beal 2004; Mugglestone 2007). As in a number of other professions—broadcasting, law, teaching, singing, politics, the clergy—voice training is vitally important to actors, for whom the voice is their stock-in-trade. Such training forms part of the curriculum at all drama schools, and blocks of the voice and speech courses for trainee actors are devoted to intensive dialect and accent work. A variety of authoritative publications have emerged from this tradition, and not surprisingly their authors have enthusiastically exploited the multimedia possibilities afforded by cheap sound-reproduction technology, initially on vinyl and later on cassette tape and CD/DVDs. Most recently, professional dialect coaches have started to realize the potential of the internet for hosting tutorial videos and delivering live coaching sessions (see especially www.paulmeier.com). The innumerable online videos of amateur dialect enthusiasts touting their skills are inconsistent in quality, but it testifies to the perennial social attractiveness of the ability to mimic the vocal habits of other individuals and groups that the number of hits, likes and other measures of approbation that videos receive can run into the millions (for example, at the time of writing, Truseneye92’s *The English Language in 24 Accents* had garnered over 26 million hits; of the approximately half-million ratings awarded, nearly 98% were positive).²

Popular volumes for actors seeking to expand their repertoire of dialects (e.g., Machlin 1992; Herman and Herman 1997; Turner and Morrison 2000; Blumenfeld 2013) often concentrate first on general principles of voice and speech production, and basic vocal techniques, before describing the features of individual varieties. Typically, speech production is broken into stages that will seem familiar and logical to phoneticians: firstly, matters relating to initiation (breathing, posture), then phonation and prosody (projection, tone, “tune”), and finally supralaryngeal articulation (pronunciations of individual speech sounds).

The approach developed by Edda Sharpe and Jan Haydn Rowles, two prominent coaches based in London, serves as a good example of contemporary practice in dialect coaching for the acting and voiceover professions. The method illustrated in their manual *How to Do Accents* and on its accompanying website (www.howtodoaccents.com) stipulates that readers

must first learn the rudiments of speech anatomy and articulation. This knowledge can then be related to how individual accents are broadly defined by the vocal settings typically used by their native speakers. Only then should the learner attempt to build upon these fundamental layers by focusing on individual sounds. The aim is to develop “muscle memory” through constant practice, such that with sufficient practice the learner will be able to produce the accent convincingly and reflexively, much as one learns to coordinate one’s unconscious movements when driving a car. Before the foundation-laying stage can begin, anatomical terms must be committed to memory, and the importance of learning the symbols of the International Phonetic Alphabet is stressed. As Sharpe and Rowles observe, many dialect manuals do not advocate learning phonetic terminology and notation because it is seen to be offputtingly tedious, or unnecessary for actors possessing “a good ear.” However, Sharpe and Rowles argue strongly for grasping this rote-learning nettle while considerably trying to minimize technical details.

Whether there is comparable phonetic substance to the classification scheme Sharpe and Rowles then propose is harder to judge. In “The Foundations” they present “four essential elements” (2009, 33) called “The Setting,” “The Zone,” “The Tone,” and “The Direction.” “The Setting” refers to articulatory setting, that is, the overall configuration of the facial and oral muscles. The learner is encouraged to focus on how the positioning of the cheeks, lips, jaw, tongue, and soft palate varies between accents. For Yorkshire accents, Sharpe and Rowles argue, the cheeks are loose, the lips slack, the jaw dropped, the tongue heavy and flat, and the soft palate high. In Scottish English, the tongue is “rolled forward and gripped,” whereas in London the cheeks are just “held” (2009, 41).

“The Zone” also refers to the accent’s “placement” in the mouth. There is good empirical evidence of the association of habitual articulatory postures with certain accents of English (e.g., Knowles 1973), but Sharpe and Rowles apparently mean something different. They talk of a “resonant focal point” (2009, 35) toward which the learner should attempt to direct his/her voice. If the desired zone is the hard palate, we must “[a]im the voice straight up into the roof of the mouth. Feel the vibrations driving up onto the hard palate and *through the bone*” (2009, 37). Whether such directives will reliably result in the intended auditory effect is unknown, though if learners implement a similar articulatory strategy each time they imagine (say) aiming the voice at the palate, it may help them produce the target accent more consistently even if what the articulators are actually doing bears scant relation to what the speaker believes is happening. Under such circumstances, the *How to Do Accents* CD may prove more useful than the text, as the learner can proceed by trying to imitate the samples provided.

The third foundational element, “The Tone,” is “the balance of tonal frequencies” or, more simply, “the *tone* of the voice” (2009, 38). It is likened to the drone of bagpipes, a sustained note that underlies the voice’s melody, and we are cited a variety of adjectives used by laypeople, and sometimes also by phoneticians and voice specialists, to describe vocal timber (“brassy,” “throaty,” “nasal,” “plummy,” “hoarse,” etc.). These subjective labels usefully capture the overall auditory impression created by the combination of a host of acoustic properties (e.g., Köster *et al.* 2007), but again it must be assumed that listeners can agree upon what the terms denote, if they are to be of practical value. “Plumminess” for one person may be manifested by a markedly velarized tongue body, minimal jaw lowering, a spread lip posture and low vocal pitch, for instance, whereas for another it might correlate with pharyngeal resonance/larynx lowering, a dentalized setting, and a wide pitch range. Alternatively, to observers from certain parts of the English-speaking world the perception of a “plummy” voice might rest more on segmental features that are associated with being middle-class or educated, rather than supra-segmental properties as such. It is hard to escape the conclusion that achieving the desired effect is more feasible if the learner has recordings to imitate, inviting the question of whether the accompanying text hinders rather than accelerates the learner’s progress.

It would be difficult to disentangle the Setting, Zone and Tone, since all three concern long-domain phonatory and articulatory properties of the voice. With some effort, however, one can see how each of them might relate to consistently observable attributes of a talker's speech. The fourth element—"The Direction"—is more perplexing. Here, Sharpe and Rowles assert that the way the sound leaves the mouth varies systematically across different accents. In London English, we learn, the sound travels forward onto the hard palate and teeth, whereas in Liverpool speech it "spills sideways." In Newcastle it "lurches back and forth" but in Standard English (RP, i.e., Received Pronunciation), it travels "forward and out like a wave" (2009, 42). Phonetic substantiation for these peculiar claims is not offered.

There is also some predictability to Sharpe and Rowles's assertion that in RP, by contrast with non-standard accents, the sound wave (perhaps airstream is meant) exits the talker's mouth cleanly, smoothly and directly. Statements in this vein are common in actors' dialect manuals, the implication being that mastery of RP is synonymous with acquiring "good diction." Similarly, we often encounter the (articulatorily baseless) claim that RP is optimally suited for the theatre and public speaking because of its "forward placement." An example is Rodenburg (1998, 107), who says, "One of the positive characteristics of RP is that all sounds are clearly placed as far forward as possible [...] These positions are athletic, energized and useful as a work-out." Such blurring of the distinction between accent and the clarity and "projectability" of spoken English is evident in many manuals (e.g., McCallion 1989; Morrison 2001), but not all sources justify a preference for RP for this reason, or indeed any other. It seems simply to be assumed that mastery of RP is incumbent upon those seeking careers in the dramatic arts. Also implicit is a tendency to treat being a native speaker of a non-standard accent as a hurdle to be overcome. Rodenburg's prefatory declaration "What I do not classify as debilitating habits are native or regional accents or colloquial speech patterns" (1998, 10) implies that many of her fellow coaches, and perhaps also her readers, would view such linguistic traits as a handicap.

In reality, career actors are much more likely to need competence in RP than in any other accent. It is also hard to dispute the claim that the accent enjoys currency and prestige in a way that other British accents do not. The reverence accorded to RP is magnified through its association with "the great writers," in particular Shakespeare, who is quoted disproportionately often in the voice manuals. It is doubtlessly useful for trainee actors to practice reciting Shakespearean texts, given the playwright's unassailable centrality in the English literary canon, but it seems ironic that it is thought so appropriate to teach correct theatrical diction using plays and poems written centuries before RP existed (see further Crystal 2005; Milroy and Milroy 2012).

Nevertheless, there is evidence that RP is no longer as exalted as it once was (Coupland and Bishop 2007; Trudgill 2008), and some have even complained that it is now "too posh" for British TV and radio (Odene 2012). Acknowledging this potential stigma, recognising that RP has undergone significant changes over its lifetime, and in reaction to requests from directors and agents, Sharpe and Rowles have repackaged contemporary RP as the "Neutral Standard English Accent" (NSEA) in their *How to Do Standard English Accents* (2011). They contend that NSEA, an accent "still considered to be an invaluable skill for an actor," is "free from indicators of gender, race, age, class or region" (2011, 11). Sustaining this claim may, however, be difficult given the overwhelming sociolinguistic evidence showing precisely the opposite. Standard accents of English are still inseparably associated with the educated middle and upper classes in English-speaking countries (e.g., Ashton and Shepherd 2012; Lippi-Green 2012), as well as with white "Anglo" ethnicity and to some extent with region (RP with south-eastern England, e.g., Trudgill 2008). Distancing NSEA from its elitist origins might serve to make learning it more appealing to speakers of other accents, but portraying the accent as "a neutral standard" could also be argued to be a substitute for the ideology of correctness that was used to rationalize and justify the privileged status of RP

in earlier periods. Equally, it can be contended that bestowing neutrality upon NSEA makes the accent easier to promote with respect to "clarity." Consumers of products like those discussed above might not object to the idea that they are learning how to speak clearly if they believe that the phonological model they are attempting to approximate is free of the taints of elitism and social snobbery. This, indeed, is a marketing ploy that has been followed by the publishers of materials designed for dialect reduction, which we examine in the following section.

12.3 Dialect Reduction

The dialect coaching tradition in the Anglophone world is rooted in the drive toward greater standardization in English speech that took place in the eighteenth and nineteenth centuries. This movement placed a heavy emphasis in both public and private education on the teaching of "approved" (or "received") pronunciations of words, and a surge in the production of pronunciation dictionaries and elocution handbooks for members of the aspirant middle classes. Among the latter, it became imperative for those wishing to consolidate or improve their social standing to speak in a manner considered "proper," "refined," and "elegant." For some this might have meant making small adjustments to their existing accents, but for others it might have been a question of learning what was effectively a foreign accent. As Beal (2004) points out, the authors of these manuals were themselves often from peripheral areas of the British Isles (Thomas Sheridan was Irish and Thomas Spence from Newcastle, for instance), implying that the necessity of speaking with an accent like that of the affluent south-east of England was sensed more acutely by authors from the far-flung provinces than it was by those closer to the metropolis. London vernacular speech was nonetheless frequently the target of criticism, it being thought that those living alongside speakers of the best English had no excuse for using linguistic vulgarisms (Beal 2004, 173). Anyone acquainted with Shaw's play *Pygmalion* (1916) will be familiar with features of London speech that were particularly stigmatized around the beginning of the twentieth century. Examples are /h/-dropping in content words like *hat* or *house*—a habit Rippmann (1909, 51–52) scorned as defective, careless and vulgar—or MOUTH-monophthonging and the use of the diphthongs [æɪ] and [ʌʊ] for FACE and GOAT respectively (see e.g., Behnke (1897, 10, 140ff.), who lambastes these and other "wrong vowel sounds" of "uncultured speech" as "detestable," "lazy," "hideously ugly," even "evil"). Pronunciation manuals like these were aimed as much at native speakers of English as they were at non-native speakers, and voicing these scathingly judgmental views of the aesthetic properties of non-standard English accents and the questionable moral fiber of their speakers can only have helped to create larger markets for their books. Rippmann's 1909 guide was expressly for use in teaching training colleges, while various regional editions of his *English Sounds* (1911), designed to help school pupils avoid graduating with an "unpleasant voice" that would "mutilate the language" (Rippmann and Robson 1913, iv), were used throughout the British Empire. The quotes above come from the edition adapted specifically for Scottish schools, with its content tailored in a relatively forgiving way toward the standard Scottish English of the day. It is nevertheless made abundantly clear that RP was the pinnacle of linguistic achievement by virtue of its correctness, elegance, expressiveness, and clarity, and was essentially a requirement for anyone hoping to enter a profession involving public speaking.

The linguistic insecurity that fed the appetite for pronunciation manuals and their endorsement by the education authorities was heightened further by the linguistic policies of the BBC. Indeed, according to the *Oxford English Dictionary* the term "BBC English" as a synonym for RP was already in use in 1928, just six years after the corporation's founding. From 1926, pronunciation standards were set out explicitly by the Advisory Committee on

Spoken English, of which the phoneticians Daniel Jones and Arthur Lloyd James were members. That tradition continues today in the form of the BBC Pronunciation Research Unit (Sangster 2008), which advises “accurately and consistently,” but not overly prescriptively, on the pronunciation of words and names in English and other languages. The Unit’s disavowal of linguistic dogmatism does not, however, dissuade the Unit’s admirers from treating its recommendations and those of sources such as the *Oxford English Dictionary* or the *Cambridge English Pronouncing Dictionary* as uncontestedly authoritative.

While the dialect reduction/correction materials targeted at linguistically insecure native speakers formerly tended to stress the advantages that better diction would bring in terms of social mobility, these days the key selling point is the “flexibility” and “empowerment” conferred by a more standard mode of speech. Rather than seeking to eradicate the reader’s normal accent by replacing it with a new one, modern guides (e.g., Meier 2011, 2012a,b) prefer to encourage readers to expand their linguistic repertoires by acquiring an accent closer to Sharpe and Rowles’s “neutral standard.” Nonetheless, the titles and impressive sales figures of the *Get Rid of Your Accent* series (James and Smith 2006, 2011, 2012) suggest a keen appetite for self-help manuals taking a more traditional approach to training users to replace their habitual accents with RP. Like their Victorian antecedents, James and Smith’s volumes make no distinction between RP (“simply a neutral pronunciation of educated Southern English... sometimes called Standard English”; 2006, 2) and good/correct/proper pronunciation. RP, they say, “allows you to become a pleasant communicator; is a good basis for public speaking; will enable you to enjoy speaking more; gives you confidence... open[ing] up for you all sorts of opportunities” (*ibid.*). Their books are not just for non-Anglophone learners, either: James and Smith (2006, 11) warn that “it is easy to revert to your original... regional accent if you do not continue making an effort to pronounce correctly,” and in one volume identify the following groups of native English speakers under the banner “Who this book is for”:

- Pronunciation and speech teachers
- Actors with non-RP accents who wish to widen their range
- Hollywood actors who need to develop a British accent
- Professionals for whom a high standard of English and clarity of speech are important
- Public speakers.

(James and Smith 2011, 9)

The website of Smith and James’s company *Business and Technical Communication Services*, which aims to “empower people by making their speech clear and interesting,” features training courses, a list of high-powered industrial and governmental clients, and smart-phone apps to “neutralise your accent within 1-3 months [and] improve your job prospects.” Similar promises appear on the *Executive Voice* website (www.executivevoice.co.uk), which links to a series of podcasts called *Superstar Communicator*. In one, “Speaking Clearly,” Susan Wright offers advice on “diction” and “modifying your accent.” For both she makes essentially the same point: words should always be pronounced clearly, whatever accent one uses. However, under “diction,” Wright focuses on RP/BBC English, and although RP is not explicitly promoted it appears that improving one’s diction equates to making one’s non-standard accent more RP-like.

Self-help materials of this kind generally gloss over what is meant by “diction,” “clarity,” “accent improvement,” and “correct pronunciation”; the standard language ideology they endorse is, it appears, assumed to be shared by the consumer. Learners might justifiably feel puzzled to be told at one moment that there is nothing intrinsically superior about RP, and at the next to hear that unless they master something effectively identical to this neutral,

optimally clear, classless, genderless accent their regionally accented speech will remain only partially intelligible, their opinions will not sound convincingly or confidently expressed, their employment prospects will be limited, and their competence—even their personal integrity—will forever be in doubt (see further, Thomson 2012).

The extent to which adult native speakers *can* eliminate or disguise non-standard accents tends to be skirted around in the commercial sources like those described above. The technical literature, however, reports a considerable amount of experimental work on accent disguise, principally from the psycholinguistic and forensic phonetic standpoints. We turn to the latter of these in the final section of this chapter.

12.4 Forensic Phonetics

The foregoing discussion has dwelt on accent imitation in entertainment, business and marketing, teaching, and self-improvement, but an area that should be included under the banner of applied dialect studies is “forensic dialectology”: the application of dialectology in the forensic arena. Forensic speech scientists specialize in the analysis of speech recordings that are of potential evidential importance in the investigation and prosecution of crimes. For example, a telephoned ransom demand may be delivered by a speaker who is attempting to disguise his normal voice. Disguise may be achieved by modifying voice quality (use of whisper or falsetto register, say), using an electronic voice changing device, or just holding an object in the mouth (see Eriksson 2010). Occasionally, speakers will disguise their voices by adopting an accent/dialect other than their own. It need not sound authentic; it may suffice simply that the talker’s normal voice is masked (Markham 1999). The consensus among speech experts is that we cannot (yet) identify any indelible acoustic hallmarks in people’s voices such that individuals can be identified uniquely irrespective of vocal changes brought about by disguise, shouting, illness, and so on (Foulkes and French 2012).

The talker may strive to do a convincing job when feigning the target accent, of course. An example case from 2007 involved an individual who sent a message on audio tape to a company director, threatening his family and employees with physical harm if a large sum of money were not handed over. The talker purported to be a debt collector representing a notoriously uncompromising Irish paramilitary organization, and spoke in what sounds like a Northern Irish accent. The vowels, consonants, and prosody of Northern Irish English (Corrigan 2010) are mimicked sufficiently well that a first-pass listening does not lead one to question the accent’s authenticity. However, more focused analysis reveals numerous inconsistencies, as well as breaks in the signal, which imply that the talker was dissatisfied with his performance and so re-recorded the faulty portions. Accent imitation can be difficult to sustain, and it is perhaps not surprising that as the recording progresses the “inauthentic” pronunciations start to accumulate.

Treating the case firstly as an exercise in speaker profiling—that is, compiling linguistic evidence that might point to an unknown talker’s geographical and social origins (e.g., French, Harrison, and Windsor Lewis 2007; Köster *et al.* 2012)—it could be asserted with some confidence that the speaker was probably not a native speaker of Northern Irish English. This was of little help in determining what his usual accent might be, although occasional lapses in the use of post-vocalic /r/, a defining feature of Northern Irish English, suggested that he might have been from somewhere in England or Wales. Because the police had also provided a recording of a man who had been arrested and interviewed in connection with the offence, the case also involved speaker comparison, whereby two or more recorded speech samples are compared with a view to estimating the relative support that the evidence gives to competing hypotheses (here, “same speaker, or different speakers?”; Gold

and Hughes 2014). The suspect spoke in the police interview with a consistently /r/-ful Scottish accent, and his voice quality differed substantially from that of the offender.

Detailed acoustic measurements of selected segmental features—chiefly stressed vowels—and fundamental frequency (vocal pitch) were made to accompany our auditory observations, but the acoustic data were of limited assistance. Nothing in them argued compellingly in favor of the same-speaker hypothesis, but it was also true that in respect of vowel pronunciations and pitch the differences between the samples fell in the ranges within which an adult male speaker's speech productions might vary if he were disguising his voice. The evidence, we concluded, did not lend adequate support to the view that the offender and the suspect were the same person, but neither could we discount that possibility. Adopting a different dialect was a very potent form of disguise here, even if the criminal had failed to perform it in a wholly authentic way.

We should remember, of course, that accent "authenticity" is something of an idealization. After all, it is usually agreed by dialectologists and sociolinguists that the full sets of features that define dialect X are not necessarily possessed by any individual speaker. In this sense, accent and dialect systems are abstractions across the community of speakers of the variety in question. Even the most conservative variety differs somewhat from speaker to speaker, in line with factors such as the speaker's age, gender, education, network of social contacts, or occupation. It is part of our sociolinguistic competence to recognize this fact: we know that people's accents can change if they move to a new area (Chambers 1992; Nyčz 2013), or if they live and/or frequently socialize with speakers of other accents (Evans and Iverson 2007; Pardo *et al.* 2012). It is worth reflecting on the nature of the knowledge that underpins the ability to spot an "accent fake." How and when do we acquire it? How much previous exposure to a variety do we need to judge another talker's authenticity accurately? What role is played by stereotypes?

We cannot properly engage here with these intriguing questions, but there is no doubt that as listeners we are sometimes surprisingly tolerant of deviations from the accent/dialect "models" that we have internalized. This is possibly because we learn when interacting with speakers that our expectations are often confounded. Given that these expectations are often based on outdated stereotypes (Preston 1989; Campbell-Kibler 2009), this is perhaps not surprising. Nonetheless, in hindsight it is remarkable that the individual discussed below could for so long pass himself off as a speaker of a variety of English other than his native one.

The man in question was arrested in 2005 on suspicion of attempting to enter the United Kingdom on a false passport. He spoke with what sounded like an RP accent, and claimed to be the Earl of Buckingham. He said that he had been born and educated in London, but had spent many years working in various European countries and had traveled extensively elsewhere. Checks, however, showed that his given name (Christopher Buckingham) and the date of birth shown on his papers were in fact those of an infant who had died in the 1960s, and whose birth certificate had been used to obtain the false passport. The detainee had been posing as Lord Buckingham since 1983 despite the fact that the title had lapsed in the eighteenth century. The man was imprisoned for traveling on a false passport, but still refused to divulge his real identity. Meanwhile, a worldwide appeal was launched to find someone who might recognize the impostor from photographs, and a recording of the 2005 interview was sent to several forensic speech experts in an effort to establish where "Lord Buckingham" might really be from.

As with the extortion case discussed earlier, the initial impression formed of the impostor's speech by many expert listeners was that his accent was genuine. However, on subsequent passes through the recording it was noticed that there were features present that are not normally associated with RP: for instance, the speaker quite frequently used post-vocalic /r/ where it would typically not be pronounced by RP speakers, he yod-drops (elides the /j/) in *news* and *institutions*, and flaps intervocalic /t/ in, for example, *what I,*

that I. It is not difficult to postulate plausible reasons for these habits, however. "Lord Buckingham" had been working for international companies in Switzerland for many years, and so doubtlessly had had plentiful opportunities to mix with speakers of varieties of English in which the non-RP-like features are found, and had perhaps adopted their patterns via linguistic convergence. He was, furthermore, married to a Canadian woman. Other features led analysts to hypothesize that he was originally Australian or South African, or perhaps not a native speaker of English at all. Some speculated that he might be an East German spy living under an assumed identity. The eventual conclusion that the experts drew based on the phonetic evidence was that "Lord Buckingham" was probably from the United States.

They were right: the family of Charles Stopford, who had disappeared from his Florida home in 1983, recognized him from photographs circulated in the media. Stopford was raised in the US but as a young man became an obsessive Anglophile who affected an English accent. He fled abroad after having caused an explosion while serving in the US Navy. Though he was deported to the US following his stretch in prison, Stopford later returned to Switzerland, where he allegedly continued to live under several bogus identities (Falconer 2009).

As an example of how speaker profiling is done the Stopford case is unusual, in that the speaker in the interview recording was not "unknown" *per se*, and unlike many suspects in criminal investigations he willingly cooperated with the police. Stopford's subterfuge was lent credibility by his accent, but it is not normally a crime to disguise one's voice, unless a crime depending on that deceit (e.g., impersonation to access a bank account) is thereby perpetrated. Indeed, as we have already seen, accent modification to conceal one's geographical origins or to persuade others that one possesses qualities one does not actually possess—a high level of education or an upper-class upbringing, say—is positively encouraged in some quarters. If a speaker's assumed accent is accurate enough to dupe native speakers into believing it is genuine, we might ask in what sense it is inauthentic. Presumably, being able to use a feigned accent undetected involves acquiring and mobilising effectively the same sociophonetic knowledge as that used by individuals who grew up speaking with that accent.

The question of when during the speaker's life an accent, dialect, or language was acquired is far from trivial in the context of what has become known as *language analysis for the determination of origin* (LADO; see Zwaan, Verrips, and Muysken 2010; Wilson and Foulkes 2014). The purpose of LADO is to help governments to decide whether to grant asylum seekers residence in the country in which they are claiming refuge. To try to distinguish economic migrants posing as asylum seekers from genuine claimants who face persecution, imprisonment, or even death if they are returned to the country they have fled from, the immigration authorities will interview claimants on a variety of topics, including matters relating to language competence. Of particular interest to LADO specialists are "languages of socialization"—those that asylum seekers learned to speak in early life—as these are thought to provide the most reliable indications of geographical origins (Patrick 2012). Verifying the claimant's story involves checking that he or she is familiar with basic terms (e.g., numbers, or parts of the body) in the language or dialect in question. Given the enormous linguistic diversity of some of the areas of the world from which people flee (sub-Saharan Africa and the Middle East are particularly well-represented among claimants seeking entry to European countries) it is therefore vital that detailed and current information on the languages and dialects of those areas be available to the LADO agencies. Much debate surrounds the value of involving native speakers of the varieties in question, as although these individuals may not have had a significant amount of formal linguistic training they may nevertheless possess knowledge that could help to resolve whether the asylum claim is genuine (Cambier-Langeveld 2010; Nolan 2012). Dialectological scholarship has obvious

applications in LADO, although regrettably for many regions there exist no detailed or up-to-date dialect surveys, and because of the political situation in some of these areas it may be too dangerous to undertake fieldwork to gather the necessary data.

We have so far touched on only a subset of the areas of forensic linguistics and phonetics for which high-quality dialectological data are of pivotal importance. Speaker profiling, speaker comparison, and LADO all depend upon the availability of information and expertise in the domains of dialect variation and sociolinguistics, if they are to meet acceptable standards of scientific rigour, but this is also true of cases which involve transcription of difficult or disputed content in speech recordings (French 1990; Fraser and Stevenson 2014), constructing voice parades (Nolan and Grabe 1996), or evaluating the reliability of ear-witness testimony (Yarmey 2012).

In the first of these scenarios, the transcriber may be unable to identify the word(s) being spoken unless reference materials or the knowledge of a native speaker are available to help disambiguate the utterance. Putting together a voice parade, the speech equivalent of a visual identity parade, necessitates obtaining recordings of foil voices to play to the witness alongside a sample of the voice of the suspect. They must be constructed so as to strike the right balance in terms of the similarity of the voices in the parade: if these are too dissimilar it may disadvantage the suspect, whose voice may stand out from the others in an unfair way, but if the samples are too similar the witness may erroneously identify a foil speaker. Knowledge of the parameters of variation in human voices that make them resemble or differ from one another is of prime importance, which is why the McFarlane Guidelines (Nolan, 2003) for the construction of voice parades in the UK specify that the task should be carried out by appropriately qualified linguists rather than police officers. Research on the perception of dialects and accents is equally crucial if we are to get closer to a full understanding of how trained and untrained listeners attune to the features of different varieties, and store these in memory for later retrieval.

Taking account of dialect and accent variation may also help to ensure that court proceedings are conducted fairly and objectively. Linguistic prejudices can sometimes run deep in otherwise fair-minded people, including highly educated and experienced legal professionals, and their views may be shared by jurors who are easily influenced by positive and negative stereotypes associated with certain subvarieties (Lippi-Green 2012). Dixon, Mahoney, and Cocks's (2002) matched-guise study showed systematic accent bias among their participants, who rated a talker for guilt after listening to him answer questions in a mock police interview. The "suspect" spoke in either RP or the stigmatized accent of Birmingham, UK. Listeners who heard the Birmingham guise assigned significantly higher guilt ratings. In a follow-up study, Dixon and Mahoney (2004) found that the Birmingham-accented suspect was perceived to be more "typically criminal" and more prone to be re-accused of wrongdoing than the (same) talker speaking RP. They conclude that "raters treat accent as criminally diagnostic" (2004, 70), which ought to concern anyone interested in the fairness and transparency of the criminal justice process.

The body of dialect resources available to forensic speech and language experts grows apace. Reference books, dictionaries, atlases, and online tools based upon data from the major national surveys have great potential to assist in speaker profiling tasks. For the purposes of abstracting population statistics for varieties of British English the availability of large speech corpora will give forensic speaker comparison a quantitative grounding that it previously lacked (examples are the *Intonational Variability in English* (IViE) corpus (Grabe 2004), *Accents of the British Isles* (ABI-1/-2; www.thespeechchark.com/abi-1-page.html), the *Dynamic Variability in Speech* (DyViS) database (Nolan *et al.* 2009), and *York Variation in Speech* (YorViS; McDougall 2014)).

The need for representative background population data is becoming acute as the field starts to embrace automated methods (Foulkes and French 2012). Automatic speaker

recognition (ASR) systems like *Nuance Forensics* (www.nuance.co.uk) work by computing an estimate of the similarity of the questioned and the known samples, and then assessing their typicality in the context of the broader population. Obtaining adequate numbers of reference recordings for individual accents is a serious obstacle, given how expensive and time-consuming the recordings can be to collect, and their potentially short shelf-life as accents change over time (Gold and Hughes 2014). In the UK, large-scale investment in collating such resources is urgently needed if practitioners are to meet the ever more stringent demands imposed by the judiciary.

12.5 Conclusions

Dialectology lost favor among academic linguists in the late twentieth century, becoming viewed as a trivial, outmoded preoccupation that had been overtaken by developments in modern linguistics. The sneering analogies that linguists in other disciplines drew between dialectological surveys and harmless but ultimately futile pursuits like stamp collecting will probably be familiar to older readers. It should be clear from the foregoing discussion, however, that dialectological research is vital for multiple applications, some of them very non-trivial indeed. The list of applications explored above is not exhaustive: we have not considered the clinical sphere (Watt 2012), or fields such as speech synthesis (Loots and Niesler 2011), speech recognition (Leemann and Kolly 2013), computer game and audio book design (Ensslin 2012), marketing and advertising (Morales, Scott, and Yorkston 2012), film and television production (Rittmayer 2009), translation/interpreting (Mazrui 2012), or education (Snell 2013).

One could reasonably argue that dialectological data yield a much better practical return on the investment needed to collect them than does the output of other sorts of linguistic inquiry. As the chapters in this volume show, the range of data collection, analysis, and presentation techniques developed for the purposes of applied dialect study show that the field is anything but stagnant or moribund. The rehabilitation of dialectology—a branch of linguistics deserving of respect as a rigorous and innovative subfield capable of making important contributions to our understanding of the structure, acquisition, and use of language—is fully underway. The indispensability of dialect data in spheres of activity beyond the academy lends additional momentum to this process.

NOTES

- 1 I am grateful to the dialect/dialogue coaches Brendan Gunn (www.brendangunn.co.uk), Andrew Jack (www.andrewjack.com) and Paul Meier (www.paulmeier.com) for insightful discussions of their working practices. Any misrepresentations of the field are entirely my own responsibility.
- 2 https://www.youtube.com/watch?v=dABo_DCIdpM (accessed 18 November 2016).

REFERENCES

- Ashton, Helen, and Sarah Shepherd. 2012. *Work on Your Accent: Clearer Pronunciation for Better Communication*. London: Harper Collins.
- Beal, Joan. 2004. *English in Modern Times: 1700–1945*. London: Hodder.
- Behnke, Kate. 1897. *The Speaking Voice: Its Development and Preservation*, 5th ed. London: Curwen.

- Blumenfeld, Robert. 2013. *Teach Yourself Accents – A Handbook for Young Actors and Speakers: The British Isles*. Milwaukee: Limelight.
- Cambier-Langeveld, Tina. 2010. "The role of linguists and native speakers in language analysis for the determination of speaker origin". *International Journal of Speech, Language and the Law*, 17(1): 67–93.
- Campbell-Kibler, Kathryn. 2009. "The nature of sociolinguistic perception." *Language Variation and Change*, 21(1): 135–156.
- Chambers, Jack. 1992. "Dialect acquisition." *Language*, 68(4): 673–705.
- Corrigan, Karen. 2010. *Irish English, Vol. I: Northern Ireland*. Edinburgh: Edinburgh University Press.
- Coupland, Nik, and Hywel Bishop. 2007. "Ideologised values for British accents." *Journal of Sociolinguistics*, 11(1): 74–93.
- Crystal, David. 2005. *Pronouncing Shakespeare: The Globe Experiment*. Cambridge: Cambridge University Press.
- Dixon, John, and Berenice Mahoney. 2004. "The effects of accent evaluation and evidence on perception of a suspect's guilt and criminality." *Journal of Social Psychology*, 144: 63–74.
- Dixon, John, Berenice Mahoney, and Roger Cocks. 2002. "Accents of guilt? Effects of regional accent, race, and crime type on attributions of guilt." *Journal of Language and Social Psychology*, 21(2): 162–168.
- Ensslin, Astrid. 2012. *The Language of Gaming*. Basingstoke: Palgrave.
- Eriksson, Anders. 2010. "The disguised voice: Imitating accents or speech styles and impersonating individuals." In *Language and Identities*, edited by Carmen Llamas and Dominic Watt, 86–96. Edinburgh: Edinburgh University Press.
- Evans, Bronwen, and Paul Iverson. 2007. "Plasticity in vowel perception and production: A study of accent change in young adults." *Journal of the Acoustical Society of America*, 121(6): 3814–3826.
- Falconer, Bruce. 2009. "Escape from America." *Lost Magazine*, 31. Online resource: <http://www.lostmag.com/issue31/escape.php> (retrieved 18 November 2016).
- Foulkes, Paul, and Peter French. 2012. "Forensic speaker comparison: A linguistic-acoustic perspective." In *The Oxford Handbook of Language and Law*, edited by Peter Tiersma, and Lawrence Solan, 557–572. Oxford: Oxford University Press.
- Fraser, Helen, and Bruce Stevenson. 2014. "The power and persistence of contextual priming." *The International Journal of Evidence and Proof*, 18(3): 205–229.
- French, Peter. 1990. "Analytic procedures for the determination of disputed utterances." In *Texte zu Theorie und Praxis Forensischer Linguistik*, edited by Hannes Kniffka, 201–213. Tübingen: Niemeyer.
- French, Peter, Philip Harrison, and Jack Windsor Lewis. 2007. "Case report – R -v- Humble, J.S.: The Yorkshire Ripper hoaxer trial." *International Journal of Speech, Language and the Law*, 13(2): 255–273.
- Gold, Erica, and Vincent Hughes. 2014. "Issues and opportunities: The application of the numerical likelihood ratio framework to forensic speaker comparison." *Science and Justice*, 54(4): 292–299.
- Grabe, Esther. 2004. "Intonational variation in urban dialects of English spoken in the British Isles." In *Regional Variation in Intonation*, edited by Peter Gilles, and Jörg Peters, 9–31. Tübingen: Niemeyer.
- Herman, Lewis, and Marguerite Herman. 1997. *American Dialects: A Manual for Actors, Directors, and Writers*. London: Routledge.
- James, Linda, and Olga Smith. 2006. *Get Rid of Your Accent: The English Pronunciation and Speech Training Manual*. London: Business and Technical Communication Services.
- James, Linda, and Olga Smith. 2011. *Get Rid of Your Accent, Advanced Level: The English Speech Training Manual, Part 2*. London: Business and Technical Communication Services.
- James, Linda, and Olga Smith. 2012. *Get Rid of Your Accent for Business: The English Speech Training Manual, Part 3*. London: Business and Technical Communication Services.
- Knowles, Gerry. 1973. *Scouse: The Urban Dialect of Liverpool*. PhD thesis, University of Leeds.
- Köster, Olaf, Michael Jessen, Freshta Khairi, and Hartwig Eckert. 2007. "Auditory-perceptual identification of voice quality by expert and non-expert listeners." *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken: 1845–1848.
- Köster, Olaf, Roland Kehrein, Karen Masthoff, and Yasmin Boubaker. 2012. "The tell-tale accent: Identification of regionally marked speech in German telephone conversations by forensic phoneticians." *International Journal of Speech, Language and the Law*, 19(1): 51–71.
- Leemann, Adrian, and Marie-José Kolly. 2013. *DialäktÄpp*. Online resource: <https://itunes.apple.com/ch/app/dialakt-app/id606559705?mt=8> (accessed 18 November 2016).

- Lippi-Green, Rosina. 2012. *English with an Accent: Language, Ideology, and Discrimination in the United States*, 2nd ed. London: Routledge.
- Loots, Linsen, and Thomas Niesler. 2011. "Automatic conversion between pronunciations of different English accents." *Speech Communication* 53(1): 75–84.
- Machlin, Evangeline. 1992. *Speech for the Stage*. New York: Routledge.
- Markham, Duncan. 1999. "Listeners and disguised voices: The imitation and perception of dialectal accent." *Forensic Linguistics*, 6(2): 289–299.
- Mazrui, Alamin. 2012. "Language ideology, translation/interpretation, and courts." In *The Encyclopedia of Applied Linguistics*, edited by Carol Chapelle. New York: Wiley. DOI: 10.1002/9781405198431.wbeal0627.
- McCallion, Michael. 1989. *The Voice Book: For Actors, Public Speakers, and Everyone Who Wants to Make the Most of Their Voice*. London: Faber and Faber.
- McDougall, Kirsty. 2014. "Listeners' perception of voice similarity in Standard Southern British English versus York English." Paper presented at the Annual Conference of the International Association for Forensic Phonetics and Acoustics, Zürich, September 2014.
- Meier, Paul. 2011. *Accents and Dialects for Stage and Screen*. Lawrence, KS: Paul Meier Dialect Services.
- Meier, Paul. 2012a. *Dialects of the British Isles*. Lawrence, KS: Paul Meier Dialect Services.
- Meier, Paul. 2012b. *The Standard British English Dialect*. Lawrence, KS: Paul Meier Dialect Services.
- Milroy, James, and Lesley Milroy. 2012. *Authority in Language: Investigating Standard English*, 4th ed. London: Routledge.
- Morales, Andrea, Maura Scott, and Eric Yorkston. 2012. "The role of accent standardness in message preference and recall." *Journal of Advertising*, 41(1): 33–46.
- Morrison, Malcolm. 2001. *Clear Speech: Practical Speech Correction and Voice Improvement*, 4th ed. London: Black.
- Mugglestone, Linda. 2007. *Talking Proper: The Rise of the English Accent as Social Symbol*, 2nd ed. Oxford: Oxford University Press.
- Nolan, Francis. 2003. "A recent voice parade." *Forensic Linguistics*, 10(2): 277–291.
- Nolan, Francis. 2012. "Degrees of freedom in speech production: An argument for native speakers in LADO." *International Journal of Speech, Language and the Law*, 19(2): 263–289.
- Nolan, Francis, and Esther Grabe. 1996. "Preparing a voice lineup." *Forensic Linguistics*, 3(1): 74–94.
- Nolan, Francis, Kirsty McDougall, Gea de Jong, and Toby Hudson. 2009. "The DyViS database: Style-controlled recordings of 100 homogeneous speakers for forensic phonetic research." *International Journal of Speech, Language and the Law*, 16(1): 31–57.
- Nycz, Jennifer. 2013. "New contrast acquisition: Methodological issues and theoretical implications." *English Language and Linguistics*, 17(2): 325–357.
- Odene, Cristina. 2012. "Without Charlotte Green and Harriet Cass, Radio 4 will turn into Radio 5 Live." *The Telegraph*, 5th September.
- Pardo, Jennifer, Rachel Gibbons, Alexandra Suppes, and Robert Krauss. 2012. "Phonetic convergence in college roommates." *Journal of Phonetics*, 40: 190–197.
- Patrick, Peter. 2012. "Language analysis for determination of origin: Objective evidence for refugee status determination." In *The Oxford Handbook of Language and Law*, edited by Peter Tiersma, and Lawrence Solan, 533–546. Oxford: Oxford University Press.
- Preston, Dennis. 1989. *Perceptual Dialectology: Non-Linguists' Views of Areal Linguistics*. Dordrecht: Foris.
- Rippmann, Walter. 1909. *The Sounds of Spoken English: A Manual of Ear Training for English Students*, 3rd ed. London: Dent.
- Rippmann, Walter. 1911. *English Sounds: A Book for English Boys and Girls*. New York: Dutton.
- Rippmann, Walter, and Bessie Robson. 1913. *English Sounds: Adapted for Use in Scottish Schools*. London: Dent.
- Rittmayer, Allison. 2009. "Translation and film: Slang, dialects, accents and multiple languages." *Comparative Humanities Review – Translation: Comparative Perspectives*, 3, Article 1. Online resource: <http://digitalcommons.bucknell.edu/chr/vol3/iss1/1> (retrieved 18 November 2016).
- Rodenburg, Patsy. 1998. *The Actor Speaks: Voice and the Performer*. London: Methuen.
- Sangster, Catherine. 2008. "The work of the BBC Pronunciation Unit in the 21st century." *Arbeiten aus Anglistik und Amerikanistik*, 33(2): 251–261.
- Sharpe, Edda, and Jan Haydn Rowles. 2009. *How to Do Accents*, 2nd ed. London: Oberon.
- Sharpe, Edda, and Jan Haydn Rowles. 2011. *How to Do Standard English Accents: From Traditional RP to the New 21st-Century Neutral Accent*. London: Oberon.

- Shaw, George Bernard. 1916. *Pygmalion*. New York: Brentano.
- Snell, Julia. 2013. "Dialect, interaction and class positioning at school: From deficit to difference to repertoire." *Language and Education*, 27(2): 110–128.
- Strong, Gwyneth, and Penny Dyer. 2007. *Access Accents – Received Pronunciation: An Accent Training Resource for Actors* [audiobook]. London: Methuen.
- Thomson, Ron. 2012. Accent reduction. In *The Encyclopedia of Applied Linguistics*, edited by Carol Chapelle. New York: Wiley. DOI: 10.1002/9781405198431.wbeal0004.
- Trudgill, Peter. 2008. "The historical sociolinguistics of elite accent change: On why RP is not disappearing." *Studia Anglica Posnaniensia*, 44: 3–12.
- Turner, J. Clifford, and Malcolm Morrison. 2000. *Voice and Speech in the Theatre*, 5th edn. London: Black.
- Watt, Dominic. 2012. "Sociolinguistic variation in vowels." In *Handbook of Vowels and Vowel Disorders*, edited by Martin Ball, and Fiona Gibbon, 207–228. New York: Psychology Press.
- Wilson, Kim, and Paul Foulkes. 2014. "Borders, variation and identity: Language analysis for the determination of origin (LADO)." In *Language, Borders and Identity*, edited by Dominic Watt, and Carmen Llamas, 218–229. Edinburgh: Edinburgh University Press.
- Yarmey, A. Daniel. 2012. "Factors affecting lay persons' identification of speakers." In *The Oxford Handbook of Language and Law*, edited by Peter Tiersma, and Lawrence Solan, 547–556. Oxford: Oxford University Press.
- Zwaan, Karin, Maaike Verrips, and Pieter Muysken, eds. 2010. *Language and Origin: The Role of Language in European Asylum Procedures*. Nijmegen: Wolf.

Section 2 – Methods

Introduction

JOHN NERBONNE

Methods

It is best to begin an introduction to the methods section of the handbook, which deals with gathering and analyzing data, with two near-platitudes that are not always kept in mind: first, that methods are subordinate to research questions (as well as to other factors), and second, that some methods are much more costly than others to apply.

It is easy to see that data collection depends on one's research question. To address any question in historical dialectology, we need data from different time periods—or from people of different ages. To consider a question on the diffusion of changes in dialects, we need to examine a substantial area in which diffusion patterns have a chance to emerge. These are obvious, almost commonsensical points, but others have become obvious only in the course of research. So work in social dialectology often tries to detect ongoing language changes, which (research has shown) are always accompanied by periods in which more than one linguistic form is used. To analyze such changes, frequency data is indispensable. For a further example, we note that work on diffusion has noted the importance of large population centers, meaning that such centers have to be part of the sample studied.

The basic point that methods are subordinate to research questions is a good principle, but not a hard and fast constraint. On the contrary, researchers have been able to exploit available material for research questions that had not even been posed when the data was collected. Nerbonne (2010) examines Trudgill's famous "gravitational view" of diffusion (Trudgill 1974) *inter alia* using data from the *Linguistic Atlas of the Middle and South Atlantic States* (Kretzschmar 1994), most of which was collected in the 1930s.¹ Eisenstein's chapter (below) analyzes the Twitter stream for evidence of geographic cohesion, and it is clear that neither the company Twitter nor the users of its services intended their communication as a contribution to dialectological data collection. Furthermore, there is also a tradition of collecting data in large cooperative efforts in dialectology in ways designed to serve many research questions—this is the tradition of the dialect atlas (see Kretzschmar's chapter, this volume). In atlas projects it is impossible to anticipate every research question, of course, and some newer questions turn out to be difficult, if not impossible to answer using atlas material. For example, the role of frequency in diffusion is theoretically interesting (Bybee 2002), but we don't have lexical frequency data for dialects. We might attempt to use frequency from standard language data (newspaper corpora), but that would entail risks.

It is also easy to understand how costs—in time and money—often play a role in data collection. For many purposes (see below), researchers would like to sample a good amount of material from many people over a large area. There are differences, of course, but in dialectology, just as in many other modern areas of research, the best data is often more data! Typical data collection efforts therefore exceed the capacity of individuals, making it necessary to seek funding. Most dialect research is funded through three- to six-year grants by research councils, foundations, and granting agencies, where funds are scarce and difficult to obtain, and where the time is always quite limited, too. Scientific academies are honorable exceptions, where longer-term projects may be conducted, but even the academies' funds and time are limited. So it is small wonder that dialectologists have sought help, for example, through crowd-sourcing. The dictionary of Flemish dialects has a team of volunteers to help in tasks such as digitizing field records (see <http://www.wvd.ugent.be/>). The data collection chapters also report on the use of telephones, internet, and smart phones to streamline data collection (especially the chapters on written surveys and on interviews). The big promise of the new techniques is their efficiency: they allow data to be collected much more quickly and in larger amounts than old methods.

Before continuing to more specific dialectological concerns, it will be useful to recall one basic statistical concept, that of *noise*, or unsystematic variation. In variationist linguistics, encompassing dialectology and sociolinguistics, we are always interested in variation, whether it be in pronunciation, lexical choice, morphology, syntax, or elsewhere. We develop and test hypotheses about the systematic variation, for example, the variation in the pronunciation of a word such as *bike* as one travels from North to South in the United States. Good hypotheses explain some of the variation we find. But in addition, we encounter variation beyond our hypotheses, which we therefore cannot explain. The additional variation may depend on how carefully the speaker spoke, on when in an interview a word was elicited (in the beginning or at the end), on the linguistic context the word was used in, and on other factors that data collectors may not have even tried to control. The variation that is not part the focus of the study is noise, or what statisticians call unsystematic variation, and we must be particularly sensitive to it given our focus on its converse, systematic variation. We should also bear in mind that what one study treats as unsystematic variation may nonetheless be structured with respect to variables not included in the study.

Returning to data collection, and continuing beyond its costs and dependence on research questions, dialectological data collection efforts may be seen as positioned along two major dimensions, naturalness, and commensurability, which of course leads to some tension. In characterizing dialects, we would like to hear how the people in an area *normally* speak, that is, when not accommodating to visitors from outside or projecting an ideal (with respect to education, etiquette or conformance to standards) that may be quite idiosyncratic. Labov has dubbed the very presence of an interviewer “the observer’s paradox” (Labov 1972: 209), noting that distortion of normal speech patterns increased as speakers’ self-awareness was heightened by formal, controlled elicitation methods. This brings us to the issue of ensuring that data is commensurable, that is, that one may compare items to one other without suspecting the influence of confounding factors. Simplifying a bit, many dialect atlases aimed to provide data that one may compare with respect to geography alone—where all the other factors that influence variation (age, gender, educational level, etc.) are fixed. Other atlas efforts attempt to vary some of the demographic factors systematically, too. One may ensure commensurability by using very strict protocols for interviews, including lists of concepts whose lexical realization is to be noted or words whose pronunciation is to be recorded (see the Chapter below on questionnaires). This allows us to study variation along the lines used in the lists. If the items are chosen carefully, we can eliminate some sources of noise that may arise in spontaneous speech, such as the context

dependence of lexical choice and pronunciation. The tension then arises as we recall that strict protocols tend to dampen the spontaneity of the interview and of the speech.

Data Collection

The first five chapters in this section of the book concern data collection. Data sampling concerns how to ensure even geographical coverage, the choice of respondents to interview, and how to approach and engage respondents. A major choice arises as to whether to collect data via an interview or via a written survey, which can be administered remotely (by mail or by web questionnaire), and which is therefore quicker and cheaper than interviews. Questionnaires are not restricted to use in written surveys, however; they often play a role in structuring oral interviews as well. The chapter on questionnaires focuses on the linguistic choices in the data collection effort regardless of the mode (oral or written) the collection assumes.

The field interview is a *primus inter pares* among data collection methods. In contrast to written survey techniques, the interviewer is present with the respondent, and ideally working hard to ensure that the conditions from interview to interview are commensurable. For example, an interviewer can note that there was an interruption in the process at some point, something written surveys have not been able to do.² As the chapter on field interviews documents, the interview can also steer the conversation toward topics where un-self-conscious production is most likely. The field interview likewise guarantees the broad bandwidth inherent to face-to-face interaction between the interviewer and the respondent. It allows the interviewer to observe details of pronunciation that are difficult to notice in recordings and to catch subtle signals that may indicate how comfortable a respondent is with a particular formulation.

A newcomer to the team of data collection methodologies is corpus linguistics, in which data is usually extracted from corpora, that is, large collections of speech and/or text. These may be corpora of dialect speech, the focus of the chapter in this section, but good work has also been done on a corpus of letters to the editor in American newspapers (Grieve 2011). In all cases one examines genuinely occurring speech or text, and, while the source of the corpora is essential to questions about representativeness, the methods of analysis used are not tied intimately to the source of the corpora. While practitioners have occasionally claimed that the speech in corpora are more natural or less self-conscious than, for example, dialect atlas data, it should be clear that this varies, depending on the particular choice of corpus and atlas. Nonetheless, extracting features (e.g., lexical realizations, morphological forms) from genuinely occurring speech or text entails dealing with the great skew in word frequency distributions (Baayen 2001), and this leads to problems in identifying comparable material. Zipf (1932) noted that if one sorts words by frequency, then the frequency of the n -th most frequent word is roughly $1/n$ times the frequency of the most frequent word. Of course, there have been subsequent attempts to reformulate and refine this (Baayen 2001), but the rough relationship suffices to make the following point. Since adults have vocabularies of tens of thousands of words, this means that most words occur rather infrequently. If we leave it to chance to elicit comparable words from speakers in different areas of a survey, then the chance of hearing any but the most common words in all of say, 20 sites, is only negligibly above zero. It is no accident, therefore, that corpus-based techniques have either focused on frequent elements, or have aggregated over classes of elements, sometimes combining the two, for example, by examining contracted versus uncontracted forms (since contracted forms are always forms of *be*, *have* or one of the modals, they are fairly frequent). The chapter on social media might also be regarded as a corpus-based analysis.

Linguistic and Geographical Methods

Two chapters are devoted to the instrumental and computational analysis of linguistic data and one to the various sorts of mapping techniques popular in dialectology. Acoustic phonetics has long been advocated as an analytical tool in dialectology (Labov, Yaeger, and Steiner 1972), especially as a way to obviate the need for phonetic transcription, which is notoriously difficult and subjective. The analysis of vowels is well established in dialectology and sociolinguistics and is now carried out automatically on large sets of vowels (Rosenfelder *et al.* 2011). Work on consonants is progressing and is reported on in the chapter. For text-encoded material, including phonetic transcriptions, several techniques from computational linguistics are potentially useful. Edit-distance measures are becoming standard; lemmatization may facilitate morphological studies and stimulate them further; and some simple syntactic analyses are robust enough for some purposes (Wieling and Nerbonne 2015).

We would predict that automatic analysis of the sorts presented in these two chapters will be of increasing interest to dialectologists for several reasons. One is simply that the increasing volumes of data available for analysis necessitate increasing automation in analysis. Leinonen (2010) was able to extract the formants of nearly 20,000 vowel tokens in the SweDia corpus (Eriksson 2004) but only because she applied reliable, automatic procedures for extracting formants. Second, automating (parts of) analyses improves them with respect to replicability, a requirement more difficult to fulfill in manual work, especially work requiring judgment on the part of the researcher, such as phonetic transcription or superficial syntactic analysis in terms of parts of speech. Third, the automated analyses are more and more capable of identifying the latent structure that a great deal of linguistic discourse revolves around. Identifying the parts of speech of words is an excellent example of uncovering the sort of latent structure that on the one hand may be reliably identified and on the other may suffice to support dialectological analysis (Wolk 2014).

We need not belabor the desirability of including a chapter on maps in a handbook of dialectology, except to note that automatic procedures for making good quality maps are also improving, not only in the graphic quality of the maps, but also in the ease with which the software may be used, and finally in the range of functions available for adding information to the geography—including political and physical boundaries, population sizes, densities of occurrence, optimum tiling for networks of data collection sites, and more. In short, maps are essential to dialectology.

Statistics for Variationist Studies

The statistical chapters may be the most challenging technically, but it is clear that a field as data-rich as dialectology could never forego statistics. Because the chapters are challenging, it will be worthwhile to remind readers of the foundation they are built on.

The chapters assume some familiarity with null hypothesis testing (NHT), a practice that has been the basis of a great deal of statistical analysis, but which has increasingly come under fire (see Field *et al.* 2012: Ch.2 and references there). It is still not clear what will take the place of NHT, but focus in statistical analysis is shifting toward model comparison. In NHT one contrasts the hypothesis of interest, perhaps a proposed difference in the mean recognition times for two classes of words, with a *null hypothesis*, which assumes that there is no effect, that is, no difference. Whether such a difference is regarded as improbable will naturally depend on the (systematic and unsystematic) variation in the data, what is called the spread in the distribution of the data. For numerical data such as reaction times (RT), the common measure of spread is the *standard deviation*. If the sample turns out to be improbable when one assumes the null hypothesis, this is regarded as evidence for the hypothesis of interest. The probability of the sample given the null hypothesis is called the *p-value*, and the

lower the *p*-value, the less likely the null hypothesis, and the stronger the evidence for the alternative (the hypothesis of interest).

When the *p*-value falls below an agreed on threshold, say, 0.01, we say that the study is *statistically significant*. This use of the word “significant” must not be confused with its everyday sense, that is, “meaningful, or having important consequences.” A *p*-value is a probability and is crucially influenced by how large a sample was studied. Statistically insignificant differences in small samples (say, a 6-to-4 difference in a sample of 10, or a 5 millisecond difference in RT in a sample of 20 reaction times) inevitably become significant at some (perhaps very large) sample size.³ This is not just a theoretical possibility, but one which is frequently seen as variationists examine ever larger data sets. It means that the intelligent reader of statistical analyses not only examines *p*-values, but also tries to gauge **EFFECT SIZE**. How effect size is measured depends on the research question and the analysis technique, but for a comparison of mean values, the difference in means, expressed in terms of the number of standard deviations, is a common measure. If we notice a difference of 40 milliseconds in recognition time for two classes of words, and the standard deviation is 80 ms., then we would express the effect size as Cohen’s *d* (=40 ms./80 ms.), or about 0.5 standard deviations.

A lot of linguistic data is categorical, that is, it occurs in various categories, where there is no intrinsic order among them. Examples are different parts of speech (Noun, Verb, etc.) or different realizations of /t/ in English ([t], [ɾ], [ʔ]). For many years, the analysis of such categorical data was limited to the χ^2 test of independence, which is ill-suited to analyzing multiple influences, which Bayley (2013) has dubbed “the principle of multiple causes.” Language variationists are convinced that choices in variation may be influenced by many factors, and that combinations are possible and important. For the sake of completeness, the chapter on logistic regression also explains χ^2 analysis and its limitations.

Sociolinguists pioneered the use of logistic regression in linguistics and have continued to use it for 40 years, so that it is only fitting that we include a chapter devoted primarily to the models underlying logistic regression, its most important prerequisites, its application, and the interpretation of its results.

Ignoring honorable exceptions (Woods 1979, Gregg *et al.* 1981), most sociolinguistic analyses focus on a small number of variables. In contrast dialectology often focuses not on individual variables (features) but on large numbers of linguistic variables simultaneously, for example, when atlases are compiled or corpora collected. Furthermore, the individual variables are often quite noisy, and one would wish as a researcher not to hand-pick variables with all the dangers of subjectively selecting exactly those variables that support the case one is making (Nerbonne 2009). This has promoted the dialectometric perspective (see chapter by Goebel, this volume), which has consistently emphasized the virtues of adopting an aggregate perspective on variation, and inspecting individual features only from that vantage point. The chapter on aggregate analyses also reports on first efforts at including geographical and social variables in a single statistical model, following the vision of Chambers and Trudgill (1998: Ch. 12).

In general, dialectologists have not made use of the specialized field of spatial statistics or what is also known as geo-statistics, and this is certainly a shortcoming of the field of dialectology. Geo-statistical techniques are often used in fields that ask similar questions about diffusion and barriers to diffusion, such as demography and epidemiology. One example of a point where they have gone beyond dialectometry concerns the selection of variables to analyze, where measures of spatial autocorrelation have been brought to bear. Dialectological method has much to learn from these fields.

Finally, we have included a chapter on social media in this section both because of its intrinsic interest, including the demonstration of the importance of geography even in the age of digital, world-wide communication, and because it uses usually sophisticated statistical reasoning from machine learning in order to draw conclusions from its data. We hope that its presence here will inspire more collaboration between variationist linguists and statisticians.

Future Challenges

Readers of this section should not come away with the impression that the methods in dialectology have stabilized to a point where we expect little innovation in the future. We suspect that just the opposite is the case. Data collection is likely to turn more and more to methods already in use such as web questionnaires and smart phone apps. Vaux and Golder (2003) pioneered the use of web questionnaires to collect English lexical data, and Möller & Elspass's (2008) questionnaire aims at everyday German variation not only in vocabulary, but also with respect to pronunciation and syntax. This new work will require a good deal of analysis to validate its methods and also to compare the results to older work. Work using smart phones is even newer, but Sherrer *et al.* (2012) and Leemann *et al.* (2015) (and other references they cite) describe Dialekt Äpp, an iOS application for collecting speech data that has already shown great promise.

On the analysis side, we certainly expect to see machine learning techniques make their way further into the analysis of dialect data, just as they have in other areas of statistics, especially exploratory statistics. The relation between single variable analyses and aggregate analyses also deserves further attention, as do further techniques for including geographic and social variables in single analyses (see chapter on aggregate analyses). To date, the more encompassing perspectives are regression analyses aimed at predicting the aggregate differences of a sample of varieties to a single alternative, the standard language. Techniques aimed at analyzing all pairs of varieties would improve our understanding further, as Wieling and Nerbonne (2015) also urge.

NOTES

- 1 Nerbonne and Heeringa (2007) examine the gravity hypothesis using data from the *Reeks Nederlandse Dialektatlassen* (RND), collected 1925–1982.
- 2 Some web-based surveys are implemented in Java programs, for example, Charlotte Gooskens' MICReLa project on mutual comprehensibility (www.let.rug.nl/gooskens/project/), which are definitely able to keep track of time, enabling some checks on the conduct of written surveys. Still, the check is minimal when compared to the presence of an interviewer during data elicitation.
- 3 Naturally apparent effects also depend on the particular sample studied, and how representative the sample is of the population of interest. Another point is being made in the main text, however, concerning differences which are relatively stable as sample size increases (the 6-to-4 ratio, or the five millisecond RT difference). Even if these remain stable, the statistical significance may change as the sample size grows.

REFERENCES

- Baayen, R. Harald. 2001. *Word frequency distributions*. Berlin: Springer (Springer Science & Business Media, Vol. 18).
- Bayley, Robert. 2013. "The quantitative paradigm." In: J.K. Chambers & Natalie Schilling-Estes (eds.) *The handbook of language variation and change*. Boston: Wiley. 117–141.
- Bybee, Joan. 2002. "Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change." *Language variation and change*, 14(3): 261–290.
- Chambers, J.K., & Peter Trudgill. 1998. *Dialectology*. Cambridge: Cambridge University Press.
- Eriksson, Anders. 2004. SweDia 2000: "A Swedish dialect database." In: P. J. Henrichse (ed.), *Babylonian Confusion Resolved. Proc. Nordic Symposium on the Comparison of*

- Spoken Languages. Copenhagen Working Papers in LSP.* Copenhagen.
- Field, Andy, Jeremy Miles, & Zoë Field. 2012. *Discovering statistics using R.* London: Sage.
- Gregg, Robert J., Margaret Murdoch, Erica Hasebe-Ludt, & Gaelan de Wolf. 1981. "An urban dialect survey of the English spoken in Vancouver." *Papers from the fourth international conference on methods in dialectology* (pp. 41–65). University of Victoria: Victoria, British Columbia.
- Grieve, Jack. 2011. "A regional analysis of contraction rate in written Standard American English." *International Journal of Corpus Linguistics*, 16(4): 514–546.
- Kretzschmar, William A. (ed.). 1994. *Handbook of the Linguistic Atlas of the Middle and South Atlantic States.* Chicago: The University of Chicago Press.
- Labov, William. 1972. *Sociolinguistic Patterns.* Philadelphia: University of Pennsylvania.
- Labov, William, Malcah Yaeger, & Richard Steiner. 1972. *A quantitative study of sound change in progress.* US Regional Survey, Vol. 1.
- Leemann, Adrian, Marie-José Kolly, David Britain, Ross Purves, & Elvira Glaser. 2015. "Documenting sound change with smartphone apps." *The Journal of the Acoustical Society of America*, 137(4): 2304–2304. 10.1121/1.4920412
- Leinonen, Therese. 2010. *An acoustic analysis of vowel pronunciation in Swedish dialects.* PhD thesis, Groningen.
- Möller, Robert & Stephan Elspaß. 2008. "Erhebung dialektgeographischer Daten per Internet: ein Atlasprojekt zur deutschen Alltagssprache." In: S. Elspaß & W. König (eds.) *Sprachgeographie digital. Die neue Generation der Sprachatlanten (mit 80 Karten).* Hildesheim: Olms. 115–132.
- Nerbonne, John. 2009. "Data-driven Dialectology." *Language and Linguistics Compass*, 3(1): 175–198.
- Nerbonne, John. 2010. "Measuring the diffusion of linguistic change." *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1559): 3821–3828.
- Nerbonne, John & Wilbert Heering. 2007. "Geographic distributions of linguistic variation reflect dynamics of differentiation." In: Sam Featherston and Wolfgang Sternefeld (eds.) *Roots: Linguistics in Search of its Evidential Base* Berlin: Mouton De Gruyter, 267–297.
- Rosenfelder, Ingrid, Josef Fruehwald, Keelan Evanini and Jiahong Yuan. 2011. FAVE (Forced Alignment and Vowel Extraction) Program Suite [Computer Program]. Available at: <http://fave.ling.upenn.edu/>.
- Scherrer, Yves, Adrian Leemann, Marie-José Kolly, & Iwar Werlen. 2012. "Dialäkt Äpp - A smartphone application for Swiss German dialects with great scientific potential." 7ème Congrès SIDG - Dialect 2.0, Vienna.
- Trudgill, Peter. 1974. "Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography." *Language in Society*, 2: 215–246.
- Vaux, Bert and Scott Golder. 2003. *The Harvard Dialect Survey.* Cambridge, MA: Harvard University Press.
- Wieling, Martijn, and John Nerbonne. 2015. "Advances in dialectometry." *Annual Review of Linguistics*, 1: 243–264.
- Wolk, Christoph. 2014. "Integrating aggregational and probabilistic approaches to language variation." PhD thesis, University of Freiburg. Available at: <https://www.freidok.uni-freiburg.de/data/9656/>.
- Woods, Howard B. 1979. *A Socio-dialectology survey of the English spoken in Ottawa.* National Library of Canada.
- Zipf, George K. 1932. *Selected studies of the principle of relative frequency in language.* Cambridge, MA: Harvard University Press.

13 Dialect Sampling Methods

RONALD MACAULAY

If you want to make tiger stew, the old saying has it, you must first catch your tiger. Since the middle of the nineteenth century dialectologists have been designing traps to catch their prey. As in many other endeavors, the trade-off is quantity versus quality, and as usual, the key factors are time and money, although intellectual commitment, human energy, and physical endurance are also of critical importance. Technological developments have transformed methods of collecting samples of language but the basic questions remain the same: What are the boundaries of the region to be surveyed? How many localities should be investigated? How big should the sample of informants be? How diverse should this sample be? How much should be recorded from each informant? Which aspects of language should be investigated? The answers to these questions will depend upon the goals of the survey. This chapter will examine a range of studies that have collected substantial amounts of information on dialect variation.

13.1 Sampling

The aim of many early dialect surveys was to collect information about forms of speech that might disappear because of social changes and they were looking for examples of “the uncorrupted folk speech of a given locality” (McDavid 1980, 210). Since men in country districts were believed to be more conservative in their speech than their wives and children, many surveys have focused on nonmobile, older, rural males, NORMs in the acronym created by Chambers and Trudgill (Chambers and Trudgill 1980, 33). One consequence of this aim is that dialect surveys have usually concentrated their efforts in smaller, relatively isolated communities and ignored larger urban centers.

The first task in any survey is what has been called “planning the grid” (McDavid 1980). This involves identifying the geographical area to be investigated and deciding at what locations to collect information. As the word “grid” suggests, deciding on locations is often based on the subdivision of the area into spatial cells of a fixed, constant size, with the goal of interviewing in one location per cell (Klepsch 2013, 20–22), but some “grids” varied cell size depending on population (Kretzschmar *et al.* 1993, 5–14), and researchers often report less formal attempts to ensure geographic coverage. If the aim is to collect evidence from speakers of traditional dialects who have been little influenced by the educationally promoted standard form of speech, then the target areas are smaller rural communities. The Survey of English dialects, for example, identified “agricultural communities that had had a fairly stable population of about five hundred inhabitants for a century or so” (Orton 1962, 15). The method chosen for identifying such communities was to “follow the rivers and tributaries towards their sources, and choose a place in the higher reaches” and then “choose a

place with about the same number of inhabitants below the junction of the tributary and the main stream" (Orton and Dieth 1951, 66–67). By this method, 311 communities were chosen to be investigated.

This approach would not work in the United States because of the differences in early settlement history, which from the start resulted in dialect mixture, including the influence of other languages. "There has been geographic mobility, with Americans always on the move, from older communities to new, and back again" (McDavid 1980, 239). For the survey of the dialects of New England, Hans Kurath chose a dense grid but decided that rural communities should be more heavily represented than urban ones because regional differences "are greater in the homely vocabulary of the family and farm than in the vocabulary of 'society' and of urban areas" (Kurath *et al.* 1939, 1). He also chose to give greater attention to older areas of settlement than to newer ones. Kurath stressed the importance of settlement history because "a realistic interpretation and a true understanding of local, regional and social 'dialects'—including the 'dialects' of the cultured, the regional standard languages—cannot be attained without a thorough knowledge of the population history" (Kurath *et al.* 1939, ix). To achieve this aim, Kurath and his team of eight fieldworkers completed 416 interviews in 213 communities and the target population was not restricted to NORMs.

Dialect surveys usually cannot employ strict random sampling methods because of the difficulty of identifying the target population in a way that would provide only the appropriate candidates for the purpose. Instead, the investigators have identified the kind of individuals they wanted as informants and then chosen individuals they considered suitable. For example, in the linguistic survey of New England (Kurath *et al.* 1939) three types of informant were to be selected:

1. A simple but intelligent farmer or farmer's wife in rural districts, a workman, tradesman, or shopkeeper in larger villages and in cities.
2. A middle-aged man or woman, native to the community, who had received better schooling (high school or academy in addition to grammar school), read more widely, or enjoyed contacts with the better educated.
3. Cultured informants, with a college education or the equivalent, were to be chosen in most of the larger cities, including all the older cultural centers, and in a number of smaller communities.

This is a much wider range of speakers than that chosen for the Survey of English Dialects with its emphasis on NORMs, although 72% of the speakers interviewed were at least 70 years old at the time they were interviewed. The instruction to the fieldworkers to choose one folk speaker and one common speaker in each community (plus in some communities a cultivated speaker) constitutes a stratified or quota sample. However, the choice of informant was made by the investigator who identified "the intelligent farmer" or the person "who read more widely." Since the choice also depended on the willingness of those identified in this way to be interviewed, this method comes into the category of convenience sampling (Baxter and Babbie 2004, 134–135) and from this kind of sample it is risky to make reliable generalizations. This weakness is less important in studying traditional dialects because dialect surveys generally do not aim at statistically significant differences, but rather to "study regularities... in a population as exemplified by the individuals in the sample" (Cochran, Mosteller, and Tukey 1954, 17).

Sociolinguistics has in general not shared the (historical) goals of dialect surveys, but has rather focused on identifying speech variation that is associated with social differences and in finding examples of language change in progress. In order to be able to claim that their results are not due to chance, they have been interested in statistical significance and have

often employed quasi-random sampling methods. Labov (1966) in his study of Lower East Side New York drew most of his informants from the sociological survey Mobilization for Youth but according to Davis (1990, 6) it cannot be claimed that "he analyzed data from a random sample" since the changes he made in making his selection destroyed its randomness. Labov was less concerned about this because he held that "the regularities observed when degrees of variability are taken seriously are so profound as to make statistical tests irrelevant" (Fasold 1972, 35). Nevertheless, Labov's use of a non-subjectively identified sample set the pattern for later sociolinguists.

Trudgill in his study of Norwich (Trudgill 1974) chose a method of systematic sampling from the local register of electors. Trudgill chose his sample from four wards and one suburb to ensure the inclusion of a range of speakers from different types of social background. From those registered Trudgill drew a random sample of 125 speakers and, despite a high number of refusals, succeeded in obtaining 50 individuals to interview. Since only people aged 21 or over were listed on the register of electors, Trudgill added a sample of 10 schoolchildren, aged 10 to 20, from two of the schools in Norwich. A comparison with earlier census information shows that Trudgill's sample was similar to that of the general population of Norwich. According to Linn (1983, 232) systematic sampling "has an advantage over random sampling in that it assures a more even spread of informants over the population." Trudgill's sample is probably as close to a random sample as any linguist could hope to achieve by direct selection.

Occasionally, the opportunity arises for a true random sample:

"For surveys of large, socially complex populations, random samples offer the best opportunity for getting accurate data. The only problem, of course, is obtaining such a sample." (Bailey and Dyer 1992, 3)

Bailey and his associates solved this problem by piggy-backing on a random-sample telephone survey of the state of Texas. The Texas Poll selects a sample of 1,000 randomly chosen Texans over the age of 18 from a list of all possible telephone numbers in the region, both listed and unlisted. Then to randomly selected households, the interviewer asks to speak with the person with the most recent birthday. For the Grammatical Investigation of Texas Speech, Bailey and his associates were able to add a section asking for information on language use.

One way to obtain a quasi-random sample is to develop a judgment sample in which speakers are identified by membership in certain categories such as social class, age, gender, religion, or race. For example, an investigator might to choose to interview 10-year-olds, 15-year-olds, and adults, both male and female, from objectively determined different social classes, but the individuals chosen in each category would be identified by some non-subjective process so that there would be no direct influence of the investigator in the selection of speakers (Macaulay 1977a).

An alternative to sampling speakers on the basis of some social category is to make use of the existence of social networks (Milroy 1987). Instead of selecting a sample on the basis of some rigid criteria, this approach makes use of the social relationships that create a group of speakers whose language can be compared with that of other groups of speakers. A variant of this approach is to examine the speech of those who come together for some mutual interest or activity, a category that has come to be known as Communities of Practice (Ekert and Wenger 2005).

An extreme example of a convenience sample was an Australian project that employed a simple technique for eliciting brief samples of speech by recording speakers in the streets or parks or other gathering places (Horvath and Horvath 2011). A word list and a set of reading passages were administered to 312 speakers, with a range of ages, sex, and social class, in

sessions lasting no more than six minutes. This was not a dialect survey since the project was directed at a single phenomenon (/l/ vocalization) but this quick and dirty technique could be adapted for wider dialect purposes.

13.2 Dialect Surveys

Apart from Joseph Wright's lexical surveys for his *English Dialect Dictionary* (Wright 1898–1905), the first systematic approach to studying dialects in Britain was by Alexander Ellis. Ellis sent out word lists and reading passages to a range of people in different parts of the British Isles but he admitted "I have not swept the country and most of my brooms so far as I went were not of perfect construction" (Ellis 1889, Part V, p. 8). He collected information from 811 people reporting on 1,145 locations but the distribution was highly skewed to the north of England with, for example, 93 informants in Yorkshire but only 7 in Middlesex. Although the quality of Ellis's work has often been criticized, his work still proves useful to those who are interested in linguistic change. Before 1950, there was so little systematic investigation of dialects in Britain that Sever Pop in his monumental survey *La Dialectologie* (1950) allocated only 4 pages (out of a total of 1,334 pages) to Britain. (Despite Kurath's substantial work in New England, the United States received a paltry nine pages.)

Two pioneering European dialect studies in the nineteenth century employed diametrically opposed methods. In the 1870s, Georg Wenker sent out a questionnaire to every German village with a school and received 44,251 completed forms from 40,736 communities, a remarkable rate of return. Wenker's questionnaire consisted of 44 sentences in Standard German and asked for the equivalents in the local dialect. Although designed to elicit fine phonetic data, "the collections proved notably refractory for that purpose" (McDavid 1966, 9).

In contrast, Jules Gilliéron employed Edmond Edmont to record in fine phonetic detail information from a local speaker in 639 communities in France, a task that took him five years. Gilliéron's principles, as set out by McDavid (1980, 211), are:

1. A network of selected communities.
2. Representative local informants in each community.
3. A questionnaire of selected items.
4. Interviewing by trained investigators.
5. Interviewing in a conversational situation.
6. Recording of responses in finely graded impressionistic phonetics.

Gilliéron's questionnaire included about 1,500 items (Gilliéron 1902, 10). Few subsequent investigators have attempted to match these heroic efforts.

Gilliéron's and Wenker's investigations illustrate the two challenges for dialect surveys: 1) How to collect information from a wide range of speakers; and 2) How to collect enough information from each respondent. Chapter 14 of this Handbook treats the choice of linguistic material in dialect surveys, whereas this chapter focuses on the choice of sites and respondents. But as the two questions above illustrate, the issues interact. Gilliéron's survey provided much more information from each location but reached many fewer communities. The trade-off between quantity and quality is a constant challenge in dialect surveys. The contrast between Wenker's and Gilliéron's approaches can be illustrated by comparing two twentieth century investigations: the Survey of English Dialects (SED) and the Linguistic Survey of Scotland (LSS).

The fieldwork for SED was carried out from 1950 to 1961 in 311 localities, with preference given "to agricultural communities that had had a fairly stable population of about five hundred inhabitants for a century or so" (Orton 1962, 15). Consequently, with the exceptions of Leeds, York, Sheffield, and Hackney (London) no urban communities were investigated.

The number of communities recorded varied from 2 in Roxburghshire to 34 in Yorkshire, but there is no explanation provided for the variation in density of sampling, though the location of the survey in Leeds may account for the large number of Yorkshire interviews.

Since the survey was designed to obtain information from dialect speakers, most of the informants were men aged 60 or over, because "in this country men speak vernacular more frequently, more consistently, and more genuinely than women" (Orton 1962, 15). The questionnaire included 1,322 items, just over half dealing with lexical items. Given the size of the questionnaire more than one informant was interviewed in each location, so the responses are not those of a single speaker. The majority of interviews involved three to four speakers and a few from five to as many as seven. The questionnaire "could be recorded satisfactorily and conveniently in some four days" (Orton 1962, 17). A total of nine fieldworkers participated in interviewing, with the number of localities they visited ranging from 9 to 118.

There could not be a greater contrast to the methods used in the SED than those used for the Linguistic Survey of Scotland (LSS). McIntosh (1952) in his description of the Survey of Scottish Dialects takes a more sober view of the situation: "questionnaires designed to be comprehensive, or said to be so, never in fact turn out to provide anything like exhaustive information, and it is probably wiser from the start to aim at something less ambitious" (McIntosh 1952, 65). Instead of the lengthy interviews conducted by the SED fieldworkers, the initiators of LSS designed a much shorter postal questionnaire to be sent out to the head teacher of every rural school and selected city schools in Scotland, Northumberland, Cumberland, and Ulster. There were two postal questionnaires sent out (PQ1 and PQ2). In the case of PQ1 some 3000 questionnaires were sent out and nearly two-thirds were returned. This is a lower rate of return than Wenker achieved in Germany but is still impressive. Of those questionnaires completed, 1774 were considered suitable for processing, 1337 from Scotland, out of a total population of approximately 5 million. This represents 1 informant for every 3800 inhabitants (or if the four major cities are excluded, since they were barely sampled, 1 informant for about 2400 inhabitants). The proportion for SED is one locality for every 144,000 inhabitants, so the density of coverage for the two surveys is very different. (The response to PQ2 was much lower with 832 returns.)

The increase in coverage comes at the expense of a certain kind of quality since there is limited information on the individual who completed the questionnaire. The head teachers were instructed to put the questionnaire into the hands of "some local person who is willing and competent to undertake this work.... The person chosen should if possible be middle-aged or older and a lifelong inhabitant of your district" (Mather and Speitel 1975, 14). Unfortunately, the questionnaire did not ask for information about the respondent's education or occupation. It is therefore impossible to tell how many of the questionnaires were completed by the teachers themselves, or what kind of local person they asked. Consequently, there is no information on the social position of the respondent. Instead, for each respondent the atlas lists the place of residence, sex, age, length of residence in that locality, birthplace, and birthplace of father and mother. This information makes it possible to identify some unreliable informants who do not represent the locality (Macaulay 1977b) but given the size of the sample, their responses are insignificant.

In contrast to the large number of questions in the SED questionnaire, PQ1 contained 211 items of which 184 asked for information on lexical items. PQ2 had 246 questions of which 218 deal with lexical items. The number of items reported is, however, much lower. The first atlas volume (Mather and Speitel 1975) provides information on 90 lexical items from PQ1 and the second (Mather and Speitel 1977) reports responses to 80 items from PQ2. The SED report (Orton and Dieth 1962–1971) presents all the information collected. There is another important difference. Harold Orton was involved in planning and directing the whole project of SED. J.Y. Mather and H.H. Speitel took over the publication of the LSS materials only after the questionnaires had been designed, distributed, and collected.

Although the number of items covered in PQ1 and PQ2 is much smaller than in SED (and the number of responses reported is even fewer), the range of items is wider. Of the 170 lexical items from PQ1 and PQ2 for which responses are shown, SED includes only 92. In SED the questionnaire was specifically designed for those who had worked on a farm and despite its much greater length it has much narrower range. The LSS questionnaires were developed in consultation with the editors of the Scots dictionaries to collect a wide range of traditional terms. As a result they contain many items of general knowledge, including names of birds, insects, and children's games.

A direct comparison between LSS and SED can be seen in the responses for the English counties Northumberland (Nb.) and Cumberland (Cu.). SED has 9 informants in Nb. and 6 in Cu; LSS has 233 respondents in Nb. and 72 in Cu. Ten lexical items reveal the difference the more extensive coverage makes. For "blisters," Orton and Wright (1974, 202) show a small area in Yorkshire as using the form *blebs*, with two sporadic forms in Nb.; the latter area, however, is shown as preferring the form *blushes*. In LSS, 175 Nb. respondents are shown as supplying the form *bleb*, and 46 give *blice*; in contrast, only 62 are listed as giving the response *blush*. In Cu., 54 are listed as giving *bleb*. Another striking example is the form (*hay*)-*rick* for "haystack." Orton and Wright (1974, 193) show this as a predominantly southwestern form, with no use recorded north of Derbyshire; but LAS records it from 65 respondents in Nb. and from 18 in Cu. Almost as remarkable are the responses for "pigsty." In Nb., SED found only two uses of *cree*, one of which said to be "rare," whereas a third informant said that the form was "not used"; however, LSS received 118 responses with *cree* in Nb. These examples show the consequence of choosing one method of sampling over another. The SED is dependent on a small number of informants for each county and their responses may (or may not) be representative of the area being sampled. The LSS responses are from a much larger sample of respondents in each county and this decreases the risk that some forms will be missed.

It is not simply a matter of number of respondents but also the range of forms elicited. McDavid (1953, 566) was surprised that the SED questionnaire did not contain any question about playing truant: "are English children more dutiful in their attendance or are English truant officers more efficient?" Mather and Speitel (1975, 67) report 68 items for this activity in Scotland, 30 different names in Nb., and 18 in Cu. The SED questionnaire was much more narrowly focused on the farm and farming. The LSS questionnaires contained a wider range of items, including children's activities, wild flowers, and birds.

The construction of the SED questionnaire may have contributed to this situation. Dieth and Orton were clearly determined to make the questionnaire as comprehensive as possible, and they included words such as *shepherd* for which there were almost no dialect forms used in response. A field testing of the questionnaire could have allowed the removal of items that were unlikely to provide dialect forms and reduced the time to administer it. With a shorter questionnaire, the fieldworkers would have been able to interview more informants and provide a wider coverage. Of course, the SED fieldworkers were also recording phonetic information but a questionnaire of over 1300 items is not necessary for that purpose. The fieldworkers usually interviewed several informants but they did not record different answers to any questions, so the lexical information comes from a single speaker in each location. With the LSS responses there is no way to know whether they represent the speech of a single speaker since the questionnaire asks for "local words" rather than the respondent's own usage.

These two surveys present an interesting contrast because they are contemporaneous and overlap to some extent. The next section will deal with surveys that employed fieldworkers to collect information and a later section will deal with other examples of the use of postal questionnaires.

Kurath's account of the survey of the dialects of New England is exemplary in its description of the factors involved in designing a dialect exploration (Kurath *et al.* 1939). After

months of exploratory fieldwork, 213 communities were selected and a total of 416 informants were interviewed. Both urban and rural communities were included.

After the fieldwork had been completed, the informants were classified into five types:

- Type I. Little formal education, little reading, and restricted social contacts.
- Type II. Better formal education (usually high school) and/or wider reading and social contacts.
- Type III. Superior education (usually college), cultured background, wide reading, and/or extensive social contacts.
- Type A. Aged, and/or regarded by the fieldworker as old-fashioned.
- Type B. Middle-aged or younger, and/or regarded by the fieldworker as more modern

This information, however, was not used in analyzing the responses.

There were over 700 questions to be asked but the exact form of the question was left to the fieldworker. The fieldworkers were given detailed instructions on how to conduct the interview and afterward were ranked according to "their phonetic accuracy, their fullness and accuracy in recording vocabulary, and their practice of giving information on the currency and social status of pronunciations, words, and grammatical forms" (Kurath *et al.* 1939, 52). (It is perhaps not surprising that Bernard Bloch was rated highest on seven out of the nine items.)

Kurath *et al.* (1939) also give potted biographies of many of the informants: "Friendly, kind-hearted, but rather stern" (p. 189), "Strong-minded; intelligent, but unimaginative. Thinks well of himself" (p. 201), "Has done much to educate himself. (Sent one son to Harvard and another to Yale)" (p. 167), "Tremendous self-confidence and personal magnetism. Keen intelligence. Exhaustive information on all familiar subjects. Now blind and crippled by 'rheumatiz.' Most of the record was jotted down from his incessant flow of conversation, reminiscence and anecdote." (p. 212). Again, this information was recorded but not put to any use in analyzing the responses. It is hard to imagine a more carefully planned and administered dialect survey than Kurath's, but it has been criticized for the freedom allowed to the fieldworkers (Atwood 1971).

Kurath's work in New England was extended in the work for the Linguistic Atlas of the Middle and South Atlantic States (Kretzschmar, McDavid, Lerud, and Johnson 1993). In this survey the target population was the white adult natives of their communities. Of the 1,162 informants interviewed, only 45 were African-Americans. There was a wide age range with three below the age of 20 and ten over 90. There was gender bias in that 70% of the informants were male but this was the result of a deliberate emphasis on men in the rural communities. In the urban areas almost 60% of the informants were women.

Labov's New York study (Labov 1966) was not designed as a dialect survey but McDavid saw it as a complement to the work of dialect geographers (McDavid 1971). Labov's pioneering work had a significant impact on dialect surveys. First, by the use of a portable tape-recorder to collect samples of speech, and secondly, through the development of his sociolinguistic questionnaire, which elicited a wider range of speech than had been the norm in traditional dialect surveys. The most dramatic impact was on the Linguistic Atlas of the Gulf States (LAGS).

The LAGS project took 25 years to complete, from the initial planning to the publication of the results in seven volumes. It covers seven Southern states (Alabama, Arkansas, Florida, Georgia, Louisiana, Mississippi, and Tennessee) and part of East Texas. All the interviews were tape-recorded and the tapes are now housed at the University of Georgia where they can be examined by any interested scholar. The LAGS interviewers recorded 5,300 hours of tape from 1,121 interviews. The range of informants was wider than in most dialect surveys. Nearly half were women and just over 20% were African American (about 15 interviews

were conducted by African Americans). LAGS also paid more attention to urban areas than had been the case in earlier surveys, and an additional part of the questionnaire dealt with topics more likely to be familiar to city dwellers than the traditional items of rural life. In addition to eliciting information on a wide range of features, the interviewers also encouraged their informants to talk freely on subjects that interested them providing a richer source of speech than found in most dialect surveys.

An even larger number of informants were interviewed for the Dictionary of American Regional English (DARE). As its name indicates, this was a project aimed at lexical variety. A total of 80 fieldworkers interviewed 2,777 speakers (1,368 men, 1,409 women) in 1,002 communities. Two-thirds were aged 60 or over and only 10% were younger than 40. The overwhelming majority were white, with fewer than 7% African American, and a very few Native Americans and equally few Asians. The questionnaire contained 41 categories of subject matter and the fieldworkers were required to ask the question in fixed form to ensure comparability in the answers.

An alternative to the broad based survey is to focus on one group. An example is a study of bituminous coal mining vocabulary (Preston 1973). Preston employed a variety of techniques in interviewing a range of coal miners from Southern Illinois to Alabama. He found that the presence of a tape-recorder was inhibiting and that note-taking was a much more successful method of recording information and did not appear to upset his informants. He also found that formal group interviews were most useful for checking the accuracy of suspect forms gathered throughout the area.

Another example of a narrowly based investigation is Macafee's investigation of traditional dialect in Glasgow (Macafee 1994). Macafee was interested in the extent to which traditional Scots vocabulary survived in the context of a modern, industrialized city. She found individual interviews to be unsatisfactory and found that group interviews produced more useful information. Her work describes in detail the difficulties in attempting to carry out a traditional dialect survey in a complex modern city.

The next innovation in dialect surveys involved an instrument that had been available for over 100 years but had not been used for dialect research until recently (Bailey and Ross 1992): the telephone. Improvements in the quality of the signal and a reduction in the cost of calls made the use of the telephone viable for linguistic research. Labov developed the Telsur survey in 1992 to obtain information on speakers in all of English-speaking North America, including Canada. Using telephone calls, thirteen interviewers recorded 805 interviews with speakers ranging from age 12 to 89 during the period 1992 to 2001 (Labov, Ash, and Boberg 2006). The majority (458) were between the ages of 20 to 50 years old. In comparison with the general population in the cities surveyed the Middle Class is over-represented and the Upper Working Class is under-represented. This was a deliberate policy because the aim of the survey was to track ongoing sound changes and the target population emphasized younger women who are often in the vanguard of linguistic change. A search of local telephone directories looked for names marked by the most prominent national ancestry groups. This resulted in a sample that included 217 speakers identified as having a German background as compared with 180 from the United Kingdom. An overwhelming majority of the speakers identified themselves as part of the Anglo-American population with only 45 being African-American and 13 Hispanic. The survey investigated urbanized areas with a population of over 50,000 plus a number of smaller cities to provide a more even coverage. By concentrating on cities of over 50,000 the survey was able to cover the entire continent in less than 10 years.

After information about the speaker's background had been obtained, the questionnaire began with questions about the city and quickly moved on to specific linguistic items and concluded with more personal information and a request to read a wordlist that would be mailed to the speaker. The Telsur questionnaire was consequently quite different from the kind used in dialect surveys that sampled rural areas.

13.3 Postal Questionnaires

The use of postal questionnaires has been a frequent feature of explorations of regional variation in Canadian English. Written dialect surveys are also the focus of Chapter 15 of this Handbook. In 1972, the Survey of Canadian English distributed questionnaires to ninth-grade students (14 to 15 year olds) and their parents and received 14,228 responses (Warkentyne 1971). The questionnaire included items on pronunciation, grammatical usage, vocabulary, and spelling. In 1999, the North American Regional Vocabulary project distributed a questionnaire to every region of Canada and the entire United States (Boberg 2005). A total of approximately 6,000 responses were received. The questionnaire contained 53 items, each question asking for a choice from a set of alternatives. Question 29 asks for the word for “a long, low piece of furniture with sliding doors or drawers, for storing dishes, etc., and provided as choices: *buffet/cabinet/credenza/cupboard/hutch/server/sideboard*.“

A slightly different approach was taken for the Dialect Topography of Canada project (Chambers 1994). Chambers argues in favor of the use of a postal questionnaire on various grounds, one of them being the adverse effect of an outsider interviewing local inhabitants as illustrated by Douglas-Cowie (1978). Chambers developed a questionnaire with 81 items to be filled in, dealing with pronunciation, morphology, syntax, general vocabulary, special vocabulary, and usage. Instead of mailing the questionnaires to informants, Chambers chose two kinds of institutions in which to collect responses: community colleges and retirement homes. This provided an age range from 14 to over 80, although the majority were in the 20-60 age range. A total of 1929 questionnaires was distributed with a 53% return rate. The coverage extended to the major population centers from east to west but the vast central regions of the country were not sampled.

13.4 Internet Surveys

The next use of new technology will obviously be the internet. In the Ottawa Intensifier Project (Van Herk 2008) students from large lower-level undergraduate classes collected linguistic samples of intensifier use from the internet and coded them for linguistic and social factors. This kind of work now continues at the Memorial University of Newfoundland. In addition to providing large amounts of data in digital form, this kind of project has obvious pedagogical benefits for the students. Another project at the Memorial University of Newfoundland is the interactive online Dialect Atlas of Newfoundland and Labrador (Clarke 2014). While the primary purpose of the online atlas is to diffuse information to the wider public, visitors to the site are also encouraged, via comment forms, to contribute information on their own usage of the linguistic features investigated, as well as general information on their regional background and age level. In addition, a frequently updated “Featured Word” section solicits users’ lexical and semantic input relative to regional usage.

Another innovative use of technology was the BBC’s *Voices* project where 51 radio researchers were sent out to find groups of people in their area who might provide responses to a survey questionnaire (Upton and Davies 2013). A total of 321 interviewing sessions were held with 1,200 participants from a wide range of backgrounds. Sessions lasted from 45 minutes to 2 hours. All the sessions were recorded on tape. The focus of the discussions was on a set of 36 lexical areas and participants were encouraged to send their examples to a dedicated BBC website. This resulted in a total of 734,000 responses from about 84,000 participants. Respondents supplied information age, gender, and location. For the online respondents the average age is about 33, over 60% are younger than 30, and 57.3% are female (Wieling, Upton, and Thompson 2013).

It is too soon to say what impact the internet will have on dialect research. The social media such as Facebook and Twitter are clearly having a significant impact on how (many)

people communicate. This presents an opportunity to dialect researchers but the ethical questions are unclear (D'Arcy and Young 2012). It is also unlikely that the internet will provide as much access to the older, conservative speakers who have been the main target of dialect geographers.

13.5 Conclusion

In dialect surveys it is probably not true that you get what you pay for. The results will depend upon the match between the goals of the survey and the methods employed. Obviously, if you can piggy-back on a survey that is being conducted for another purpose, that will save money and time, but it will constrain the project in ways that may not be ideal. Even when funds are available to support an extensive investigation, the danger is that a narrow conception of the situation to be studied may result in too large a data set that is not sufficiently useful. Large scale projects such as Kurath's study of New England and the Survey of English Dialects produced a great deal of information and it had been suggested that these surveys would provide the basis for many interpretive studies but this has hardly happened. This situation may change with newly developed computerized methods of analyzing the data (Kretzschmar and Schneider 1996).

A slightly jaundiced view of many research projects suggests that the danger is overkill and that too many resources are expended on simple questions that could have been answered more economically. Narrowly focused projects such as the Dialect Topography of Canada can achieve useful results on a modest budget. One problem with the use of carefully administered questionnaires, however, is that you get answers about the distribution of items that you know about in advance. There is little chance that you will uncover uses that you did not expect. A small scale sociolinguistic project in the Scottish town of Ayr (Macaulay 1991) recording free conversations with a range of speakers revealed syntactic and discourse features that had not been observed before in dialect surveys. This kind of approach requires tape-recording the sessions and transcribing the conversations in a form that can be mechanically searched.

A more basic problem, however, is the notion of "dialect" itself (Kretzschmar 1998; Macaulay 2002). The aim in many surveys has been to provide information on which to draw an isogloss for a particular word or sound and then to group isoglosses together in order to determine dialect boundaries. Unfortunately, it has proved difficult to provide a well-defined account of coherent dialects in this way (Macaulay 1985) and the notion of isogloss may have passed its sell-by-date. The alternative approach that is now being adopted by some investigators is to choose a location and attempt to characterize the language within that area, for example, Ocracoke (Wolfram and Schilling-Estes 1997), Washington, DC (Schilling 2013), and Pittsburgh (Johnstone 2013). This is a more economical approach to the question of regional variation and is likely to be more widely adopted in the future. As with the possibilities for hunting tigers, the prospects of large-scale surveys may be diminishing.

REFERENCES

- Atwood, E. Bagby. 1971. The methods of American dialectology. In Harold B. Allen and Gary N. Underwood (eds.) *Readings in American Dialectology*, 5–35. New York: Appleton-Century-Crofts.
- Bailey, Guy and Cynthia Bernstein. 1998. Methodology of a phonological survey of Texas. *Journal of English Linguistics* 22: 6–12.
- Bailey, Guy and Margie Dyer. 1992. An approach to sampling in dialectology. *American Speech* 67: 3–20.

- Baxter, Leslie A. and Earl Babbie. 2004. *The Basics of Communication Research*. Belmont, CA: Wadsworth/Thomson Learning.
- Boberg, Charles. 2005. The North American Regional Survey: New variables and methods in the study of North American English. *American Speech* 80: 22–60.
- Chambers, J.K. 1994. An introduction to dialect topography, *English World Wide* 15: 35–53.
- Chambers, J.K. and Peter Trudgill. 1980. *Dialectology*. Cambridge: Cambridge University Press.
- Clarke, Sandra. 2013. Adapting legacy regional language materials to an interactive online format: the Dialect Atlas of Newfoundland and Labrador English. In Alena Barysevich, Alexandra D'Arcy, and David Heap (eds.): *Proceedings of Methods XIV: Papers from the Fourteenth International Conference On Methods in Dialectology 2011*. Frankfurt: Peter Lang.
- Cochrane, William G., Frederick Mosteller, and John W. Tukey. 1954. Principles of sampling. *Journal of the American Statistical Association* 49: 1–12.
- D'Arcy, Alexandra and Taylor Marie Young. 2012. Ethics and social media: Implications for sociolinguistics in the networked public. *Journal of Sociolinguistics* 16: 532–546.
- Davis, Lawrence M. 1990. *Statistics in Dialectology*. Tuscaloosa, Ala: University of Alabama Press.
- Douglas-Cowie, Ellen. 1978. Linguistic code-switching in a Northern Irish village: social interaction and social ambition. In Peter Trudgill (ed.) *Sociolinguistic Patterns in British English*, 37–51. London: Arnold.
- Eckert, Penelope and Etienne Wenger. 2005. Communities of practice in sociolinguistics. *Journal of Sociolinguistics* 9: 582–589.
- Ellis, Alexander. 1889. *On Early English Pronunciation: Part V*. London: Trübner.
- Fasold, Ralph W. 1972. *Tense Marking in Black English: A Linguistic and Social Analysis*. Arlington, VA: Center for Applied Linguistics.
- Gilliéron, Jules (ed.). 1902–1910. *Atlas linguistique de la France*. Paris: Champion.
- Horvath, Barbara M. and Ronald J. Horvath. 2001. A multilocality study of a sound change in progress: The case of /l/ vocalization in New Zealand and Australian English. *Language Variation and Change* 13: 37–57.
- Johnstone, Barbara. 2013. *Speaking Pittsburghese: The Story of a Dialect*. New York: Oxford University Press.
- Klepsch, Alfred. 2013. Wie entstand der *Sprachatlas von Mittelfranken?* Planung,
- Exploration und Publikation. In: Horst Haider Munske and Andrea Mathussek (eds.) *Handbuch zum Sprachatlas von Mittelfranken. Dokumentation und Auswertung*. 19–37. (*Schriften zum Bayerischen Sprachatlas*, Vol.9) Heidelberg: Universitätsverlag Winter.
- Kretzschmar, William A., Jr. 1998. Analytical procedures and three technical types of dialect. In Michael B. Montgomery and Thomas A Nunnally (eds.) *From the Gulf States and Beyond: The Legacy of Lee Pederson and LAGS*, 167–185. Tuscaloosa: University of Alabama Press.
- Kretzschmar, William A., Jr. and Edgar W. Schneider. 1996. *An Introduction to Quantitative Analysis of Linguistic Survey Data: An Atlas by the Numbers*. Thousand Oaks, CA: Sage.
- Kretzschmar, William A., Jr., Virginia G. McDavid, Theodore K. Lerud, and Ellen Johnson. 1993. *Handbook of the Linguistic Atlas of the Middle and South Atlantic States*. Chicago: University of Chicago Press.
- Kurath, Hans, Marcus L. Hansen, Julia Bloch, and Bernard Bloch. 1939. *Handbook of the Linguistic Geography of New England*. Providence, R.I.: Brown University.
- Labov, William. 1966. *The Social Stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.
- Linn, Michael D. 1983. Informant selection in dialectology. *American Speech* 58: 225–243.
- Labov, William, Sharon Ash, and Charles Boberg. 2006. *The Atlas of North American English: Phonetics, Phonology and Sound Change*. Berlin: De Gruyter.
- Macaulay, Ronald K.S. 1977a. *Language, Social Class, and Education: A Glasgow study*. Edinburgh: Edinburgh University Press.
- Macaulay, Ronald K.S. 1977b. Review of *The Linguistic Atlas of Scotland*, vol.1. *Language* 53: 224–228.
- Macaulay, Ronald K.S. 1985. Linguistic maps: Visual aid or abstract art? In John M. Kirk, Stewart Sanderson, and J.D.A. Widdowson (eds.) *Studies in Linguistic Geography*. 172–186. London: Croom Helm.
- Macaulay, Ronald K.S. 1991. *Locating Dialect in Discourse: The Language of Honest Men and Bonnie Lasses in Ayr*. New York: Oxford University Press.
- Macaulay, Ronald K.S. 2002. I'm off to Philadelphia in the morning: A Scotsman looks at dialect in America. *American Speech* 77: 227–241.
- Macafee, Caroline. 1994. *Traditional Dialect in the Modern World: A Glasgow Case Study*. Frankfurt am Main: Peter Lang.

- Mather, J.Y. and H.H. Speitel (eds.) *The Linguistic Atlas of Scotland: Scots Section*, 2 vols. 1975, 1977. London: Croom Helm.
- McDavid, Raven I. Jr. 1953. Review of *A Questionnaire for a Linguistic Atlas of England* by Eugen Dieth and Harold W. Orton. *Journal of English and Germanic Philology* 52: 387–391.
- McDavid, Raven I. Jr. 1966. Sense and nonsense about American dialects. *PMLA* 82: 7–17.
- McDavid, Raven I. Jr. 1971. Planning the grid. *American Speech* 46: 9–26.
- McDavid, Raven I. Jr. 1980. *Varieties of American English: Essays by Raven I. McDavid, Jr.* (Selected and edited by Anwar S. Dil). Stanford: Stanford University Press.
- McIntosh, Angus. 1952. *An Introduction to a Survey of Scottish Dialects*. Edinburgh: Thomas Nelson and Sons.
- Milroy, Lesley. 1987. *Language and Social Networks*, 2nd ed. Oxford: Blackwell.
- Orton, Harold. 1962. *Survey of English Dialects: Introduction*. Leeds: E.J. Arnold and Sons.
- Orton, Harold and Eugen Dieth. 1951. The new survey of dialectal English. In C.L. Wrenn and G. Bullough (eds.) *English Studies Today*, 63–73. London: Oxford University Press.
- Orton, Harold and Nathalia Wright. 1974. *A Word Geography of England*. London: Seminar Press.
- Pop, Sever. 1950. *La Dialectologie: Aperçu historique et méthodes d'enquêtes linguistiques*. Louvain: Bibliothèque de l'université.
- Preston, Dennis R. 1973. *Bituminous Coal Mining Vocabulary of the Eastern United States*. Alabama: University of Alabama Press.
- Trudgill, Peter. 1983. *On Dialect*. Oxford: Blackwell.
- Schilling, Natalie. 2013. *Sociolinguistic Fieldwork*. Cambridge: Cambridge University Press.
- Upton, Clive and Bethan Davies. 2013. *Analyzing 21st Century British English: Conceptual and Methodological Aspects of the 'Voices' Project*. London: Routledge.
- Van Herk, Gerard. 2008. A very big class project: collaborative language research in large undergraduate classes. *American Speech* 83: 222–230.
- Warkentyne, H.J. 1971. Contemporary Canadian English: A report of the survey of Canadian English. *American Speech*. 46: 193–199.
- Wieling, Martijn, Clive Upton, and Ann Thompson. 2013. Analyzing the BBC Voices data: Contemporary English dialect areas and their characteristic lexical variants. *Literary and Linguistic Computing* 28: 1–11.
- Wolfram, Walt and Natalie Schilling-Estes. 1997. *Hoi Toide on the Outer Banks: The story of the Ocracoke Brogue*. Chapel Hill, NC: University of North Carolina Press.
- Wright, Joseph. 1898–1905. *The English Dialect Dictionary, being the complete vocabulary of all dialect words still in use, or known to have been in use during the last two hundred years*. London: Henry Frowde.

14 The Dialect Questionnaire

CARMEN LLAMAS

14.1 Introduction

Since the nineteenth century, when the systematic, large-scale study of dialects began, the questionnaire has been the fundamental instrument of surveys and studies that seek to analyze and document variation in accents and dialects. From the beginning, data have been collected from informants within the framework of a questionnaire to enable a common core of linguistic data to be collected, even when fieldworkers and circumstances differ. The tool of the questionnaire comes in a multitude of shapes and sizes, which depend largely upon the focus of the study. Data can be elicited directly or indirectly; formally or informally; through open-ended or closed questions; in a structured or an un-/semi-structured way; under tightly controlled, experimental conditions or in a much freer, looser way; on lexical, grammatical, or phonetic/phonological levels of variation, and so on. Techniques for administering questionnaires are similarly varied: postal, telephone, and online questionnaires have all featured as the available or the preferred option for the administration of the questionnaire in studies both large and small. The face-to-face interaction of fieldworker and interviewee is also the favored technique of administration of the questionnaire in a very large number of studies.

In general terms, a questionnaire consists of a series of prompts, usually questions, used to elicit some kind of response or measure. This is largely true in dialect research, but the type and design of prompt or question is highly variable and depends, to a very great extent, on the nature of the data to be elicited and on how the questionnaire is to be administered. A dialect questionnaire can be based around the elicitation of single target words or grammatical constructions, or it can be designed to prompt extended stretches of speech. It can be used to assess perceptions of variation between accents and speakers, or it can be used simply as a tool to collect unscripted speech for linguistic analysis. In the latter case, such speech can be content-relevant, in that the information given or attitude expressed may play a role in the interpretation of production differences, or it can be content-irrelevant, wherein the only thing of interest is the linguistic/phonetic content of the speech produced.

Despite the considerable variation in form, the core function of the questionnaire remains the same—that is, to elicit comparable data from informants. In this chapter we will consider how this is achieved and what factors bear on the content, design, and

administration of a dialect questionnaire. To begin, we take an historical perspective through examination of the questionnaires used in some early traditional dialectological surveys of the nineteenth and twentieth centuries, focusing particularly on the *Survey of English Dialects* (Orton and Dieth 1962–1971). We then turn to an examination of some recent approaches to the construction of questionnaires designed to elicit information on how one speaker's dialect may differ from another's, both in how it is produced and how it is perceived.

14.2 Historical Approaches and Their Limitations

The first large-scale dialect survey undertaken in Germany was begun in 1876 by Georg Wenker and resulted in the *Sprachatlas des Deutschen Reichs*. Wenker's interest lay in the phonetic differences between dialects. In order to investigate this, he compiled a questionnaire consisting of 40 sentences that were to be translated into the local dialect by his informants.¹ The sentences were written in Standard German, and Wenker's informants, who were largely schoolmasters, were asked to provide equivalent sentences in their local dialect.

Each sentence offered various possibilities for a dialectal variant to be used (see example sentence below).

1. Im Winter fliegen die trocknen Blätter durch die Luft herum (*In winter the dry leaves fly around through the air*).

By 1887, Wenker had sent out around 50,000 questionnaires to school districts and had managed to achieve extensive coverage across the entire nation. He received close to 45,000 completed responses. The considerable geographical coverage and vast amount of data amassed through these responses demonstrate the major benefits of the technique of using a postal questionnaire in large-scale surveys, such as Wenker's. The lack of control over the accuracy and consistency of responses, however, arguably counteracts these advantages. (For details on the outputs from Wenker's large-scale survey and implications of the findings, see Chapter 27, *Dialects of German, Dutch and the Scandinavian Languages*; for further discussion of the postal questionnaire technique, see Chapter 15, *Written Dialect Surveys*.) The obvious way to achieve some level of accuracy and consistency in the recording of the data is to send trained linguists and phoneticians out into the field to undertake face-to-face interviews with informants (detailed discussion of field-work methods can be found in Chapter 16, *Dialectological Field Interviews*). This, indeed, was the technique used to administer the questionnaire in the majority of large-scale dialect surveys that followed, although notable exceptions, for example, the initial stages of the *Survey of Scottish Dialects* begun in 1952 by Angus McIntosh, still made use of the technique of the postal questionnaire.

Weigand, who undertook the first survey of the Romanian language, made systematic use of a questionnaire consisting of a list of 114 words. The resultant *Linguistischer Atlas des daco-rumänischen Sprachgebietes* (1909) was the product of Weigand's collection of the material himself, which had the obvious advantage of ensuring a level of consistency. Weigand's questionnaire, and hence his phonetic enquiry, was limited in scale, however, and not all sounds and all environments were covered in the 114 words used.

It was Jules Gilliéron's contemporaneous *Atlas linguistique de la France* (1902–1910), which served as a model for future works. In terms of size and scope, the dialect questionnaire used for Gilliéron's survey was much greater than those which had gone before.

It not only ran to over 1,900 items, but, unlike others, it was designed to elicit data concerning variation in morphology, syntax, and vocabulary, as well as pronunciation. As with Wenker's survey, Gilliéron used direct questioning, for example, *What do you call a cup?* In terms of the vocabulary investigated, terminology for common everyday objects was used as it was felt that a higher number of variants would be available for more rather than less familiar objects. Variation in vocabulary was not limited to this, however, and a number of neologisms were included to ascertain their spread outward from Paris. Gilliéron was an influential figure and, indeed, two of his students, Karl Jaberg and Jakob Jud, went on to produce the *Sprach- und Sachatlas Italiens und der Südschweiz*, which appeared from 1928 to 1940.

In the first half of the twentieth century, many other surveys were begun (e.g., in the United States and Canada, Italy, Spain). Fieldwork for the *Survey of English Dialects* (SED) was undertaken between 1950–1961. The length of time taken to complete the fieldwork reflects the large-scale geographical coverage of the survey (313 localities across England) and the lengthy data elicitation tool, which was administered in a face-to-face interview. After a series of pilot studies, the questionnaire was settled on as the "fundamental instrument" (Orton 1962, 15) to be used in the survey.

The design of the questionnaire was planned with reference to existing dictionaries and glossaries, and drew particularly on Joseph Wright's (1898–1905) *English Dialect Dictionary*. Possible notion words and grammatical constructions were collected with a focus on prompts that were likely to yield a large variety of dialectal forms. The questionnaire was worked on for five years. Extensive piloting of the questionnaire took place and the second version, which was ready by the end of summer of 1948, was trialed by various fieldworkers in six different counties. In its final form (which was the sixth version), the dialect questionnaire used in the SED consisted of 1,332 questions. Of these, 387 elicited a response of phonological interest, 128 were morphological, 77 were syntactic, and 730 were lexical. The majority of the questions eliciting a morphological or syntactic response were assembled in the last two of the nine books of questions.

The questionnaire is published in its entirety in *Survey of English Dialects: An Introduction* (Orton 1962). Taking approximately 24 hours to complete (Chambers and Trudgill 1980, 27), the interview was administered over several days, often with two or three different informants from the same locality completing different parts. The questions were grouped together by subject matter, rather than arranged alphabetically or randomly. Johnston (1985, 83) argues that grouping questions by subject matter allows for a level of spontaneity in the responses. By asking questions in a random or alphabetical order, the informant's attention is drawn to the language elicited, whereas when questions are grouped by subject, the informant's attention is called to the grouping.

The nine books of the SED questionnaire were assembled around semantic fields, reflecting, to some extent, the focus on rural localities. As dialect was felt to be "best preserved by the farming community" (Dieth and Orton 1952, v), the questionnaire was designed with this in mind, centering on husbandry, home life, and nature. Specialized industries, such as fishing and mining, were felt to be too technical to be included. The nine books of the questionnaire were: I. The Farm; II. Farming; III. Animals; IV. Nature; V. The House and Housekeeping; VI. The Human Body; VII. Numbers, Time and Weather; VIII. Social Activities; and IX. States, Actions Relations.

The questionnaire was formal, that is questions were scripted and standardized. It used indirect questioning, for example, *What do you call this? (holding a cup)*, as opposed to direct questioning, *What do you call a cup?* The use of indirect questions was adopted in reaction to the problems associated with direct questioning, as used in some of the surveys that had come before. The direct questioning technique involves a translation from the standard variant. The influence of the standard may, therefore, affect the

response given. Also, direct questioning of this type rests on the assumption that the informant is bidialectal and is aware of the difference between the standard and the dialect. This is not always the case. One of the advantages of using direct questioning, however, is that it is much quicker, and the fieldworker working on Gilliéron's survey of French dialect, Edmond Edmont, was probably able to complete one interview per day (Chambers and Trudgill 1980, 27), unlike the several days needed for the fieldworkers on the SED.

The basic types of questions used in the questionnaire were *naming* and *completing*. Naming questions usually make use of pointing to objects or visual stimuli and begin, *What do you call this?* This type of question was considered by Dieth and Orton to be "the simplest and the best" (1952, vi). Subtypes of naming questions involve *talking* questions, which elicit more than one word, for example *How do you mark your sheep?*, and *reverse* questions, which ask for the meaning or meanings the informant attaches to a word of multiple meanings, for example, *What do you mean by broth?* The talking question was claimed to be the ideal way of eliciting dialectal material, but was only used sparingly in the questionnaire as it "produces results too slowly for our purposes" (Dieth and Orton 1952, vi). For completing questions, the fieldworker supplies a blank for the informant to fill in, for example *You sweeten your tea with...* to which the informant completes with his/her word for "sugar." A subtype of completing questions, *converting* questions, was also used. This type of question was used, for example, to obtain the tenses of irregular verbs where the target verb in the present tense is elicited. The fieldworker introduces the temporal adverbs *yesterday*, *always*, and so on, and the informant converts his/her verb accordingly. The majority of the questions used in the SED questionnaire were naming questions. Completing questions were the next most frequent, followed by conversion questions and other types (see examples of all types of questions used in Table 14.1 below).

The way the SED questionnaire was set out was clear and easily navigable for the fieldworker. The keywords were printed in bold, as in the examples given in Table 14.1, and a book, page and question number reference system was used to index the individual target forms. The fieldworkers noted responses to the questions on special quarto sheets divided in half vertically. Narrow phonetic transcriptions were made of responses and noted on the left-hand side of the recording sheets. The survey questions naturally led to passages of spontaneous conversation. During this, as many expressions of interest as was feasible to collect were noted on the right-hand side of the fieldworkers' recording sheets and labeled *Incidental Material*. Tape recordings of unscripted speech of suitable informants were also included in the survey, although it was prohibitively expensive to re-record all of these onto discs; only a selection were chosen by the fieldworker. Responses to the questions of the questionnaire were included in the volumes of *Basic Materials* to come from the survey. Included along with the response were abbreviated notes on whether the response had been elicited under strong pressure, whether the fieldworker had suggested the form or the word, or whether it had not been asked, known or found. This demonstrates some of the problems encountered with the use of a lengthy questionnaire, which uses a series of questions asked in the context of a direct, formal questionnaire. The consequences of possible fatigue and waning attention and interest on the part of the informant, combined with the restricted type of speech style elicited constitute major disadvantages associated with this type of dialect questionnaire. The amount of comparable data collected by such surveys is enormous, however, and analysis of the SED *Basic Materials* (responses to the questions) and *Incidental Material* (that produced in spontaneous, unscripted interaction) has yet to be exhausted to this day.

Table 14.1 Examples of the different question types used in the Survey of English Dialects questionnaire (Orton 1962).

Naming questions

Book IV. Nature; 10. Trees, Bushes; Question 13.

½ What do you call this plant? It grows on moist ground; its leaves are long and feathery.
Fern*.

Book II. Farming; 1. The Land; Question 1.

What do you call land that you have ploughed but that you leave unsown for some time?
Fallow-land.

Talking questions

Book V. The House and Housekeeping; 5. The Dairy; Question 4.

What can you make from milk? **Butter***, **cheese***.

Book IV. Nature; 10. Trees, Bushes; Question 1.

What trees have you round here? **Birch***, **oak***, **elm***, **elder**, **willow**.

Reverse questions

Book V. The House and Housekeeping; 6. Baking; Question 9.

What do you mean by **loaf***?

Book VI. The Human Body; 14. Clothing; Question 1

What do you mean by a **bonnet**?

Completing questions

Book VII. Numbers, Time and Weather; 2. Ordinals; Question 14.

In this room there's just you and me, no third person; in other words, just **we two**†

Book IV. Nature; 11. Berries, Fruits; Question 4

If you know a berry will kill you if you eat it, you say it is **poisonous**†

Converting questions

Book IX. State, Actions, Relations; 3. Verbs: Irregular; Question 7

John Smith had the chance to go to college, but didn't **take*** it.

But his brother was given the chance too and he gladly **took**† it.

If their sister had had the same chance, she certainly would have **taken**† it.

In fact, she never misses any chance; every chance she gets, she **takes**†.

½ means that a picture is to be shown, if necessary.

* denotes that the word has been inserted for its phonological importance and should therefore be obtained by the fieldworker

† denotes that the word has been inserted for its morphological importance

‡ denotes that the word or phrase has been inserted for its syntactical importance.

14.3 Contemporary Approaches

The move away from a focus on variation across geographical space toward investigation of socially stratified variation within localities entailed changes and innovations in the design of the questionnaire. Although some work which was begun in the traditional dialectological tradition incorporated social factors in speaker sampling, for example, the Linguistic Atlas of the United States and Canada (see further Chapter 26, *Dialects of North American English*), methods for eliciting samples of speech were to change radically as the focus moved toward intra-variety and intra-speaker differences. While use of forms was approached in a

categorical manner in a given locality in traditional dialectological surveys, the move toward a more urban, social dialectology saw the focus of interest fall on the examination of differences in frequency of variant usage among speakers of different demographic characteristics who, nonetheless, belonged to the same speech community. With such an aim, very short, often one-word responses to precisely targeted prompts would not elicit the data required to examine the frequency of use of a form in the contexts where it could have been used. Neither would it allow for the investigation of different frequencies in the use of forms in the same speaker in different contexts. Longer stretches of audio-recorded speech are required for such an undertaking. The move away from aims associated with traditional dialectological surveys saw a vast increase in novel ways to investigate language variation and change. Different types of questionnaires were therefore required for the diverse types of studies undertaken.

14.3.1 Phonological Variation

Many of the studies of dialectal variation, perhaps the majority, concentrate on phonological differences. The greater part of these deal with variation at the segmental level. For most the goal is to elicit extended stretches of speech in a style that is as unmonitored as possible, given the effects of the act of recording the speech. The section of a questionnaire designed to elicit the extended, spontaneous speech for analysis can be of two types, those for which the topical content is irrelevant, and those for which the responses elicited are used to interpret findings at the phonological level. Questions in the former can cover any topic that is deemed by the researcher to be one to which the informant will respond without offence, without reluctance to speak in a recorded interview, and with an opinion or narrative, which will produce an extended sample of speech. An early example of one such topic is Labov's classic question as used in his study of Lower East Side, New York: "*Have you ever been in a situation where you thought you were in serious danger of being killed—where you thought to yourself, 'This is it'?*" (1972, 93). The belief that the recounting of such a narrative would result in an unmonitored speech style approximating the vernacular was later questioned along with the suitability and appropriateness of the question to other contexts (e.g., the study of Belfast English by the Milroys (J. Milroy 1992, L. Milroy 1987) and the study of Norwich English by Trudgill (1974)). However, topics or modules of conversation may be based around themes of personal interests, and many examples can be found of questions used to elicit extended stretches of unmonitored speech from informants (early examples taken from Wolfram and Fasold (1974) include, "What are your favorite TV programs? Describe a recent program"; "What is your favorite movie of all time? What happens?"; "If someone came up to you and said, 'Here's all the money in the world', what would you do with it?").

Regarding asking specific questions about language and linguistic variation, some argue that these should appear as part of the interview (or module); "[a] sociolinguistic module should focus on matters of importance for the subjects and have a series of language questions, including specific questions about language variation patterns of the area" (Hazen 2001, 777). Others, however, take a different view and see the elicitation of personal narratives and topics as central to the sociolinguistic interview. Indeed, Tagliamonte suggests that discussion of language should be used almost as a last resort, "[i]f you are going to include a module on Language, always put it at the very end of the interview when your informant has exhausted all the more personal topics" (2006, 39). She further stresses that the content of the interview is of little consequence: "[k]eep in mind that you are not asking questions to get information: you are asking questions that reach the 'real' sentiments of your speakers and which elicit natural, spontaneous speech" (2006, 43). Nonetheless, some questionnaires are constructed precisely to collect information, evaluations, and attitudes via the medium of natural, spontaneous speech. Such questionnaires are designed around the belief that accessing attitudes toward and

perceptions of linguistic variation and the social context in which the speaker operates allows greater insight into motivations for linguistic behavior, the causes of sociolinguistic variation, and the socio-indexical information carried by language forms.

A questionnaire with questions for which the response is relevant to the study is one that takes the informant's understanding of the context, the place, the situation, the sense of "self" and "other" as central concerns in understanding the linguistic patterns uncovered. An example of this would be that used in the *Accent and Identity on the Scottish/English Border* study (see further Watt *et al.* 2014a, 2014b). As the name suggests, the informants' attitudes toward the border and the social categories it divides are of central importance to an understanding of the progression of particular sound changes in the borderland region. As such, the questions used to elicit extended samples of spontaneous speech were of the sort shown in Table 14.2.

In terms of eliciting a sample of unscripted, spontaneous speech that can be analyzed at various levels of linguistic analysis, such questions typically produce a reasonably lengthy and engaged response. Further to this, the opinions given in the responses will allow a profiling of the speakers as individuals who have attitudes, identities, and evaluative responses to the social context in which they find themselves.² Informants have an awareness of the differences in how speakers use language and what this variation signifies. They have a sense of where language boundaries may be drawn (see further discussion of these methods in Chapter 10, *Perceptual Dialectology and Subjective Evaluation of Dialects*). Examination of these opinions allows the researcher to establish correlations between the use of linguistic forms and particular identities, attitudes, or ideologies. Without such opinions, the researcher would have to second guess the orientations, affiliations, and identifications that may be implicated in the motivations for variation and change in progress in the dialects under investigation. Also, without such information, the researcher can only assume that particular linguistic forms of interest carry socio-indexical information and cue particular social identities in the mind of the speaker.

Whether content-irrelevant or content-relevant, the questions used in a questionnaire of this type should always be designed to elicit an extended stretch of spontaneous speech on a topic about which the informant feels comfortable talking.

The unself-conscious, unscripted, spontaneous speech elicited in this way is often compared to read speech from the same speaker to target specific forms and to infer the evaluation of such forms by ascertaining the differences in pronunciation achieved through the effects of the self-monitoring of speech. A typical dialect questionnaire, then, used in a study of production differences at the phonological level would generally contain prompts to elicit extended, spontaneous stretches of speech and materials to elicit read speech. The latter of these could be in the form of reading passages presented to the informant (*The Boy who Cried Wolf* is an example of a reading passage for the elicitation of all phonemic contrasts that occur in English (Deterding 2006)). Along with a reading passage, a bespoke word list is

Table 14.2 Example questions from the Identity Questionnaire used in the *Accent and Identity on the Scottish/English Border* project (Watt *et al.*, 2014a, 2014b).

1. What accent would you say you had, and do you like it?
2. Are there any pronunciations or ways of saying things that you would hear and think, that sounds really Scottish or really English?
3. Where, geographically, would you say people stop talking the same as you and start sounding different?
4. What best describes you? British/Scottish/English/Borderer/none
5. Do you think people speak differently depending on what side of the border they are from?

commonly administered. Such a word list would usually contain target forms either presented as single words or phrases, or embedded into carrier phrases, such as "Say [test word] again," in order to provide consistent parameters for segmentation for acoustic analysis (see further Müller and Ball 2013). The choice of forms used depends entirely on the focus of the particular study. Typically, however, proper nouns and infrequent words are avoided, as they may elicit idiosyncratic or uncertain pronunciations.

Added to this, there are tools that can be included in a dialect questionnaire that fall somewhere between these two standard techniques and allow some control over the elicitation of target forms while producing a certain amount of unscripted, unread speech. A map task (see Figure 14.1 below for an example as used in the *Intonational Variation in English* study (Grabe *et al.* 2001)) ensures that certain forms will be produced several times by the two informants participating in the task as one attempts to guide the other through the route from start to finish. As the participants unknowingly have slightly different maps, an amount of confusion and negotiation ensues which results in unscripted speech.

In addition to production data, informants' judgments on pronunciations can be elicited in a number of ways. Examples of question types taken from Maguire's online *Survey of Accents of English in Britain and Ireland*,³ which began in 2009, include:

1. Minimal Pair tests, for example, "Do you pronounce **fur** and **fair** the same?"
2. Rhyme tests, for example, "Do **mate** and **eight** rhyme for you?"
3. Three-way minimal pair and rhyme test, for example, "How do **merry**, **Murray** and **Mary** rhyme for you?" (Maguire 2009)

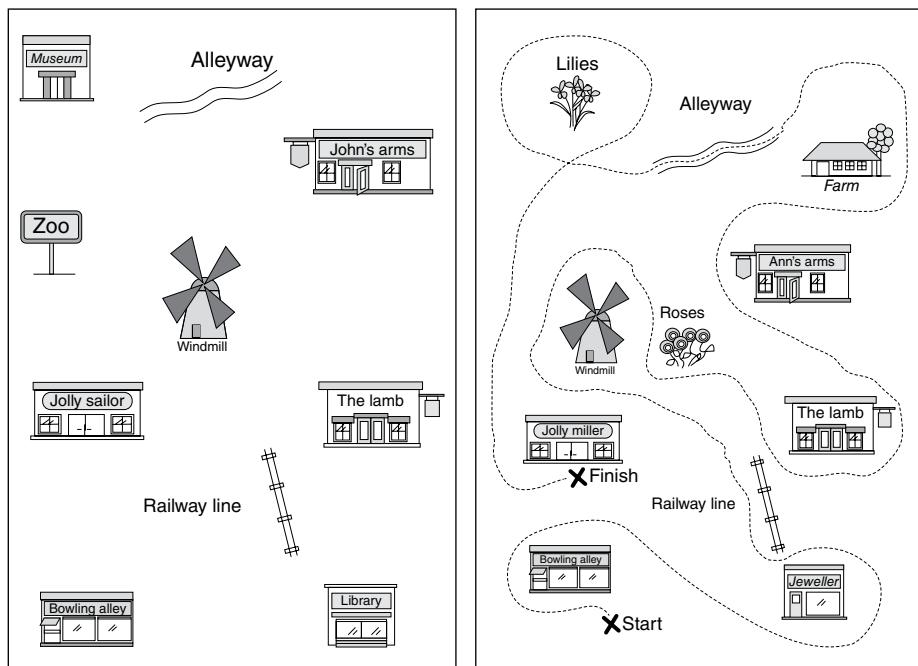


Figure 14.1 Example of map task materials as used in the *Intonational Variation in English* project (Grabe *et al.*, 2001). The map on the right is used by the direction giver, the map on the left is used by the direction receiver.

Findings from such tests can reveal broad patterns of variation across geographical space, including the distribution of phonological distinctions and mergers.

14.3.2 Grammatical Variation

Morphosyntactic features can also be obtained from extended stretches of spontaneous speech, in whatever way they are elicited. However, unlike segmental features at the phonological level of analysis, an extended sample of free speech will not ensure that the target forms in a morphosyntactic analysis will emerge in any quantity, if at all. For the majority of vocalic and consonantal variables, a sufficient number of tokens for analysis can usually be obtained from under an hour of speech as they emerge regularly and predictably. If the focus of the study is on grammatical variation in production, however, some targeting of the questionnaire may be necessary so as to increase the likelihood that particular forms will emerge. For example, questions that elicit narratives of past experiences will be of use if past tense forms are of particular interest.

Rather than a focus on production, many studies of grammatical variation use questionnaires that examine the informants' perceptions of whether or not grammatical structures would be used by them or members of the speech community to which they belong. An example can be seen in Cheshire *et al.*'s (1989) *Survey of British Dialect Grammar*. This large-scale, grammatical survey was administered as a postal questionnaire and sent to 87 schools located throughout Britain. The aim of the survey was to increase the current state of knowledge of the morphology and syntax of British English dialects such that the features that were widespread throughout Britain and those that had a more restricted distribution could be determined. The questionnaire used contained 196 features and was completed by school pupils who worked in groups on one page (the questionnaire was three pages in length). The pupils discussed the forms listed and recorded those that they had heard regularly in the community. A wide range of linguistic phenomena was represented in the questionnaire including modal constructions, imperative forms, present and past tense verb forms, among others. Most of the grammatical features that had been included in the SED were included for comparative purposes. The items on the questionnaire were presented in a clear and graphically appropriate way, with tick boxes clearly presented and illustrations used to engage interest (example items used to assess the distribution of expressions of negation are given in Table 14.3 below).

This type of questionnaire allows statements about the distribution of forms across geographical space to be made. As with some of the early dialectological surveys that utilized a postal questionnaire, the findings from responses reflect informants' perceptions of the use of dialect forms. Actual usage is not addressed through such a method, and the researchers

Table 14.3 Example prompts from Cheshire *et al.* (1989) *Survey of British Dialect Grammar*.

-
1. **Dinna** run too fast
 2. Count on me, I **won't** do **nothing** silly
 3. You **shouldna** go in there!
 4. You've **no** to go in there!
 5. **Anyone** mustn't go in there
 6. My friend broke that, I **never**
 7. No, I **never** broke that
 8. **Will you not** try to mend it – we need an expert
 9. That **ain't** working
 10. That **in't** working
 11. That **ay** working
-

clearly state that such a survey is not “intended to take the place of empirical investigations of language as it is used in daily life” (Cheshire *et al.* 1989, 214). Nonetheless, valuable information can be gathered on the perceived distribution of forms over a wide geographical area through use of such a questionnaire. Furthermore, the questionnaire allows for variation in a given locality to be recorded as more than one form can be documented.

14.3.3 Combining Lexical Variation

Excluding work done toward the compilation of dialect dictionaries (see Chapter 2, *The Dialect Dictionary*, for further discussion), contemporary dialect studies that focus on lexical variation may have a relatively narrow focus on a specific style or register (see, e.g., Millar *et al.* (2014) for a recent example of lexical attrition among Scottish fishing communities). Others may have a broad, large-scale focus making use of online surveys (see, e.g., Vaux and Golder 2003). Many of the latter use indirect questions, such as “What do you call the wheeled contraption in which you carry groceries at the supermarket?” as used in Vaux’s *Harvard Survey* and *The Cambridge Online Survey of World Englishes*.⁴ Although these surveys include data on phonological and morphosyntactic variation, the methods employed are not designed to examine regionally or socially correlated lexical variation in combination with phonological and morphosyntactic variation as produced in unscripted, relatively natural conversation. A combination of this sort presents a methodological problem. The exercising of control over the specific vocabulary used in an interaction results in speech that is no longer wholly spontaneous. The targeting of unscripted, casual speech can, therefore, be seen as incompatible with the elicitation of comparable lexical data.

One elicitation method designed to combine data on lexical variation with elicitation of spontaneous speech was created for use in a proposed survey of variation in spoken British English, the Survey of Regional English (SuRE) (Llamas 1999). The method was designed and first used in a study of Middlesbrough English (Llamas 2001, 2006, 2007), and it has been used in various studies of other varieties of British English including the *Accent and Identity on the Scottish-English Border* project discussed above (see also, Asprey 2007; Burbano-Elizondo 2008; Pichler 2008; Finnegan 2011, for examples). It was also used in an adapted form in the popular *BBC Voices* series on language variation throughout the British Isles (see further, www.bbc.co.uk/voices and also Elmes (2005), Upton and Davies (2013)).

Rather than asking scripted questions to elicit particular lexical items, as in many of the traditional dialectological questionnaires, speech is elicited by the fieldworker through the prompting and encouraging of discussion about words used in the area. The ensuing conversation results in informants discussing and often disagreeing about their perceptions and definitions of dialect words. As well as producing informal conversation from which phonological and grammatical analyses can be undertaken, the conversation produces a mass of information on the lexical data produced. This can include age and gender differences in usage, connotational and collocational information, perceived social variation in usage, supposed etymologies, perceived geographical distribution of usage, knowledge and use differentiation, and attitudinal information on dialect connotations.

In itself, this approach is unusual. In fieldwork, discussion about lexical variation that encourages informants to elaborate on attitudes toward their responses is relatively rare. An exception to the view that direct questions and discussions about lexical items should be avoided comes from Pratt, who suggests “One must get the informants to TALK about the word, use it in different contexts, pass judgement on it, in short, to display knowledge of it” (1983, 153) (for further discussion, see Chapter 10, *Perceptual Dialectology and Subjective Evaluation of Dialects*).

Although much information can be gathered by allowing informants to discuss what they consider local lexical items to be, this alone does not guarantee the collection of data that are comparable. An instrument is still needed to control the specific words elicited in order to

gain data on lexical variation that can be compared across speakers and across varieties. The dialect questionnaire designed for these purposes does not consist of naming or completing questions about dialectal lexical items. Rather, the principal tool of this method is the Sense Relation Network (SRN) sheet (see Figure 14.2 for an example of a completed SRN).

As can be seen in Figure 14.2, networks are designed such that standard "notion words" are connected to subdivisions. The subdivisions, in turn, are connected to the semantic field of the SRN. Under the standard notion word, space is provided for the insertion of a dialectal variant. Visually, the aim is for the SRN to be inviting and engaging, and each of the three core SRNs used in interviews (*People; Feelings, Actions and States; and The Outside World*) is printed in a different colour to maximise visual appeal.

As regards content design, the SRNs are built around semantic fields and, as such, are akin to the grouping of questions by subject matter in the SED questionnaire. The selection of semantic fields and standard notion words used in the SRNs is the result of trialling and revision of the method during which standard notion words producing little or no variation were removed.

Figure 14.2 provides an example of the amount of data on lexical variation elicited through an SRN. As can be seen, this type of questionnaire offers potential for the study of distinctions between dialectal variants, regional slang, national slang and standard colloquialisms, as well as the study of nonstandard orthography.

As well as the three core SRNs, additional or alternative SRNs can be used where relevant to the context of a particular study. For example, recent research on phonological variation in four coal mining villages in the North East of England (Devlin, French, and Llamas, forthcoming) has included an SRN based around the subject of *Mining* in order to examine the effects of conversational topic on the retention of traditional phonetic forms. This has allowed the topic of mining to be highlighted and delimited. Use of traditional forms (for example, realisations of the MOUTH vowel with a raised, close-mid first element) have then been analyzed both in specific lexical items related to the mining industry and also in conversation about the topic of mining (see Drager, Hay, and Walker (2010) and Love and Walker (2013)) for other examples of the effects of conversational topic on patterns of phonological variation). The findings reveal that highly localized, traditional forms, which are being lost in the speech communities, are produced most frequently when the conversation is on the topic of the traditional mining way of life compared with other conversational topics, although not necessarily within the context of specialised vocabulary. Such findings have important implications for our understanding of sound change within a community. Thus, the dialect questionnaire, in its very design, can be a tool that allows us to better understand processes of language change, as well as allowing us to gain a descriptive insight into the dialect of a community.

With advances in technology, we see an increasing use of online surveys to amass information on dialectal variation. Such surveys allow for the collection of very large amounts of data in relatively short periods of time. The questionnaires used in many such surveys are, in many respects, similar to those used in the early large-scale traditional dialectological surveys. We see the use of indirect written questions (as in Vaux and Golder (2003) noted in Section 14.3.4), and also indirect questions using pictorial stimuli. We also see use of minimal pair and rhyme tests to assess the distribution of phonological distinctions and mergers (as in Maguire's (2009) online survey reported in Section 14.3.1). Multiple-choice onscreen answers are presented to the informant, and responses may be mapped dynamically, allowing the respondent to view his or her response among the mapped data. Advances in technology allow the respondent of the dialect questionnaire to engage with it in a more immediate and meaningful way than has been the case in the past. Such technological developments combined with the diversity of approaches to the study of dialectal variation and change ensures that the dialect questionnaire, and how the informant interacts with it, will continue to evolve.

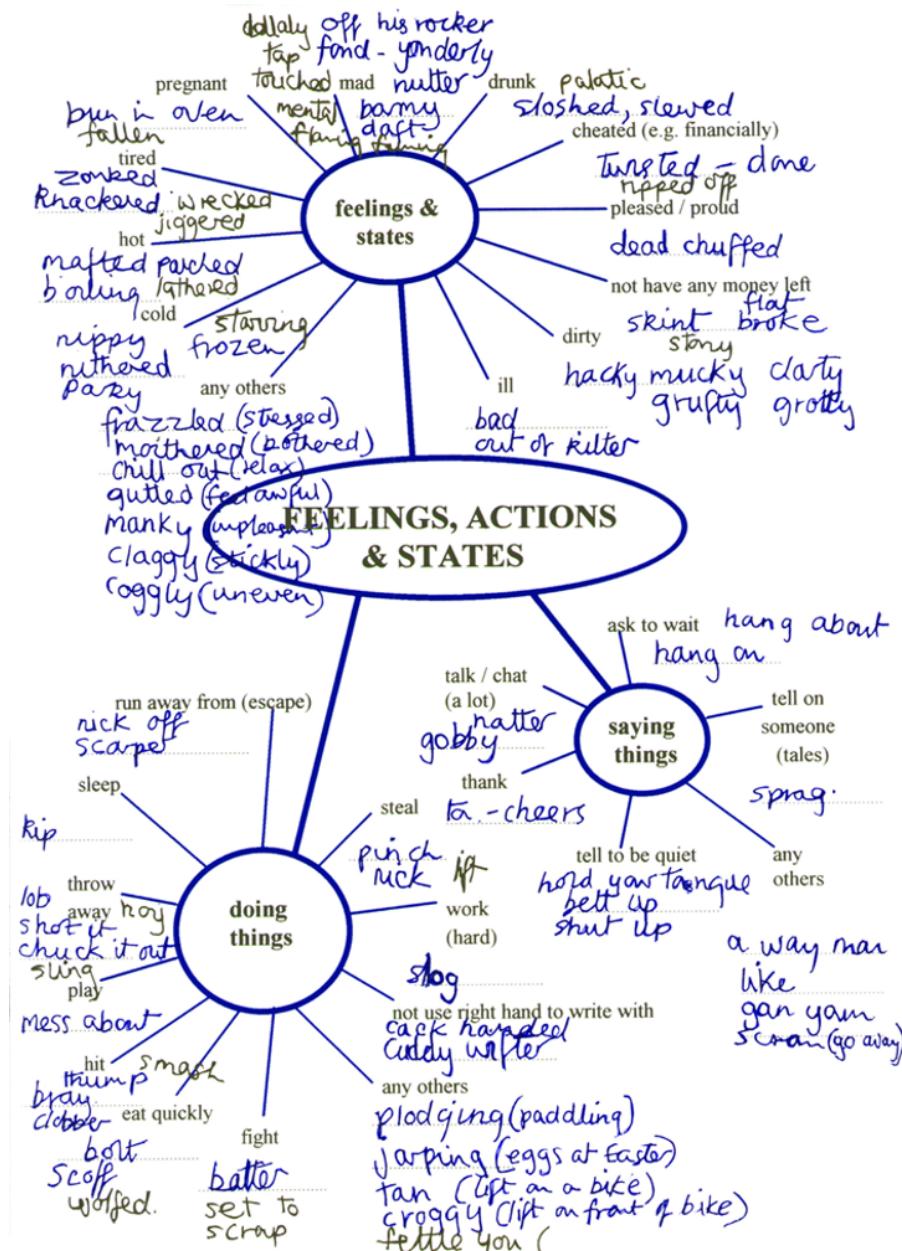


Figure 14.2 Completed Sense Relation Network sheet (one of three) as used by a female speaker of Middlesbrough English (Llamas 1999, 114).

14.4 Conclusion

Back in the early part of the nineteenth century, Jaberg and Jud, editors of the *Sprach- und Sachatlas Italiens und der Südschweiz*, laid out what they saw as the requirements of a good questionnaire:

- a. It should include as rich and characteristic a selection as possible of the linguistic peculiarities of the dialect region to be studied.
- b. It should represent the cultural circumstances of the dialect region to be studied.
- c. It should guarantee at the same time the spontaneity and indigenousness of the answers on the one hand, and their comparability on the other. (Jaberg and Jud 1928, 175, in Francis (1983, 53))

These requirements are perhaps too broad and all-encompassing to have ever been satisfied. Indeed, one could argue that no dialect questionnaire could ever cover all of these elements to an adequate extent. As well as the enduring interest in obtaining a characteristic selection of linguistic features in dialect studies, the need to collect spontaneous speech and the importance of the social and cultural context in which the speakers operate have both taken the dialect questionnaire forward in interesting ways since these early days, as the study of dialects has evolved.

This chapter has given a taste of the variety of dialect questionnaires available for use in studies of language variation. The vast differences in questionnaire type reflect the variety of approaches and concerns in studies of how one person may use language in a different way to another. Nonetheless, whatever the size, shape, or type of dialect study, the questionnaire used will probably never be infallible, as each interview situation has the potential to produce unexpected results and to reveal flaws in the questionnaire design. It is impossible to know in advance all the parameters of variation that will be of interest in a study, and some knowledge and acceptance of the limitations of the questionnaire must form part of its design in order for the survey to proceed. To this extent, then, however far we move from the earliest days of large-scale traditional dialectological studies, there will probably always be truth in Gilliéron's wry observation: "the questionnaire, in order to be clearly the best, ought to be made after the survey" (Gilliéron 1915, cited in Francis 1983, 52).

NOTES

- 1 These "Wenker sentences" are still in use today, and have been made available electronically via the Digital Wenker Atlas (DiWA) (see further, <http://diwa.info/>).
- 2 Further techniques designed to produce quantitative data on speaker attitudes in this project can be seen in Llamas and Watt (2014).
- 3 For further details, see <http://www.lel.ed.ac.uk/~wmaguire/survey/survey.html>.
- 4 For further details, see <http://edparkerjr.mml.cam.ac.uk/BertVaux/DialectsPage2.html>.

REFERENCES

- Asprey, Esther. (2007). *Black Country English and Black Country Identity*. Unpublished PhD thesis, University of Leeds.
- Burbano-Elizondo, Lourdes. (2008). *Language Variation and Identity in Sunderland*. Unpublished PhD thesis, University of Sheffield.
- Chambers, J.K. & Trudgill, Peter. (1980). *Dialectology*. Cambridge: Cambridge University Press.
- Cheshire, Jenny, Edwards, Viv K., & Whittle, Pamela. (1989) Urban British Dialect Grammar: the question of dialect levelling. *English World Wide*. 10(2): 185–225.
- Dieth, Eugen, & Orton, Harold. (1952) *A Questionnaire for a Linguistic Atlas of England*. Leeds: Chorley & Pickersgill.
- Deterding, David. (2006) The North Wind versus a Wolf: short texts for the description and measurements of English pronunciation. *Journal*

- of the International Phonetic Association.* 36: 187–196.
- Devlin, Thomas, French, Peter, & Llamas, Carmen. (forthcoming) Vowel change across time, space and conversational topic: The use of localised features in former mining communities.
- Drager, Katie, Hay, Jennifer, & Walker, Abby. (2010) Pronounced rivalries: Attitudes and speech production. *Tē Reo.* 53: 27–53.
- Elmes, Simon. (2005) *Talking for Britain: A Journey Through the Nation's Dialects*. London: Penguin Books.
- Finnegan, Katie. (2011) *Phonological Variation and Local Identity in Sheffield*. Unpublished PhD thesis, University of Sheffield.
- Francis, W. Nelson. (1983) *Dialectology: an introduction*. Harlow: Longman.
- Gilliéron, Jules, & Edmont, Edmond. (1902–1910) *Atlas Linguistique de la France*. Bologna/Paris: Forni/Champion.
- Grabe, Esther, Post, Brechtje, & Nolan, Francis. (2001) *The IViE Corpus*. Department of Linguistics, University of Cambridge (ESRC grants R000237145 and RES000230149).
- Hazen, Kirk. (2001) Field Methods in Modern Dialect and Variation Studies. In Mesthrie, Rajend. (ed.) *Concise Encyclopedia of Sociolinguistics*, pp. 776–779. Oxford: Elsevier.
- Jaberg, Karl, & Jud, Jakob. (1928–1940) *Sprach- und Sachatlas Italiens der Südschweiz*. (AIS) 8 vols. Zofingen: Ringier.
- Johnston, Paul A. (1985) Linguistic atlases and sociolinguistics. In Kirk, John M., Sanderson, Stewart & Widdowson, J. D.A. (eds.) *Studies in Linguistic Geography*. Beckenham: Croom Helm, pp. 81–93.
- Labov, William. (1972) *Sociolinguistic Patterns*. Philadelphia: University of Philadelphia Press.
- Llamas, Carmen. (1999) A new methodology: data elicitation for social and regional language variation studies. *Leeds Working Papers in Linguistics and Phonetics* 7: 95–118.
- Llamas, Carmen. (2001) *Language Variation and Innovation in Teesside English*. Unpublished PhD thesis, University of Leeds.
- Llamas, Carmen. (2006) Shifting identities and orientations in a border town. In Omoniyi, Tope, & White, Goodith (eds.) *Sociolinguistics of Identity*. London: Continuum, pp. 92–112.
- Llamas, Carmen. (2007) 'A place between places': language and identities in a border town' *Language in Society*. 36(4): 579–604.
- Llamas, Carmen, & Watt, Dominic. (2014) Scottish, English, British?: Innovations in attitude measurements. *Language and Linguistic Compass*. 8(11): 610–617.
- Love, Jessica, & Walker, Abby. (2013) Football versus football: Effect of topic on /r/ realization in American and English sports fans. *Language and Speech*. 56(4): 443–460.
- Maguire, Warren. (2009) A Survey of Accents of English in Britain and Ireland. Edinburgh: University of Edinburgh School of Philosophy, Psychology and Language Sciences Available at: <http://www.lel.ed.ac.uk/~wmaguire/survey/survey.html>.
- McIntosh, Angus. (1961) *Introduction to a survey of Scottish dialects*. Edinburgh: Nelson.
- Millar, Robert McColl, Barras, William, & Bonnici, Lisa. (2014) *Lexical Variation and Attrition in the Scottish Fishing Communities*. Edinburgh: Edinburgh University Press.
- Milroy, James. (1992) *Linguistic Variation and Change: On the Historical Sociolinguistics of English*. Oxford: Blackwell.
- Milroy, Lesley. (1987) *Language and Social Networks* (2nd edn.). Oxford: Blackwell.
- Müller, Nicole, & Ball, Martin J. (2013) (eds.) *Research Methods in Clinical Linguistics and Phonetics: A Practical Guide*. Oxford: Wiley-Blackwell.
- Orton, Harold. (1962) *The Survey of English Dialects: Introduction*. Leeds: Arnold.
- Orton, Harold, & Dieth, Eugen. (1962–1971) *The Survey of English Dialects*. Leeds: E.J. Arnold.
- Pichler, Heike. (2008) *A qualitative-quantitative analysis of negative auxiliaries in a northern English dialect: I don't know and I don't think, innit?* Unpublished PhD thesis, University of Aberdeen.
- Pratt, T.K. (1983) A case for direct questioning in traditional fieldwork. *American Speech* 58: 150–155.
- Tagliamonte, Sali. (2006) *Analysing Sociolinguistic Variation*. Cambridge: Cambridge University Press.
- Trudgill, Peter. (1974) *The Social Differentiation of English in Norwich*. Cambridge: Cambridge University Press.
- Upton, Clive, & Davies, Bethan. (2013) (eds.) *Analysing 21st Century British English: Conceptual and Methodological Aspects of the 'Voices' Project*. London: Routledge.
- Vaux, Bert, & Golder, Scott. (2003) The Harvard Dialect Survey. Cambridge, MA: Harvard University Linguistics Department http://www.tekstlab.uio.no/cambridge_survey/
- Watt, Dominic, Llamas, Carmen, Docherty, Gerard J., Hall, Damien, & Nycz, Jennifer. (2014a) Language and identity on the Scottish/English

- border. In Watt, Dominic, & Llamas, Carmen (eds.) *Language, Borders and Identity*, Edinburgh: Edinburgh University Press, pp. 8–26.
- Watt, Dominic, Llamas, Carmen, & Johnson, Daniel Ezra. (2014b) Sociolinguistic variation on the Scottish-English border. In Lawson, Robert (ed.). *Sociolinguistics in Scotland*. London: Palgrave Macmillan, pp. 79–102.
- Weigand, Gustav J. (1909) *Linguistischer Atlas des dacorumänischen Sprachgebietes, Volume 1*. Leipzig: J.A. Barth.
- Wolfram, Walt, & Fasold, Ralph W. (1974) *The Study of Social Dialects in American English*. New York: Prentice Hall.
- Wright, Joseph. (1898–1905) (ed.) *The English Dialect Dictionary*. 6 vols. London: H. Frowde.

15 Written Dialect Surveys

J.K. CHAMBERS

Written dialect questionnaires have played a role in data-gathering since the beginning of dialectology, and they continue to play that role today. The purpose, method, and administration has remained the same throughout this long history, but technological changes have altered the dissemination and collection of written questionnaires so thoroughly that it might be difficult at first glance to recognize the continuity of the tradition. Nowadays, written dialect surveys are normally distributed as on-line questionnaires that invite respondents to specify choices among dialect variants from a menu of possibilities or by typing in free choices. For more than a century before digitization became widely available at the turn of the twenty-first century, written surveys were distributed as hard copies on which respondents indicated choices by checking off appropriate variants in a list or by filling in blanks.

The essential similarities of the task, whether the medium is electronic or manual, are obvious. I will have little to say about the media, although I will include examples from both. In both media, written questionnaires elicit self-reports about linguistic variables from respondents based on their answers to a set of questions or linguistic tasks with little or (usually) no mediation from a trained linguist or fieldworker. Usefulness of the data from written dialect surveys depends upon judicious choices in the questions that are posed and concentration on topics for which the method is well suited. Those factors are crucial, regardless of the medium of distribution and will be given due weight in what follows. I begin with a selective survey of some landmarks in the use of written dialect surveys.

15.1 Georg Wenker and the Tradition of Written Questionnaires

The dialect survey that is widely credited with being the first-ever comprehensive, scientific application of dialect geography used postal questionnaires. It was carried out by Georg Wenker (1825–1911) in Germany beginning in 1876 and it initiated a dialect project that continued well beyond Wenker's lifetime. Although Wenker's survey was not literally the first ever,¹ it was undeniably a watershed in the discipline in terms of its influence. Dialectologists routinely concede its importance. Chambers and Trudgill (1998, 15) credit it as "the first dialect survey that can properly be called dialect geography." Petyt (1980, 40) called it "the first great dialect survey." Francis (1983, 67) called it "the first broad survey of a whole language community." Wells (1987, 364) says it put dialectology "on a serious, scholarly footing." The consensus is clear.

Georg Wenker was a schoolteacher in Düsseldorf with an ear for dialect, and he nurtured an ambition to plot dialect differences in the surrounding Rhineland, the boundary between High and Low German. To do so, he devised 40 sentences that included forms he thought would elicit regional words, pronunciations, or grammatical features. The sentences were generally homely and familiar, like these:

*In winter the dry leaves fly around through the air.
It will stop raining in a minute, and then the weather will get better.*

Wenker sent the list of sentences, written in standard German, to every school in the region and asked teachers to translate the sentences into local dialect. He then took selected features from the returns and published a monograph with a hand-drawn map called *Das Rheinische Platt* (1877).

Wenker's results caught the fancy of the Berlin Academy, which sponsored the expansion of his survey to all parts of the nation. Thus encouraged, Wenker sent his questionnaire to schools in every town and village, and eventually the schoolteachers returned more than 50,000 completed questionnaires. Wenker was then appointed to the University of Marburg as director of the Linguistic Atlas of the German Empire, where he assembled a staff for selecting and analyzing the results. Even with institutional support, the task was overwhelming. At the time of Wenker's death in 1911, numerous maps had been drawn and deposited in the Atlas archives. The project continued for 45 more years until 1956 resulting in thousands of hand-drawn maps in the Atlas archives at Marburg.

Wenker's questionnaire, hallowed though it is among dialectologists, was doomed from the outset. Essentially, almost every word in all 40 sentences on his questionnaire constituted an open question to which his schoolteacher-respondents might supply any one of several possible answers. Depending upon the schoolmasters' discretion or perhaps whimsy, a word might be "translated" with a lexical variant or a pronunciation variant or both. It might in some circumstances be morphologically restructured or syntactically altered following regional usage. If the variation was phonetic or phonological, the schoolmaster's acumen in manipulating spelling conventions came into play, at a time when phonetic notation was not standardized. In many cases, Wenker and his associates must have struggled trying to discern if variants cited on different questionnaires were the same or subtly different—and the problems of interpretation were multiplied as many as 50,000 times.

The profusion of data hindered the progress of Wenker's project, but in another sense his questionnaire provided too little data. Because choices were open, Wenker's successors were faced with gaps from region to region. Lexicographers, for instance, found the reported lexical variation too spotty for their purposes. In 1939, Wenker's successor Walter Mitzka designed a new questionnaire specifically for eliciting lexical variants with 200 questions on semantic fields such as body parts, flora and fauna, farm implements, and the like. He maintained Wenker's distribution system but pointed the schoolteachers at specific areas of variation. The coverage almost matched Wenker's with about 48,000 returns, but its productivity greatly exceeded Wenker's, eventually resulting in 21 volumes of the *Deutscher Wortatlas* (1953–1978) in a steady output that carried on some two decades beyond the official existence of the project itself. The success of the *Wortatlas* in terms of results and efficiency provided a much better model for the use of written dialect surveys than Wenker's prototype, which, it must be admitted, provided valuable negative lessons.

While Wenker and his associates were busy grappling with the superfluity of data in his questionnaires, an alternative methodology came into being that seemed to many dialectologists to be more promising. In 1896, Jules Gilliéron at the Sorbonne began laying the groundwork for the linguistic atlas of France. His questionnaire was designed for face-to-face elicitation of specific items, and he recruited a fieldworker, Edmond Edmont, a grocer

with a keen ear and apparently unbounded energy. For four years, Edmont cycled around the French countryside stopping at villages where he recruited informants and elicited about 1,500 items using Gilliéron's questionnaire. Edmont periodically bundled his results off to Paris, and publication of *Atlas linguistique de France* (1902–1910) began two years after the fieldwork was completed and ended eight years later with volume 13.

Gilliéron may have chosen his alternative methodology partly in reaction to Wenker's problems, but in any case the success of his use of elicitation by fieldworkers became the principal method for data-gathering in dialectology for the next century. Its dominance apparently led to facile assumptions about its superiority to written dialect surveys, but as we shall see in §2 below those assumptions do not always hold up to critical scrutiny.

Although written dialect surveys became infrequent, they continued to figure in the progress of dialect studies. Marius Kristensen immediately followed Wenker's lead by polling schoolteachers in Denmark. With a more focused questionnaire and a more compact national region, Kristensen produced results that should have refuted any suspicion that the method suffered from intrinsic problems (Valdemar and Kristensen 1898–1912).

For several decades, written questionnaires were used mainly to supplement fieldworker research, in an attempt at exploiting their efficiency and breadth of coverage alongside one-on-one interviews. Among the comprehensive surveys, the Survey of Scottish Dialects (Mather and Speitel 1975, 1977) began by distributing a postal questionnaire to all the schools in Scotland as the first phase of its research and followed it up by dispatching fieldworkers into the main regions. They were careful to justify their use by noting that they were aware of shortcomings in Wenker's method, which they took to be "eliciting material for which [it] is least suitable, namely for phonetic descriptions of common core vocabulary" (1975, 10).² Similarly, Alva Davis devised what he called a "correspondence questionnaire" (Davis 1949), which he distributed to 233 subjects in the Great Lakes region (Michigan, Ohio, Indiana, and Illinois), one of the regional projects of the Linguistic Atlas of the United States and Canada. Harold B. Allen used Davis's questionnaire as the basis for his postal questionnaire when he surveyed the adjoining region for *The Linguistic Atlas of the Upper Midwest* (Allen 1973–1976). Allen's use proved to be a fortuitous development for my purposes because it permits incisive comparison of the two methods in terms of reliability, coverage and other matters, as I will show in the next section.

Written dialect surveys went through a resurgence in the latter years of the twentieth century. One obvious factor is the rapid advance in digitization and computer technology that has solved many of the traditional problems of sorting and matching the superfluity of data. Another factor, no less obvious but seldom noted explicitly is the growth of literacy. Compulsory education came into being in most of the world around the beginning of the twentieth century, and by mid-century almost everyone in the developed countries was literate. The need for trained fieldworkers to explain, interpret, and transcribe the responses of their subjects was no longer necessary even in the more isolated byways where traditional dialectologists preferred to work. Moreover, the focus of dialectology shifted so that it sought responses of a more representative population sample for whom the intermediary was probably never really necessary.

In §4, I will demonstrate some practical aspects of written dialect surveys from the Dialect Topography of Canada (Chambers 1994, 2007), a sociolinguistic dialect survey. It began with a postal questionnaire and only in its very last phase did it use an electronic version. Data-handling and analysis were computerized all along, and it continues to yield useful results as an open-access, interactive database.

A "purer" example of an internet-based dialect survey is the Harvard Dialect Survey, one of the first surveys designed for interactive, open participation (Vaux ca. 2005, accessible at <http://dialect.redlog.net>). The Harvard Dialect Survey is regional rather than sociolinguistic. Respondents who logged onto the questionnaire were asked for their age and region.

They then answered 122 questions mostly about pronunciations and lexical choices, including these (without the percentages at the end, of course, which could only be added when the survey ended)—

7. **coupon**

- a. with [u:] as in “coop” (“cooon”) (66.86%)
- b. with [ju:] as in “cute” (“cyoopon”) (31.31%)
- c. other (1.83%)

73. What is your general term for the rubber-soled shoes worn in gym class, for athletic activities, etc.?

The results for each question are mapped with color-coded variants at the place the respondents identified as their home area. For the question about *coupon*, the maps show a sprinkling of yod (as in (b) above) in the northeast but considerable mixing in Florida and some of the central states. For the question about shoes worn in gym class, there are nine words offered and two null categories (“I have no general word for this” and “others”). The dominant answers are “sneakers” (45.5%) and “tennis shoes” (41.34%), and the maps show “sneakers” in every region but almost unanimous in the northeast, and “tennis shoes” more diffuse but densely represented in the upper Midwest.

The Harvard Survey attracted healthy attention in its active period, with its results cited in newspaper stories, online chats, undergraduate essays and many other contexts. Its popularity caused a minor logistical problem—with more than 10,000 responses per question (11,571 for *coupon*, 10,722 for *gym shoes*) the color-coded responses in the most populous regions become solid, and only grossly interpretable. (This is the kind of “problem” the designers presumably do not mind, and of course it can be eliminated simply by printing much bigger maps.) The results are not socially differentiated and the sample is inevitably skewed toward internet mavens (65% of the respondents are 14–39). Within limits, of course, it could yield sociolinguistic information. For age, for instance, it would be easy for the curators to map the over-50s (about 20%) with the 20-year-olds (34%) for any variable suspected to be changing. An ingenious application of the survey’s geographical orientation, due to Joshua Katz (Trendacosta ca. 2013), invites participants to answer 25 selected questions from the survey, and then purports to locate the participant geographically based on the survey norms. If refined, it could identify the subset of variables that determine regional allegiances.

Dialectologists might wish the Harvard Survey had gone deeper, but thousands of participants found it stimulating and debatable and intelligent fun. It integrates the written dialect questionnaire into the interview method that Labov called “casual and anonymous” (1972, 49, known informally as “quick and dirty”). Information is gathered quickly from willing participants, with minimal control on sampling. Abetted by the reach of the internet, the Harvard Dialect Survey asked over 100 questions and got thousands of answers. The results are at the very least suggestive. Dialectologists who wish to look deeper into the dialect situation using more systematic methods can find out where they might go looking from the Harvard results.

15.2 Testing Gilliéron’s Bias

After Gilliéron’s success with *Atlas linguistique de France*, dialectologists preferred fieldworker elicitation to written elicitation. When they used both methods, exploiting the obvious advantages of postal questionnaires in terms of their efficiency and coverage, they did so apologetically. Thus, Allen (1973, 29) in the *Linguistic Atlas of the Upper Midwest* (LAUM) states that he

considers the data elicited by postal questionnaires to be “distinctly supplementary and not additive” to the data elicited in field interviews. His only justification is what he calls “the contradictory data for a few items” (1973, 30), without further evidence, comparative or otherwise. To his credit, he does at one point concede that “the [postal] checklist returns exhibit rather remarkable correspondence with the findings of the field investigations.” In the context, this concession seems grudging, but he was, as we shall see, absolutely right.

Prima facie, there seems to be no theoretical basis for preferring fieldworker elicitation to postal questionnaires. At least three methodological factors, discussed in turn below, suggest that postal questionnaires should be equal to or better than field elicitations:

1. the data elicited by a fieldworker-administered questionnaire is exactly the same in content as data elicited on written questionnaires, and indeed the former was often an exact copy of the latter in its form and style
2. the presence of the fieldworker invokes the Observer’s Paradox (Labov 1971, 101) in the respondent to a greater extent than does the written questionnaire
3. even the best fieldworkers exert a directive influence on respondents to some extent

The identity in data elicited in either method is patently obvious by comparing questions on postal questionnaires and on the fieldworker’s worksheets in any survey that uses both. In the Scottish survey, for instance, Mather and Speitel (1975, 12) used the same questions for both phases. So did Allen, as we will see below. How different could they be? Questions in dialect surveys follow essentially the same format, which Labov (1971, 113) characterized as “a long question from the interviewer and a short answer from the subject.” Their purpose is to elicit regional words or structures, and even if the respondent spoke for 10 minutes or wrote an essay the primary goal is to elicit a particular lexeme (say, *darning needle* for *dragon fly*) or construction (say, *usen’t to* for *didn’t use to*) or whatever the relevant variant might be.

The points about the fieldworker as observer and as directive influence may overlap but in fact they are distinct and separable, as everyone who has worked in the field knows. The mere presence of the fieldworker automatically invokes Labov’s paradox by imposing an inquisitive outsider into a situation in which the speech we are hoping to observe is unobserved speech. Nelson Francis (1983, 70), an experienced fieldworker, notes that “the fieldworker is a stranger, with limited time and relatively little local knowledge.” The other aspect, field-worker bias, is more insidious but apparently equally inevitable. The greatest empirical demonstration of fieldworker bias in the long history of dialectology is Kurath and Bloch’s famous ranking of fieldworkers in their dialect survey of New England (1939, 52–53). They evaluated nine fieldworkers, themselves among them, on criteria such as “freedom from systematization according to the phonemic system of the field worker’s own speech” and “freedom from systematization according to the phonemic system of the informant.” For these and seven other criteria, they were able to put the fieldworkers in rank order. They graciously forego a master ranking that would show which fieldworkers showed the greatest bias, but for our purposes it is enough to say that all the fieldworkers biased the results to some extent.

In this light, we might read into Allen’s criticism of the postal questionnaire a very different implication from what he intended. “An inherent weakness in a mail questionnaire,” says Allen (1983, 30), “is the weakness due to the absence of a field worker with his ability to explain what is wanted without suggesting possible answers.” There is a fine distinction here, it seems to me, between the fieldworker explaining what is wanted and suggesting what the answer is.

The strongest test of the relative merits of postal questionnaires and fieldworker interviews would be a comparison of actual surveys in the same region eliciting the same items at around the same time. Ideally, both surveys would have the same director to ensure that the same standards were applied in both surveys.

Amazingly, exactly that test is available. Harold B. Allen directed the survey of the Upper Midwest (North Dakota, South Dakota, Minnesota, Nebraska, and Iowa) from its inception

in 1947 until its publication in three volumes almost 30 years later (Allen 1973–1976). In the peak years of the fieldwork, 1951–1953, Allen distributed a postal questionnaire with 136 lexical questions, a subset of the questions on the field questionnaire. Allen's use of the postal questionnaire was, as I have mentioned, apologetic, because he was operating in the historical context of what might be called "Gilliéron's bias," the presumed advantage of the field-worker that has its roots in Gilliéron's choice to employ a fieldworker in the formative years of dialectology. More than half a century later, Allen carried out parallel surveys, one using fieldworker elicitation and the other using a postal questionnaire, and my detailed comparison of his results show beyond a reasonable doubt that the bias is groundless.³

Allen's assertion that the postal data are "distinctly supplementary and not additive" to the field data amounts to a claim that they are representative of different populations. Whether or not the results are so different as to be incommensurable is a testable hypothesis. Indeed, it is the question that provided the original motivation for statistical testing, the very question that called t-tests and regression tests into being. The data elicited by the two methods is, in fairness to Allen and others who shared his bias, far too complex to be evaluated by inspection or any other informal means. Table 15.1 illustrates the complexity as succinctly as possible in the answers to four questions. The four questions are deliberately chosen to suggest the range of possibilities. Some variables have only two variants, like the first one in the table, *attic* or *garrett*, the part of the house under the eaves. Others have as many as six. The 35 lexical variables in LAUM for which Allen included sufficient information to allow comparison had 103 variants, an average of three each (Chambers 1998, 233, Table 3).

I chose the four variables in Table 15.1 because the responses reveal a kind of continuum from very similar to very dissimilar. For *attic/garrett*, it seems clear that the two surveys are compatible. Looking at the fourth variable, the container used for carrying coal, the answers seem less compatible. Allen explicitly notes that in this question "the mail survey does not closely correspond with the field data" (1973, 224); in the context of the bias, I note in passing that he does not single out instances where they do correspond. *Attic* and *coal scuttle* represent

Table 15.1 Responses to four questions in Field interviews and Postal questionnaires in the *Linguistic Atlas of the Upper Midwest* (Allen 1973–1976; selected from Chambers 1998, 233, Table 3). Percentages exceed 100 because multiple answers were counted equally.

<i>Variants</i>	<i>Field</i>	<i>Postal</i>
attic	96	98
garrett	08	06
blacktop	77	86
oil road	13	18
tarvia	12	07
(garbage) slop	71	89
(garbage) swill	20	26
(garbage) pail	77	82
(garbage) bucket	28	34
(coal) bucket	29	41
(coal) hod	33	14
(coal) pail	13	45
(coal) scuttle	45	19

extremes. In between, the gradations really defy impressionistic judgments. How similar are the field responses and postal responses for *blacktop/oil road/tarvia*, the name given to an asphalt roadway? The proportions for each variant differ by at least 5% and as much as 11%, but the rank order of the three responses is the same. What about *garbage slop/swill/pail/bucket*? Here the proportions differ by as much as 18%, and the rank order is different. On the positive side, the list of variants is the same in both surveys, as it also is in the *coal scuttle* question.

Judging compatibility for data like this cannot be carried out reliably by merely inspecting the lists. The paired-samples *t*-test was devised specifically to resolve this kind of analytic conundrum. Presented with two sets of responses, what is the likelihood of their coming from the same or different populations? If the field data and the postal data came from exactly the same source, their means in the *t*-test would be exactly the same, and the difference between them would be \emptyset . In other words, by the null hypothesis, the closer the *t* number is to \emptyset , the more probable it is that the two data-sets came from the same population. When we compare the field data and the postal data in Harold Allen's atlas, the *t*-value is -0.327, very close to \emptyset . The probability that the field data and the postal data were produced by the same population is 0.745, about 75%, that is, highly likely. By the paired-samples *t*-test, the subjects who sat for the field interviews and the ones who answered the postal survey cannot be shown to be different.⁴

There is no empirical support at all for Allen's claim about "the lack of congruity between the two sets of data" (1973, 29). I think it is fair to say that he was wrong because he relied on his impressions when the data were too complex to be evaluated impressionistically. Nevertheless, when the statistical evidence so convincingly supports the similarity of the results, it is surprising that his impressions led him to doubt their similarity. And I think it is fair to say, again, that he was predisposed, as was every dialectologist of his generation, to distrust the postal data. He inherited Gilliéron's bias, and even though, by incorporating a postal survey into his Upper Midwest survey, he had the empirical data to show that the bias was unfounded, he failed to see it.

15.3 Practical Advantages of Written Dialect Surveys

Given the strength of Gilliéron's bias in dialectology for almost a century, we might wonder why written questionnaires were used at all. They remained in use simply because they have inherent advantages over fieldworker elicitation that made them irresistible despite their "reputation." One advantage, already hinted at in my discussion of the observer's paradox and discussed below, is the elimination of what sociologists call "the enumerator effect." This advantage is not widely known among dialectologists and probably played no role in perpetuating the method, but it is a real advantage nonetheless, as we shall see. The most obvious advantage, however, and the one that led to its continued use against all opposition, is in terms of what can be broadly termed coverage.

Coverage takes in matters of population sampling, regional inclusiveness, deployment of personnel, and temporal efficiency. In all of these areas, the written questionnaire has distinct advantages over fieldworker elicitation. Again, there is no need for speculation on these matters because Harold Allen's accounting gives us a concrete basis for comparison. In Table 15.2, I have summarized Allen's meticulous discussion of his field and postal surveys in terms of linguistic coverage, respondents, regional coverage, personnel, and duration.

The clear disadvantage of the written questionnaire is Items Queried, shown in the first row of Table 15.2. Subjects can clearly tolerate much longer elicitation sessions when they are conducted by a fieldworker in convivial social circumstances. Allen's field questionnaire with about 550 questions is by no means long compared to other examples of the genre. Gilliéron's included about 1,500 items and the Survey of English Dialects included about 1,200 (Chambers and Trudgill 1998, 21–25). The Linguistic Atlas of the United States and Canada, which provided the prototype for Allen's questionnaire, included 700 items. The length seems

Table 15.2 Comparison of coverage and duration of the Field Survey and the Postal Survey for the *Linguistic Atlas of the Upper Midwest* (Allen 1973–1976, from Chambers 1998, 229, Table 1).

	<i>Field Survey</i>	<i>Postal Survey</i>
Items queried	584–661 words and phrases in 529–578 questions	136 lexical items
Respondents	208	1,064
Regions covered	97 “loosely defined communities”	at least 2 per county
Personnel	7 trained fieldworkers	local contacts
Duration	10 years (1947–1957)	3 years (1951–1953)

ungainly and has not escaped criticism. The American survey admitted that its 700 questions required at least two sittings, usually more, and most field records actually begin with one informant and end with someone else answering the question. How many sittings did Edmont require with 1,500 questions, and how many informant changes did he require? Passing off a field record as the response of a single subject when it is actually a composite record is notoriously bad methodology, almost fraudulent, as every social scientist knows. The integrity of the surveys would have been better served by using shorter, more manageable questionnaires.

However that may be, the fact remains that field elicitations can include many more items than a written dialect survey. Allen's 136 items comprise only about a quarter the questions on his field questionnaire, and at that it must have been pushing the upper limit. My postal questionnaire for the Dialect Topography of Canada, discussed in more detail in the next section, with 12 personal questions (age, sex, birthplace, and so on) and 73 linguistic items, was deemed long by a few respondents and short by some others but neither short nor long by most. It would be disastrous, of course, if the length of the questionnaire deterred any group (old or young people, men, or blue-collar workers) from completing it because it would affect its representativeness. It would be worse if length led many respondents to leave the last questions blank or to give slapdash answers. There is a limit to most people's tolerance for filling in answers on standardized forms, and purveyors of written dialect surveys are obliged to respect them.

In the other measurable aspects shown in Table 15.2, the postal questionnaire has certain advantages. The number of respondents is more than five times greater than the number of interviews. Its regional coverage is more representative, partly a function of the greater numbers which make the grid finer. The use of personnel is not really comparable. The time frame (duration) is much more efficient. These details come from one particular survey, of course, but generalizing the results seems fair. Simply put, postal questionnaires have a continuous tradition in dialect studies because they can blanket a region in a short time without involving trained personnel.

The other inherent advantage of written surveys is freeing the respondent from pressures, real or imagined, of an observer. Even in the highly favorable circumstances in which the interviewer is a local person with the requisite skills and training to conduct fieldwork there is likely to be a disconnect with the subjects. More pernicious than the Observer's Paradox, it turns out, is the potential for subconscious bias exerted by the fieldworker. Social scientists in poll-taking disciplines have known about this failing, called the “enumerator effect,” for many years. In 1970, the US Census completed the transition from a system whereby enumerators interviewed every citizen to a system in which citizens completed the questionnaires entirely on their own and returned them by mail. The change was not capricious. According to one study, “Far greater differences between enumerators were found than had been anticipated....

[A] complete census would have as much variability in its results because of enumerator effect as would a 25 percent sample if there were no enumerator effect" (Hansen 1978, 338).

Dialectologists have come to the same realization in their own terms. Kretzschmar (1992, 405) first identified what he called the "McDavid distribution" in the *Linguistic Atlas of the Middle and Southern Atlantic States*. It is named for Raven I. McDavid, the principal field-worker, a venerated figure in dialect studies, who conducted numerous interviews in two widely separated regions. Mapping the variables shows striking but implausible similarities "among speakers whose commonality," as Kretzschmar (2008, 345) says, "surely comes from the fact that most of them were interviewed by Raven McDavid, who conducted the interviews in Upstate New York and in the far south."

Nerbonne and Kleiweg (2003, 342–345) tried to devise a kind of normalization metric using "various corrections to try to obtain measurements which make sense from one field-worker to the next," but they were forced to conclude that "all comprehensive measurements reflected the fieldworker source of the data."

Fieldworker bias (Nerbonne and Kleiweg's term) is pervasive. Self-administered written questionnaires obviously avoid that problem, but of course they have limits and restrictions of their own.

15.4 Asking the Right Questions

The main limitation of written questionnaires is that they can only be distributed to literate subjects. That was a serious drawback years ago, and undoubtedly, it remains a problem in some societies today. Sporadic or unreliable literacy undoubtedly dictated Wenker's decision to distribute his questionnaire to schoolmasters rather than directly to the people whose speech he wanted to study. It was probably also the reason behind Gilliéron's decision to dispatch Edmonton as his "enumerator."

It is hardly a limitation in the era of mass literacy. The practical attainment of near-universal literacy can be gauged fairly reliably by the landmarks of the US Census mentioned in the previous section, which was entirely enumerator-administered in 1940 and entirely self-administered in 1970. In the decades in between, literacy increased to the point of unanimity. The Canadian government declared that "literacy is treated as a 'cultural given' for most adults in our society" (Statistics Canada 1996, 13), and so it is in all the developed nations.

Literacy is itself a graded skill, and written dialect questionnaires, like self-administered census forms, must necessarily avoid asking intricate, highly inferential, and multipartite questions. This too might seem like a limitation, especially to sociolinguists familiar with discursive, experiential interviews on a stylistic continuum (as in Tagliamonte 2006, 37ff). However, the "limits" are in fact well suited to the dialectology tradition of short answers.

Another limitation perhaps is the necessity for framing questions in such a way as to avoid any presupposition that respondents are technically sophisticated in terms of phonetic reporting, semantic nuances and other matters. Since Labov's New York study (1972, 177 et passim), we have known that self-assessments on linguistic matters are often mistaken, not willfully but naively. It is folly, for example, to ask respondents whether they have a tense vowel or lax vowel in the word *leisure*; it is essential that the question be posed indirectly, by asking, say, if the word *leisure* rhymes with *seizure* or with *pleasure* (as in Section 15.4.2 below). Sub-phonemic phonological processes such as Canadian Raising may be too nuanced for elicitation in a written questionnaire, and indeed no attempt is made in the *Dialect Topography of Canada*.

In the following sub-sections, I will discuss lexical predilections (Section 15.4.1), making pronunciation accessible (Section 15.4.2), the cost of phonology (Section 15.4.3), grammar and usage (Section 15.4.4), and happy accidents (Section 15.4.5). Examples will be drawn almost exclusively from the *Dialect Topography of Canada*, a postal survey that amassed

data from seven regions of Canada (Chambers 2007). The questionnaire included twelve questions asking for personal data: age, sex, occupation, education, place raised, place born, place of residence, mother's birthplace, mother's occupation, father's birthplace, father's occupation, and frequency of English use in four common situations. Seventy-three questions asked for linguistic information in the following categories: 30 pronunciation, 25 general vocabulary, 6 special vocabulary, 7 morphology, 5 syntax and 4 usage.⁵ I hasten to say that in looking so narrowly at the Dialect Topography project I am not espousing it as a model but simply exploiting my personal familiarity.

15.4.1 Lexical Predilections

From the beginning, dialectologists assumed that written surveys were best suited for lexical elicitation. Hence Mitzka and his colleagues perpetuated Wenker's methods in all respects except that they restricted the questions exclusively to lexical items. Eliciting lexical variants, as every dialectologist knows, is the easiest task. Words are conscious, unlike structural elements (phones, morphemes, syntagms), which are beneath consciousness.⁶ Words occur as gestalts, spellable by literates and retrievable as units by everyone. Unlike sandhi processes or passive constructions, words can be deliberately learned (for instance, in word-a-day calendars) and accidentally forgotten (as in the tip-of-the-tongue phenomenon). They are, for these very reasons, the least interesting linguistic variables. They are malleable and superficial: if variants exist, most people know more than one, and people are willing to change from their homely variant to the new favorite with a little social pressure.

If written dialect surveys were only suited to eliciting lexical variation, they would be methodologically impoverished. The rest of this chapter is devoted to showing that they are, with reasonable care, suited to eliciting the whole range of linguistic variation. All the same, there is no denying that they are very well suited to lexical elicitation.

The main consideration in eliciting lexical variants is in deciding whether the question should be closed or open. Closed questions, in which the variants are listed, have the obvious advantage of focusing the responses. Equally obviously, closed questions limit the possible answers. Only 2 (of 31) lexical questions in the Dialect Topography questionnaire are closed, and they are deliberate.

- Which do you say? She's going to bath the baby.
 She's going to bathe the baby

Here we were interested in the variants *bath/bathe*, essentially American/British differences, and not at all interested in their synonyms (*wash, clean, scrub*, etc.) that would have diluted the data-set in an open-ended question.

In most circumstances, closed questions are hard to justify in a discipline interested in discovering what people actually say. A notorious gaffe in Canadian dialectology came up in the Survey of Canadian English (Scargill and Warkentyne 1972) when they asked a closed question that omitted the main variant:

- What do you call a piece of furniture that seats two or three people in a row and has upholstered arms and back?
 A. sofa
 B. chesterfield
 C. davenport
 D. by another name

The survey took place at the moment when the Canadianism *chesterfield* was being replaced, but inexplicably the word that was replacing it, *couch*, was not listed as a choice. In order to register *couch* as a choice, the high-school students and their parents answering the question had to summon up the word *couch* (although they were likely to recognize both A. *sofa* and B. *chesterfield* as appropriate names) and choose D. Unless they then wrote *couch* in the margin, their specific variant would remain a mystery. Two decades later, in the Dialect Topography survey the equivalent question was open:

What do you call the upholstered piece of furniture that 3 or 4 people sit on in the living room? _____

The results show the predictable consequence of open questions in an unstable lexical set: in addition to *chesterfield* and *couch*, several other words were offered: *sofa*, *davenport*, *settee*, and from one speaker each *love seat*, *love couch*, *divan*, *bank*, and *chair* (Chambers 1995, 161). *Chair* is presumably a mistake, but *bank* came from an 80-year-old woman whose parents were Dutch and it definitely belongs in the data-set, if only as a footnote. In any event, the “noise” from minority choices that inevitably crop up in open-ended questions is hardly loud enough to vitiate analysis. The majority choices accounted for almost 90% of the total, and the age-correlations were statistically significant and socially coherent. *Chesterfield* was the majority choice among people 50 and over, but it became a minority choice among 40-year-olds in the survey, the 1950s in real time, and *couch* had such ascendancy that it was nearly the only variant used by teenagers in the survey.

In Section 15.4.6 below, I look at a more extreme case where an open-ended question gave considerably more noise and ended up, as I say, as a “happy accident.”

15.4.2 Making Pronunciation Accessible

Mather and Speitel claimed that Wenker’s questionnaire failed because he tried “eliciting material for which [it] is least suitable, namely for phonetic transcriptions of common core vocabulary” (1975, 10). On the contrary, vocabulary items with lexicalized sound differences, that is, pronunciation variants, are in fact nicely accessible in written questionnaires, though (pace Mather and Speitel) “phonetic transcriptions” cannot be elicited. What can be elicited are rhyme-words from which phonetic transcriptions can be inferred. Here is a straightforward example from the Dialect Topography survey:

Does LEISURE rhyme with Ømeasure, or with Øseizure?

The difference is notable in Canadian dialectology because the pronunciation with /ɛ/, rhyming with *measure*, is the prescriptive relic of nineteenth-century Briticisms that for some decades competed with the indigenous North American /ij/ variant, rhyming with *seizure* (Chambers 2004, 233–236). The 80-year apparent-time span captures the demise of the former variant in a steady-decade-by-decade diminution.

Variables with several pronunciation variants might be expected to pose challenges in terms of the respondents’ ability to make fine distinctions. However, we have not found inconsistencies or incoherence, for instance, in responses to questions like this one with four choices:

For you, does VASE rhyme with Øface, Ødays, Øcause, or Øhas?

Canadian responses are split between the middle two variants, /veɪz/ and /vəz/ in all regions. The responses, in this question and all others, have corrective value by revealing social patterns that would presumably not show up if the questions were vague or ambiguous or “difficult.”

15.4.3 The Cost of Phonology

Phonology is accessible in written surveys, but it comes at the cost of using more than one question, usually multiple questions, in order to get a plausible reading on contexts in which the process applies. Yod-Dropping, the loss of the on-glide /ju/ after coronals in words like *tune*, *dew*, *stupid*, *suit*, *nuisance*, and *lute* has been progressing for many decades, slowed though it is by the perception in some circles that yod-retention carries prestige (Clarke 2006). Phonological processes are more or less regular and so the need for multiple questions is somewhat alleviated. We asked two questions pertaining to tonic /ju~u/ after coronals (formally the same as pronunciation questions):

Does NEWS sound like Onyooze or Onooze?

Does the u in STUDENT sound like the Oo in *too*, or the Ou in *use*?

The questions were widely separated (by 10 intervening questions) to forestall looking back. We assumed that anyone who dropped yod in these words was likely to drop it in the whole set, and vice versa. This seemed reasonable, although not necessary, but adding another question or two from the same set would not make it more certain. We knew that /ju/ in non-tonic syllables sometimes behaved differently, as in the final syllables of *continue* and *retinue*, and included this question:

Does the ending of AVENUE sound like O you or Ooo?

Finally, we added a pronunciation question known to vary with the same /ju~u/ vowels:

Does the beginning of COUPON sound the same as O cue, or Ocoo?

Yod-Dropping in *student* and *news* correlated with age, from a low of about 50% among 80-year-olds to over 90% among teenagers (Chambers 2002a,b). Yod-Dropping in *avenue* was almost non-existent; yod is retained by people of all ages and is stable throughout the population at about 90%. Non-tonic yod in word-final position does not undergo Yod-Dropping in Canada. It frequently does, however, in adjacent regions of the United States, making *avenue* a Canadian-American shibboleth. *Coupon* has its own history: as a French loanword (from *couper* “to cut”) it did not enter English with yod but gained it upon nativization (yod is not dropped after velars, in *cute*, *kewpie*, *acupuncture*, etc.). The results in central Canada are bimodal, with the yod-pronunciation found more commonly in people 50 and older (stable for about 50%) and much less common among people under 50 (about 30% of them). Results like these provide an overview of the phonological process that at the least invites closer scrutiny by intensive fieldwork and at best shows systematic regularities in the speech of a large, representative population.

15.4.4 Grammar and Usage

Morphological, syntactic and usage variables are usually easily accessible in written surveys because the range of variation is circumscribed, usually binary. However, they are risky because these variables often carry prescriptive baggage and some respondents might be led to give the “correct” answer instead of the honest one. Any skewing that results from this kind of prescriptivism will show up as a bias toward prescriptive norms, hopefully negligible (as discussed in usage variables below).

When choices are binary, questions can be straightforward, as in this inquiry about the inflectional morphology of past tense of *sneak*:

- Which would you say? He snuck by when my back was turned.
 He sneaked by when my back was turned.

Fortunately, the 80-year apparent-time span of the Dialect Topography of Canada caught the change from the weak form *sneaked* to strong *snuck* from inception to completion. In the first decades of the twentieth century, about 80% of Canadians said *sneaked*, and in the final decade 95% said *snuck*. Graphically, the change forms a near-perfect S-curve in all regions (Chambers 2007, 29–32). The change takes place simultaneously across the nation, thus providing a textbook definition of a change in which a well-established form is replaced by a new standard, that is non-regional, form.

Any tendency toward prescriptive correction might be expected to show up most forcibly in usage variables since they are often subject to overt correction by teachers, parents and other arbiters. The best-known usage variable, perhaps, is nonconcord of object prepositions such as:

- Which do you say? Just between you and me, your aunt is often wrong.
 Just between you and I, your aunt is often wrong.

Results of these and other usage variables (Chambers 2009, 2010) certainly do not betray any skewing as a consequence of normative tendencies. Respondents make their choices such that the social dimensions fall clearly along lines of studies that use more discursive elicitation methods: more men use the non-prescribed *between you and I* than do women, and young people use it significantly more than do older people. The primary social correlate, however, is education—uniquely in sociolinguistics, to my knowledge, only usage variables have education as the primary correlate. University-educated people use fewer nonconcord constructions. But they do not use none at all. That fact leads to an abiding mystery: education exerts a moderating influence but it fails to eradicate them. Usage variables have been given little consideration in the sociolinguistic canon, which is understandable for those usage problems that are simply the result of old-timers carping about natural changes in the speech of young people (as was the case, for instance, with sentence-ending prepositions and *hopefully* as a sentence adverb, now generally accepted). Persistent and stable usage variables (notably object nonconcord and existential nonagreement) may persist because they pose processing problems that run afoul of grammatical rules, and thus appear to shed light on the innate language faculty (Chambers 2009, 2010). Usage variables are readily elicited in written questionnaires in closed questions.

15.4.5 Happy Accidents

One of the fundamental tenets of science is that interesting discoveries sometimes come from questions that were never asked. The classic case is antibiotics, which were discovered when an unknown fungus ruined Alexander Fleming's experiment by destroying the bacterial culture he was studying. A much more humble case comes from written dialect surveys. Open-ended questions, as discussed above, are usually necessary even though they may lead respondents to provide numerous variants. Boberg (2013, 139) notes that "fill-in-the-blank questions may elicit an overwhelming variety of minority responses, some chosen by only one or two respondents." As a case in point, he cites an open-ended question from the Dialect Topography survey about the schoolyard prank now widely known as a *wedgie*:

There is a prank (a kind of mean joke) that grade-school boys sometimes do to another boy: they grab his underpants at the back and hoist him up. What did you call that prank?

True enough, when the question was first posed in the early 1990s, the responses were (almost) “overwhelming”: specifically (as enumerated in Chambers 1994, 46–49, 53) there were four main responses and at least a dozen minor ones—and almost nobody over 50 had any word for it at all. A holy mess, we thought at the time, noteworthy for its stunning diversity. When we replicated the survey 10 years later in the same region, the results were almost as stunning but for exactly the opposite reason: this time, there was only one word for it. Almost everyone called it *wedgie* (93%). What had happened in the ten-year interval between surveys is that the *wedgie* had entered general consciousness. The schoolyard prank had previously existed in the semi-literate sub-culture of grade-schoolers, and there were almost as many words for it as there were schools. But suddenly it became known to almost everyone, and everyone started calling it by the same name. The word *wedgie* showed up in dictionaries, and it was called that by teachers, parents and some grandparents as well as by schoolchildren. The shift from the profusion of responses in the first survey to the focusing into one word in the later one charts “a prototypical standardizing change” (Chambers 2012, 471). The real-time difference in the two surveys showed the dramatic resolution of the profusion of regional responses that formerly seemed overwhelming. It merits a niche in the annals of sociolinguistics as an illustration of the dynamics of standardization, perhaps the most dramatic one so far documented.

It is one happy accident among many. Accidents like these are the rewards for asking good questions, categorizing the answers in all their profusion, discovering coherence in the correlates, and discerning their social evaluation. When they happen, they more than make up for the hypotheses that fail and the ones that yield tepid results. Those too are the result of asking questions, sometimes even good questions.

15.5 Looking Ahead

Venerable though the written dialect surveys are, they played a secondary role in the first century of systematic dialectology. Their role and (with it) their stature took an upward turn in the last decades of the twentieth century, and that role will almost certainly grow in years to come. The reasons are straightforward. The practical considerations that led to the continuous use of written questionnaires in the years when they were considered suspect are as persuasive as ever, and the social and linguistic changes that resulted in their upward turn remain in force. To these we now add their compatibility with globally accessible technology.

The practical considerations are the regional coverage and distributional efficiency that seduced dialectologists even when they felt the need to apologize for taking advantage of them. The social changes mainly involve the spread of literacy as a basic human right. So ubiquitous is literacy in many parts of the world that anybody over the age of 9 or 10 can serve as a subject in a written dialect survey. Dialectology also underwent a revolutionary change in the second half of the twentieth century with the incorporation of social correlates into the analytic framework. Dialectology is sociolinguistic dialectology with many more social correlates besides region. Eliciting social information is, if anything, easier in a written questionnaire than eliciting linguistic information. Once elicited, the computerization that has enabled the filing and sorting of nearly limitless data-sets can simultaneously cross-tabulate the social and linguistic correlates. And the internet links dialectologists to an almost unbounded population, a world of potential respondents, all of them equipped with devices for answering questions and circuits for transmitting their answers back to the database.

The simple virtues of regional coverage and distributional efficiency that beguiled Georg Wenker in 1876 beguile us still, but now both the coverage and the efficiency have expanded exponentially. As long as the enterprise of dialectology remains the survey of topographic variation, the surface features of language variation, the written dialect survey will be an important tool.

NOTES

- 1 Elsewhere I have made the case for Anto Warelius as the pioneering systematic dialectologist (Chambers 2002b, 3637). Warelius (1821–1904) was a Finnish scholar who set out in 1846 from south-eastern Finland and walked in a northwesterly direction eliciting dialect forms and expressions from the villagers he encountered. He ultimately established a dialect continuum across almost 400km. Warelius did not use written questionnaires, but face-to-face fieldworker elicitation, the method that came to dominate dialectology after Wenker.
- 2 Written questionnaires, as I show in the next section, seem to be quite adequate for eliciting pronunciation variants, that is, what Mather and Speitel call “phonetic descriptions of common core vocabulary.” As I see it, the shortcoming of Wenker’s questionnaire was the vagueness of the task he set his schoolmasters, by giving them leave to “translate” sentences with regional words, pronunciations, morphemes or grammar.
- 3 My comparison of Allen’s results using the two different methods was discussed in considerable detail some years ago (Chambers 1998); greater detail was essential at the time because my results contradicted Gilliéron’s bias at a time when it was shared by many dialectologists.
- 4 In the original article I also ran the Pearson correlation coefficient, with (obviously) equally robust results. The correlation coefficient is 0.928, that is, positive and very close to 1.0 (1998, 234).
- 5 The Dialect Topography of Canada is freely accessible in interactive databases with instructions and tutorials by Dr. Tony Pi at <<http://dialect.topography.chass.utoronto.ca>>
- 6 As a result, wordlists (dictionaries, glossaries, etc.) came into being centuries before grammars and phonologies.

REFERENCES

- Allen, Harold B. 1973–1976. *The Linguistic Atlas of the Upper Midwest*, 3 vols. Minneapolis: University of Minnesota Press.
- Boberg, Charles 2013. “The use of written questionnaires in Sociolinguistics.” In *Data Collection in Sociolinguistics*, ed. Christine Mallinson, Becky Childs and Gerard Van Herk. New York and London: Routledge, pp. 131–141.
- Chambers, J.K. 2012. “Homogeneity as a Sociolinguistic Motive.” *Canadian English: Autonomy and Homogeneity*, ed. Stefan Dollinger and Sandra Clarke. Special issue of *World Englishes* 31 (2012): 467–477.
- Chambers, J.K. 2010. “‘Bad’ grammar and the Language Faculty.” Selected Papers from NNAV 38, ed. Marielle Lerner. *Penn Working Papers in Linguistics* 16 (Fall 2010): 19–25. <[http://repository.upenn.edu/pwpl/vol16/iss2>](http://repository.upenn.edu/pwpl/vol16/iss2/)
- Chambers, J.K. 2009. “Cognition and the linguistic continuum from vernacular to standard.” In *Vernacular Universals and Language Contacts: Evidence from Varieties of English and Beyond*, ed. Markku Filppula, Juhani Klemola and Heli Paulasto. London/New York: Routledge, pp. 19–32.
- Chambers, J.K. 2007. “Geolinguistic patterns in a vast speech community.” *Modern Dialect Studies*, ed. Wladyslaw Cichocki, Wendy Burnett and Louise Beaulieu. *Proceedings of Methods XII*. Special issue of *Linguistica Atlantica* 28: 27–36.
- Chambers, J.K. 2004. “‘Canadian Dainty’: the rise and decline of Criticisms in Canadian English.” In *Legacies of Colonial English: Studies in Transported Dialects*, ed. Raymond Hickey. Cambridge, UK, and New York, US: Cambridge University Press. 224–241.
- Chambers, J.K. 2002a. “Yod-Dropping in an English accent.” *Journal of the Phonetic Society of Japan* 6: 4–11.
- Chambers, J.K. 2002b. “Dialectology.” *International Encyclopedia of Social and Behavioral Sciences*, 26 vols. Ed. Neil B. Smelser and Paul B. Baltes. Amsterdam: Elsevier Science. 3637–3642.
- Chambers, J.K. 1998. “Inferring dialect from a postal questionnaire.” *Journal of English Linguistics* 26: 222–246.
- Chambers, J.K. 1995. “The Canada-U.S. border as a vanishing isogloss: the evidence of chesterfield.” *Essays in Memory of Harold B. Allen: special issue Journal of English Linguistics* 23: 155–166.

- Chambers, J.K. 1994. "An introduction to Dialect Topography." *English World-Wide* 15: 35–53.
- Chambers, J.K., and Peter Trudgill. 1998. *Dialectology*. 2nd edition. Cambridge: Cambridge University Press.
- Clarke, Sandra. 2006. "Nooz or nyooz: the complex construction of Canadian identity." In *Canadian English in a Global Context*, ed. Peter Avery, J.K. Chambers, Alexandra D'Arcy, Elaine Gold and Keren Rice. *Canadian Journal of Linguistics* 5: 225–246.
- Davis, Alva L. 1949. A Word Atlas of the Great Lakes Region. Ph.D. dissertation. University of Michigan.
- Francis, W. Nelson. 1983. *Dialectology: An Introduction*. London and New York: Longman.
- Gilliéron, Jules, et Edmond Edmont. 1902–1910. *Atlas linguistique de France*, 13 vols. Paris: Champion.
- Hansen, Morris H. 1978. "How to count better: using statistics to improve the Census." In *Statistics: A Guide to the Unknown*, 2^e, ed. Judith M. Tanur. San Francisco: Holden-Day. 332–341.
- Kretzschmar, William. 2008. "Neural networks and the linguistics of speech." *Interdisciplinary Science Reviews* 33: 336–356.
- Kretzschmar, William. 1992. "Interactive computer mapping for the Linguistic Atlas of the Middle and South Atlantic States." In *Old English and New: Studies in Language and Linguistics in Honor of Frederic G. Cassidy*, ed. Joan H. Hall, Nick Doane and Dick Ringler. New York: Garland, pp. 400–414.
- Kurath, Hans, and Bernard Bloch. 1939. *Handbook of the Linguistic Geography of New England*. Washington, DC: American Council of Learned Societies.
- Labov, William. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.
- Labov, William 1971 "Some principles of linguistic methodology." *Language in Society* 1: 97–120.
- Mather, J.Y. and Hans H. Speitel, eds. 1975, 1977. *The Linguistic Atlas of Scotland*, 2 vols. London: Croom Helm.
- Mitzka, Walter, and L. E. Schmidt. 1953–1978. *Deutsche Wortatlas*. 21 vols. Giessen: Schmitz.
- Nerbonne, John, and Peter Kleiweg. 2003. "Lexical distance in LAMSAS." *Computers and the Humanities* 37: 339–357.
- Petyt, K. Malcolm. 1980. *The Study of Dialect*. London: Andre Deutsch.
- Scargill, Harry, and Henry Warkentyne. 1972. "The survey of Canadian English: A report." *English Quarterly* 5: 47–104.
- Statistics Canada. 1996. *Reading the Future: A Portrait of Literacy in Canada*. Ottawa: Statistics Canada. [Catalogue no. 89-551, <http://www.statcan.ca>].
- Tagliamonte, Sali A. 2006. *Analysing Sociolinguistic Variation*. Cambridge, UK: Cambridge University Press.
- Trendacosta, Katherine. ca. 2013. "This quiz pinpoints your American dialect down to the town." <<http://io9.com/this-quiz-pinpoints-your-american-dialect-down-to-the-t-1441692591>> [viewed November 2013]
- Valdemar, Bennike, and Marius Kristensen. 1898–1912. *Kort over de danske folkemål, med forklaringer*. Copenhagen: Gyldendal.
- Vaux, Bert. ca. 2005. "The Harvard Dialect Survey." Cambridge, MA: Harvard University Linguistics Department. <<http://dialect.redlog.net>> [viewed November 2013]
- Wells, C.J. 1987. *German: A Linguistic History to 1945*. Oxford: The Clarendon Press.
- Wenker, Georg. 1877. *Das Rheinische Platt mit einer autographischen Karte*. Düsseldorf: Schulte.

16 Field Interviews in Dialectology

GUY BAILEY

16.1 Introduction

Dialectology is best understood as the front end of historical linguistics. Originally conceived to test the Neogrammarian Hypothesis and later adapted to explore the spatial and social structure, linguistic content, and historical antecedents of the speech of a defined area, dialectology operates on the assumption that the speech of an area is a synchronic reflex of its linguistic and demographic history. The survey methods, approaches to fieldwork, and basic analytical techniques in dialectology have all traditionally reflected its fundamentally historical aims. However, while historical linguists typically analyze extant texts, some of which may have survived simply by chance, a major function of dialectology is the creation of linguistic “texts” that provide insight into the synchronic reflexes of an area’s linguistic and demographic history.¹ Field interviews are the primary mechanisms for the creation of these texts.

Over the last century and a half, approaches to field interviews have evolved as a result of (1) changes in our understanding about what constitutes good data (and hence appropriate linguistic texts) and (2) technological innovation. Georg Wenker’s pioneering work, begun in 1876, surely produced the densest coverage of any linguistic atlas (some 45,000 completed questionnaires from more than 40,000 villages), but later investigators believed that his approach to gathering data did not produce best evidence on the speech of an area.² To gather evidence for his *Deutscher Sprachatlas*, Wenker used a postal questionnaire to ask schoolmasters to “translate” 40 sentences from Standard German into the local dialect, an approach that was criticized for its reliance on intuitions about linguistic behavior rather than on actual observations of that behavior and for its lack of a mechanism for the systematic representation of sounds.³

Working after the development of the International Phonetic Alphabet, Jules Gilliéron addressed these criticisms in his *Atlas Linguistique de la France* (ALF). He used a trained field-worker both to do face-to-face interviews with folk informants chosen to represent points on a fixed geometric grid and also to transcribe those informants’ responses in phonetics. These innovations established best practices in dialectology for a century. Gilliéron’s approach to asking questions, however, did not. His questionnaire included queries that used the item under investigation in the question (e.g., “how do you say ‘50?’”). Recognizing that the use of target items in a question might influence an informant’s response, Karl Jaberg and Jacob Jud devised a questionnaire for their *Sprach- und Sachatlas Italiens und der Sudschweiz* that did

not include items under investigation in their queries and thus established a method for asking questions that still endures.

Although the approach to fieldwork established by Gilliéron and modified by Jaberg and Jud has been the standard in dialectology for more than a century, its actual implementation has varied significantly from project to project, and alternative approaches have emerged as new modes of investigation and different types of field instruments have been developed. An examination of the different types of linguistic evidence, investigative modes, and field instruments used across dialectology will illustrate both the issues that inform fieldwork in dialect geography and the range of approaches to field interviews in the discipline. An examination of the conduct of field interviews in recent projects will illustrate how ideas about evidence, modes of investigation, and field instruments are put into practice.

16.2 Some Issues That Guide Approaches to Fieldwork

16.2.1 Types of Linguistic Evidence

In constructing their atlases, Wenker and Gilliéron relied on two very different types of linguistic evidence. Wenker used the **intuitions** of schoolmasters about the dialect used in their villages. In the best cases, those intuitions were probably based on careful observation, but as McDavid (1983) has shown, intuitions are often unreliable:

McDavid compare[d] the intuitions he had about 86 linguistic features when he was interviewed by Bernard Bloch in 1937 (before he became a professional linguist) to evidence on those features in the Linguistic Atlas of the Middle and South Atlantic States (LAMSAS). The LAMSAS evidence show[ed] that of the 86 intuitive judgments in McDavid's interview, 29 were clearly wrong and another 15 were doubtful (Tillery 2000).

It was the unreliability of intuitions like these that led Gilliéron to focus on **observations** of actual linguistic behavior—and to have the observations recorded immediately in the field. No dialectologist disputes the superiority of the latter type of evidence, of observations of linguistic behavior over intuitions about that behavior.

Unfortunately, observations of behavior are difficult to obtain for some crucial linguistic features. For instance, multiple modals, as in “*I might could* talk to you next week,” are well-known stereotypes of English in the Southern United States and also occur in some parts of Great Britain. Obtaining systematic observations of the actual use of these forms, however, is difficult.⁴ Multiple modals are infrequent in both undirected and directed conversation, and the sample questions provided for the target item (*might could*) investigated in the relevant linguistics atlases, LAMSAS and the Linguistic Atlas of the Gulf States (LAGS), are ineffective:

Suggesting the possibility of being able to do something, you say, “I'm not sure, but I ____.” Or, you say, “If it quits raining by Thursday, I ____ get the yard work finished” (Pederson *et al.* 1974, 158).

Contrast this with the sample question provided for *green beans/snap beans*: “What do you call the beans you break in half to cook?” The latter question is tightly bounded and for the vast majority of respondents yields one of three lexical choices: *snap beans*, *green beans*, or *string beans*. The former question, on the other hand, has a wide range of possible responses and usually does not elicit a modal at all. The LAGS data demonstrates both how unproductive the question is and also how infrequent *might could* is in conversation: only 55 of the 1121

Table 16.1 Correlation of *might could* in LAGS with Different Elicitation Strategies
(Source: Bailey & Tillery, 1999).

Strategy	Rutledge Interviews	Total LAGS Corpus
Suggestion	50 (65.79%)	107 (50.47%)
Elicitation	18 (23.68%)	55 (25.94%)
Conversation	14 (18.24%)	61 (28.77%)
TOTAL*	76	212

*Note that the totals add up to more than 100% because in 11 instances in the total LAGS corpus and in 6 instances in the Rutledge interviews, *might could* appears both in response to suggestion and in conversation.

LAGS informants used *might could* in response to the questions designed to elicit it, and the form occurred spontaneously only 61 times in more than 5000 hours of recorded speech.

For *might could* (and for similarly infrequent but crucial target items) the most productive strategy was simply to ask informants directly if they used the term. The LAGS fieldworker who elicited the largest number of tokens of *might could* was Barbara Rutledge: she did 200 of the 1,121 LAGS interviews (17.84% of the total), but she elicited 76 of the 212 tokens of *might could* in the corpus (35.84%). She was able do this largely because she did not hesitate to suggest *might could* if neither conversation nor elicitation yielded the form, as Table 16.1 shows. More generally, Table 16.1 demonstrates that suggesting the form, that is, simply asking informants if they used it, was the single most productive strategy for eliciting *might could* in LAGS.

Tillery calls these responses to suggestions self-reports, and she distinguishes them from intuitions: "while self-reports are statements about one's own usage, intuitions are statements either about other people's usage or about the general currency or scope of linguistic forms" (2000, 57). (Note that the "intuitions" that often form the basis of work in formal linguistics correspond to what are here termed "self reports.") Tillery goes on to show that unlike intuitions, self-reports can be both reliable and valid indicators of linguistic behavior. To test their reliability, Tillery compared self-reports on six morpho-syntactic features in the two surveys that comprise a Survey of Oklahoma Dialects (SOD): a random sample telephone survey of 632 Oklahomans and a field survey of 144 native Oklahomans (for comparability Tillery used only the native Oklahomans in the telephone survey). Figure 16.1 summarizes the results of the comparison and shows that in every case the self-reports in the two surveys are within six percentage points of each other; in no case is the difference statistically significant. To test the validity of self-reports, Tillery examined complete typescripts of 28 interviews in the SOD field survey (roughly 20% of the sample) to look for uses of linguistic features that would contradict self-reports about them. Table 16.2 summarizes the results of her examination: in 27 instances Tillery found spontaneous uses of linguistic features elsewhere in the interview that confirmed informants' self-reports; in only five cases did she find uses that contradicted their self-reports (i.e., they used a form they denied using). Tillery (2000) concludes that when questions are framed properly, self-reports can be reliable, valid indicators of linguistic behavior. While the direct observation of linguistic behavior is clearly best evidence, self-reports can be a useful supplement for features that are difficult to elicit and infrequent in conversation. Intuitions about other people's speech, on the other hand, seem to have little value in dialectology.

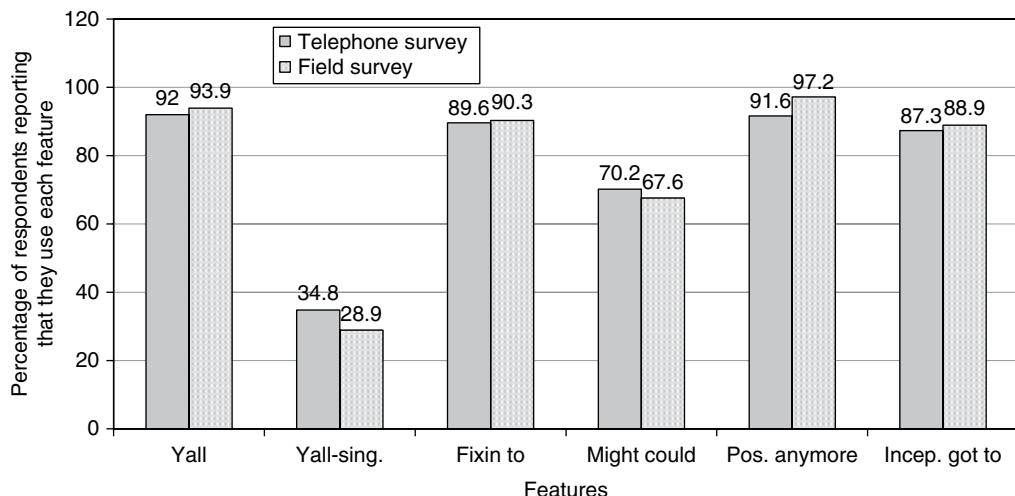


Figure 16.1 Comparison of Self Reports on Six Features in the SOD Field Survey and Among Native Respondents in Communities under 25,000 in the SOD Telephone Survey Source: Tillery, 2000.

Table 16.2 Comparisons of Self Reports on the Use of Linguistic Features with Their Actual Use in 28 Interviews in the SOD Field Survey (Source: Tillery, 2000).

Form	# self reports of use	# who report using and use in conversation	# who deny using but use in conversation
<i>yall</i>	23	3	0
<i>fixin to</i>	25	3	0
<i>got to</i>	24	9	3*
<i>might could</i>	19	0	1
<i>may can</i>	09	1	0
+ <i>anymore</i>	24	6	1
perf. <i>done</i>	11	5	0

* includes one instance in which an informant acknowledged *got to* but used *went to*

16.2.2 Modes of Investigation

The work of Wenker and Gilliéron also illustrates two different modes of investigation. The former used a postal questionnaire requiring written responses, a mode of investigation that continued to be used occasionally through the middle of the twentieth century.⁵ (See Chapter 15 in this volume for a discussion of written questionnaires.) The latter initiated the type of direct investigation in the field (fieldworkers interviewing informants one-on-one) that remains the most common mode of investigation. Technical innovations, however, have affected the operation of the latter approach in significant ways. Gilliéron and those who followed him in Europe, the United States, and England had fieldworkers transcribe the responses of informants in phonetics in the field during the interviews. The development of inexpensive, portable recording equipment changed that practice.

While the earliest mechanical recordings of dialects were apparently made in Sweden in 1897 and some mechanical recordings were made with informants in the United States as well as in Europe before World War II, the lack of portable recording equipment made the systematic recording of informants in the field impractical. As inexpensive, high-quality, portable recording equipment became available after World War II, the tape recording of field interviews became standard practice—and with it the separation of interviewing and transcription. The implications of separating fieldwork and transcription are obvious.⁶ Before their separation, the number of people who could do fieldwork was limited by the exceptional skill set required: the ability to do accurate, relatively fine-grained phonetic transcription without re-auditing responses along with the “soft skills” needed to engage strangers in informal conversation. Gilliéron’s reliance on a single fieldworker reflects the difficulty of this activity. Kurath attempted to expand the number of fieldworkers for the Linguistic Atlas of New England (LANE)—he used nine including himself—but this met with mixed success. Kurath points out that the multiple fieldworkers “produce[d] ‘personal boundaries’ on maps” (1939, 52), and as a result, in LAMSAS (which was begun shortly after the completion of fieldwork for LANE) he used just one fieldworker, although upon that fieldworker’s untimely death he had to employ a second one to complete the project.⁷ Perhaps the best indication of the value of good fieldworkers and the difficulty of the activity before the separation of interviewing and transcription is the fact that the fieldworkers for both ALF (Edmond Edmont) and LAMSAS (Guy Lowman and Raven I. McDavid, Jr.) have become legends in dialectology.

The separation of interviewing and transcription allows both for multiple fieldworkers and for multiple transcribers since fieldworkers do not need to be expert in phonetic transcription and transcribers do not need interviewing skills. Moreover, transcribers can be located in one place and work closely with each other to ensure as much agreement in phonetic norms and practices as possible.⁸ LAGS, which was begun in 1968, was the first linguistic atlas to fully operationalize the use of multiple fieldworkers and transcribers within the framework developed by Gilliéron. While the separation was generally successful, interviewer and transcriber effects do creep into the data. For example, as pointed out above, the distribution of *might could* in LAGS is partly the result of different elicitation strategies used by different fieldworkers (Bailey and Tillery, 1999). Likewise, Bailey, Tillery, and Andres (2005) show that the distribution of certain allophones of /l/ in LAGS represents the norms and practices of different transcribers rather than geographic, social, or linguistic effects. They also point out, though, that transcriber differences occur primarily with features that were not target items and that “the basic phonic record in both LAGS and LAMSAS … seems quite solid” (18).

Just as technical innovations have led to the separation of interviewing and transcription, they have also led to new modes of investigation. By the last two decades of the twentieth century, the convergence of the near universality of telephone ownership; improved telephone reception and recording technology; and the association of area codes and telephone exchanges with specific regions made it possible to use the telephone to do interviews for dialect surveys.⁹ The telephone permitted creative approaches to sampling (e.g., different types of random sampling over broad geographic areas) that were quite difficult using other modes of investigation (see Chapter 14 in this volume for a more complete discussion of sampling). Further, doing interviews by telephone significantly reduced the time and expense of surveys.¹⁰ Telephone surveys can provide comparable evidence to that obtained in field surveys, as Figure 16.1 above suggests and as the detailed analysis in Bailey, Wikle, and Tillery (1997) demonstrates. However, protocols for gathering data by telephone must be efficient and tightly structured since most informants are unwilling to spend the five or six hours needed for a linguistic atlas interview on the telephone. Nevertheless, telephone surveys produced some impressive results during the last decade of the twentieth century and the first decade of the twenty-first—note in particular Labov, Ash, and Boberg’s *Atlas of North American English* (ANAE); Bailey, Wikle, Tillery, and Sand (1991, 1993); Bernstein (1993) and Bernstein and Bernstein (1998); and Thomas (1996).

Two more recent technological innovations, the rapid expansion of cellular telephone usage and the development of the internet and other digital technologies, will make replicating these successes with telephone surveys more complicated; the broader implications of these new technologies for dialectology are unclear at this time. As cell phone ownership has increased (more than 80% of adults in the United States had at least one cell phone in 2010), the number of “wireless-only households” grew to more than a quarter of all households in 2010. Further, in 2010 more than 40% of cell phone subscribers did not live in the county associated with their exchange. These changes create two problems: first, response rates for cell phones are significantly lower than those for landlines, and second, a mismatch between area code and place of residence creates obvious sampling problems.¹¹ While telephone surveys are still possible, especially on the national level, their implementation on any smaller scale will be complex and will require careful planning.

The role of the internet and other digital technologies in dialectology is not yet clear. In some respects an internet survey simply represents a technological updating of a postal survey (email versus “snail mail”), but internet surveys offer possibilities for creating databases even larger than the one developed by Wenker. Exploiting these new technologies, though, will require appropriate mechanisms for creating both representative samples and also tightly structured protocols that elicit self-reports (or better yet, actual speech) rather than intuitions. Social class and ethnic differences in access to digital technologies pose special challenges for researchers. Work by Scherrer, Leeman, Kollu, and Werlen (2012), Kendall (2008), and Argamon, Koppel, Pennebaker, and Schler (2007) suggests some interesting possibilities for the future.

16.2.3 Field Instruments

To some extent the type of evidence and the mode of investigation determine the kind of field instrument used in any particular survey. The two most common field instruments are questionnaires and worksheets. Questionnaires focus on eliciting linguistic features in identical contexts from each informant. Hence, they include a set of pre-formed questions asked in the same way to every respondent (e.g., “what do you call the kind of beans that you break in half?” or “what is the number after nine?”). Questionnaires are particularly useful for ensuring comparable contexts for the elicitation of pronunciations. The *Survey of English Dialects [SED]* (Orton and Dieth, 1952; Orton *et al.* 1962–1971) is an excellent example of an atlas that collected its data via questionnaire.

In contrast, worksheets focus on eliciting linguistic features in ways that reflect informants’ most “natural” pronunciations. Worksheets provide fieldworkers with a set of target items for elicitation (e.g., *snap beans* or *green beans*; the pronunciation of *ten*) but leave the actual approach to elicitation up to the fieldworker. As indicated below, fieldworkers use a wide range of questioning techniques to elicit items. American atlases (e.g., LANE, LAMSAS, and LAGS) use worksheets rather than questionnaires. Projects using field interviews can employ either questionnaires or worksheets, but postal, internet, and telephone surveys are restricted to questionnaires. Postal and internet surveys require questions that cannot be misinterpreted since no fieldworker is present for clarification, and time constraints on telephone surveys (most people will not continue to respond to a survey on the telephone for more than a half hour or so) mean that protocols for collecting data by telephone must be tightly structured and efficient: questionnaires are the best mechanism for this.

Field surveys are usually flexible enough to allow for a variety of other instruments for gathering data as well. In New York City, Labov (1966) not only had informants answer questions, but he also used reading passages and word lists to create a mechanism for exploring stylistic variation (operationalized as attention paid to speech). Further, he developed question modules designed to elicit obtain longer stretches of discourse (i.e., “free conversation”) and

specific questions (e.g., “danger of death” questions) designed to elicit more informal styles—in contrast to the formal styles elicited by reading passages and word lists. Longer stretches of discourse were crucial for producing large numbers of tokens needed for the quantitative analysis of features such as post-vocalic /r/ and consonant cluster reduction.

The use of several different instruments to obtain different types of data is not uncommon in field surveys. The SOD field survey illustrates this point, and it also demonstrates how both observations of behavior and self-reports can be elicited in the same survey. The heart of the SOD field survey is a questionnaire (see Figure 16.2) based on the worksheets used in

<p>Next, I'd like to ask you about some traditional Oklahoma words and phrases. We want to know if people are still using them or if they are disappearing.</p>	
2.0	What word would you use for the beans that you break in half to cook? {If the respondent uses <i>snap bean</i> , go to 2.1; if not, go to 3.0}
2.1	Have you ever heard the term <i>snap beans</i> used for those kinds of beans? {If yes, ask 2.2; if no, go on to #3}
2.2	How often would you use that term yourself? (1) All of the time (2) Some of the time (3) Not very often (4) Never
3.0	What word would you use for the regular white bread that you buy at the store? {If the respondent uses <i>light bread</i> , go to #4; if not, ask 3.1}
3.1	Have you ever heard the term <i>light bread</i> for that kind of bread? {If yes, go to 3.2; if no, go to #4}
3.2	How often would you use that term yourself? (1) All of the time (2) Some of the time (3) Not very often (4) Never
4.1	How about the term <i>yall</i> ? Have you heard that? {If yes, ask 4.2; if no, go on to #5}
4.2	How often would you use that term yourself? (1) All of the time (2) Some of the time (3) Not very often (4) Never
4.3	Can you use the term <i>yall</i> for just one person, or does it always have to be for more than one? {If respondents will elaborate on how <i>yall</i> can be used, let them}
5.1	What do you call those little bugs that get on you in the grass and make you itch? [PROMPT: Would you call them <i>redbugs</i> or <i>chiggers</i> ?]
5.2	{If the respondent acknowledges both, ask} Is there a difference between them?
6.1	Now what do you call those bugs that light up at night? [PROMPT: Would you call them <i>lightening bugs</i> or <i>fireflies</i> ?]
7.1	What about the expression <i>fixin to</i> , as in “I'm <i>fixin to</i> go to town?” Have you heard that? {if no, go on to 8; if yes, ask 7.2}
7.2	Would you use it (1) All of the time (2) Some of the time (3) Not very often (4) Never

Figure 16.2 Sample Questions from the Protocol of the Field Survey Portion of SOD.

American linguistic atlas projects and carefully scripted to (1) determine passive knowledge as well as active use of lexical items; (2) determine how many of several synonyms informants know; (3) intersperse morphosyntactic with lexical items to minimize stigma attached to them; and (4) guide relatively inexperienced fieldworkers in obtaining the full range of information we wanted for each feature.¹² With an eye toward instrumental analysis, SOD also used a reading passage to ensure that for every informant, each vowel occurred at least five times in identical contexts in five phonological environments where applicable: before voiced obstruents or word finally; before voiceless obstruents; before nasals; before /l/; and before /r/. To explore the expansion of conditioned and unconditioned mergers in Oklahoma, the SOD field survey had informants read a list of minimal pairs and indicate whether the members of each pair sounded alike or different. Finally, SOD fieldworkers were instructed to obtain as much conversation as possible before and after the questionnaire was administered. While SOD interviews are tightly scripted, interviews in the Gilliéron tradition are often not; their conduct deserves a detailed examination.

16.3 The Conduct of Field Interviews

Although the conduct of field interviews in dialectology varies according to the type of evidence sought, mode of investigation, and type of instrument used, many atlas projects give fieldworkers considerable discretion, even to the point of informant selection. The success of an atlas depends in large part on how well fieldworkers do their work.

16.3.1 *The Selection of Informants*

The selection of good informants is critical in field surveys, and the task is not a simple one. In many projects, fieldworkers are simply given a set of criteria and told to find appropriate informants. For example, LAGS fieldworkers were told to interview at least one elderly (over 65 years old), Type I (grade-school education) white informant and one somewhat younger, Type II (high school education) white informant in each grid unit.¹³ Fieldworkers were instructed to interview to a Type III informant (college-educated) in every fifth grid, and in grids where the African American population exceeded 20% in 1930, they were instructed to interview Type I and Type II African Americans as well.¹⁴ LAGS fieldworkers used a number of strategies to find informants who met these criteria, but many (including Bailey) relied on local postmasters to serve as “talent scouts,” especially in rural communities. Postmasters are often from the local community, usually know which residents might be cooperative, and can provide directions to residences, though not actual addresses. This strategy worked so well that SOD fieldworkers used it exclusively.¹⁵

Once appropriate informants are selected, fieldworkers face the task of enlisting their cooperation. Straightforward approaches work best. The introductory script we prepared for SOD was based on what Bailey had done as a LAGS fieldworker:

Under the sponsorship of the National Geographic Society, we are conducting a survey of culture and language in Oklahoma. We are looking at how the culture and language of Oklahoma differ from one part of the state to another and at how they are changing over time. To do this survey, we have divided the state into 33 grids and are interviewing four natives of Oklahoma in each grid. Each person represents a different age group. We are asking all of the participants a series of questions about their views of Oklahoma and about some of the words that Oklahomans use for different things. The interviews last about 45 minutes, and we are tape-recording them because we cannot write down or remember everything that participants say (SOD Field Survey Questionnaire: 1).

The sponsorship of the National Geographic Society probably helped, and postmasters had already helped identify cooperative people, but we rarely had people decline to participate.

16.3.2 Doing the Interview

In conducting interviews fieldworkers have two primary tasks: eliciting the target items in the worksheets or questionnaire and eliciting speech that is representative of the population being studied. These two imperatives are not always compatible since asking questions designed to elicit linguistic features, especially grammatical features that may be stigmatized, focuses informants' attention on their speech and may lead to speech that does not reflect regional norms. As a result, most experienced fieldworkers develop strategies to mitigate informants' attention to their speech.¹⁶ One of the more effective strategies is simply to break the interview into several different sessions. The more interaction with an informant on different occasions, the more relaxed that informant becomes. Cukor-Avila and Bailey (2001) document the effect of this phenomenon, which they term *familiarity*, in their research in Springville, Texas. The following texts, taken from their work, illustrate those effects. Text #1 comes from the first interview Cukor-Avila did with Vanessa, an African American female born in 1961. It follows a simple question/answer format, with Vanessa answering questions but providing little elaboration. In fact, the questions are often longer than the responses.

Text #1

- FW: What time do you, what time does the store open? Like what time do you work from?
V: I have to come at eight an' get off at four thirty.
FW: Uh huh. So what do your kids do while they're at home? What do they do during the daytime when they're not in school?
V: Well they jus' got outta summer school an' they jus', they haven' been doin' too much of anything but layin' aroun' watchin' TV an' fussin' an' fightin' an' callin' over here for me. [laughs]
FW: Oh yeah you can hear 'em I bet, huh?
V: Yeah.
FW: So what kinda stuff do they fight over. Now lemme see now, tell me who's the oldest? The boy or the girl?
V: The girl.
FW: An' then the next one is...
V: Is the boy an' the girl.
FW: Uh huh, so who fights with who?
V: It be, it mostly be the baby fightin' with the oldes' girl. She jealous. She always want things to go her way. An' when it don't she wanna fight 'em. So it's the baby always wanna fight.
FW: Uh huh. So what about the boy? What does he play, referee?
V: Yeah he do. He uh, he's, he's like the man of the house. He watch over them an' when they get to fussin' too much he'd get on 'em. He keeps 'em in order.
FW: Does he?
V: Uh huh.
FW: Do they listen to him?
V: Yeah. They better or he'll knock 'em out. [laughs]
FW: Oh my goodness! How old did you say he is?
V: He's, he's eight. He's in between them.
FW: Uh huh. Is he big?
V: Uh huh. Real big.
FW: Hmmm. That's funny. Well do they have a lot of friends around here they play with too?
V: Nuh uh.

- FW: No?
 V: They uh, they got a niece, a cousin that come play with 'em every day. She comes.
 FW: Where does she live?
 V: She live right down the road from the school.

Compare this to Text 2, taken from Cukor-Avila's second interview with Vanessa, done a week later. In Text 2, Vanessa responds to questions with more elaboration and intimate personal information. As these two texts show, even a small increase in familiarity can increase an informant's comfort level with a fieldworker. Doing a five to six hour interview over two or three days does not eliminate the observer's paradox, but it can ameliorate its worst effects.

Text #2

- FW: Uh, so after you, after you got outta high school then what did [V. interrupts]
 V: I didn't finish high school.
 FW: Oh you didn't?
 V: Nuh uh.
 FW: How far did you go?
 V: To the, uh I was goin' into eleventh.
 FW: To the eleventh grade?
 V: Uh huh.
 FW: An' why did you decide to leave?
 V: I ran off from home.
 FW: Oh you did? You wanna tell me about it? [laughs] That's O.K. I won't tell anybody.
 V: Yeah, uh, well I [FW interrupts]
 FW: You were livin' out here in Springville, right?
 V: Yeah I felt at the time like all of my friends they would, you know, mother was givin' them a little slack to let 'em go places an' do things an' stuff. So Mama wasn't doin' that for me. She was kinda strict on me. An' I kep' astin' {askin'} her because I wanted to be around, be aroun' my friends an' uh, an' do some of the things that they were doin'. An' she seem like she didn't understand that. So, uh she let me go to uh, Brownsville to visit. It was - O.K. she had jus' met this man. An' they was down there pickin' cotton. An' so I tol' her, you know, to let me go down there. So I went down there an' I stayed with 'em a week, for two weeks. An' I was comin' back home. An' so I met these people from Mississippi. An' they seem, you know, nice an' stuff.
 FW: But you, where'd you meet 'em? Here or on the, on the highway?
 V: When I was down in uh, [FW interrupts]
 FW: Oh in Brownsville.
 V: Brownsville. Uh huh. And uh, so I got to know 'em. So I was comin' back home. But the guy that I met, he was, he say, "Oh you don't hafta catch the bus." Say, "I'll bring you, you know, back to Springville." 'Cause we came right through here to go to, uh Mississippi. So I believed him. An' uh, when I woke up we was in Houston. An' I say, "I thought you told me that I was goin', you know, get off at home." He say, "I change my min'. I want you to go to Mississippi with me." An' so I [FW interrupts]
 FW: How old were you at that time?
 V: I was seventeen.
 FW: An' how old was he?
 V: About twenty-five.
 FW: Uh huh.
 V: So uh, when we had got to Mississippi you know I was scared because I s'pose been came back home an' he didn't let me go back home. An' I didn't know what to do. So I didn't call Mama or nothin'. An' had her worried about me. An' uh, so...
 FW: She probably thought you were kidnapped or something.

V: Yeah she did. They did. You know they was goin' crazy by, by me bein' young an' not carin' about stuff. I was scared mostly. That's what it was. I didn't call or try to get in touch with 'em an' let 'em know that I was all right. So one day I, I was up there 'bout a month, two months. So one day the phone rang an' it was the police. An' they, they tol' me they say, "We can' make you come back home." But say, "Are you all right?" I say, "Yeah I'm all right." They say, "Well would you call, uh talk to your mama to let her know you is all right?" I was so scared. I say, "Yeah I'll talk to her." So after I heard her voice an' we talked I was cryin'. She say uh, "You wanna come back home?" I say, "Yeah I wanna come back home." She say, she say, "Well we'll sen' uh...." My, my brother came after me. So he came, like I talked to him that day, the nex' day he was there.

Breaking the interview up over several sessions is not the only mechanism for ameliorating the observer's paradox. The right strategy for asking questions can generate significant amounts of conversation; create contexts that lead to many of the target items occurring in conversation; and reduce the informant's focus on speech. Figure 16.3 summarizes one such approach—the interview strategy that several LAGS fieldworkers (including Bailey) used. This approach begins with questions about the area and what it was like when the informant was growing up. Using topics that occur in the directed conversation as a cue, the approach then uses "shotgun questions," broad questions designed to provide informants with the opportunity for using a number of target items. Text 3, taken from a LAGS interview done by Bailey in Sprott, Alabama, in 1978 with a white male born in 1910, illustrates the shotgun question. The cue for the question was the informant's earlier comment that he had always had a vegetable garden. The shotgun question elicits more than a dozen LAGS target items (indicated in italics in the text).

TEXT #3

FW: What-all would you usually have grown in your garden? What kind of things would you have planted there?

INF: In the springtime when *you was* plannin' a garden, this time of year you would put out some purple-top *turnips*, you'd put out some *radish*, you'd put out some *lettuce*, an' you would prepare yourself a bed for you some *tomato* seed, in a hot bed that you might have to put out after the weather got ready. Then you would plant your garden according. As the climate adjusted itself the nex' thing you would put in your garden would be some *snap beans* an' some *lima beans*, some *butter beans* we call them. *Lima beans* is the same thing. *They's* a bunch one an' a runnin' one. But you put your bunch ones out an' maybe you could stagger it along an' put your runnin' ones out later when you haven't got them comin' along together. An' then shortly after that would be your *tomato* plants that you put in there. A lot of people have put out *cabbage* an' some put out *Irish potatoes* and the bed should have *done been planted*. An' then you can plant your garden where that you'll have something comin' out of the garden all into the summer time until the frost kills it.

Many target items can be elicited in this way. More specific follow-up questions that flow naturally from the shotgun questions should provide most of the other target items. Text 4, taken from the same LAGS interview, provides a follow-up to a shotgun question

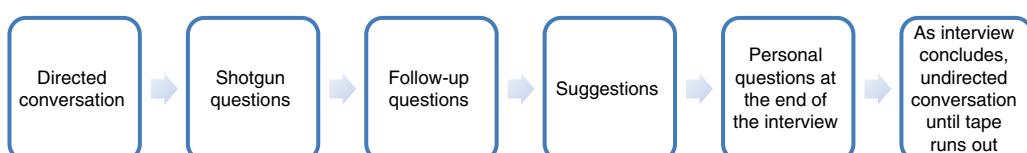


Figure 16.3

about growing cotton. The informant had mentioned clearing cotton fields of grass, but he failed to mention the names of various grasses that grew in the fields. His answer to a question about the kinds of grasses that come up in cotton fields yields several more target items:

TEXT #4

FW: What kind of grass was that that would grow up in the fields?

INF: *Crabgrass* was the main grass that they had, but you had other types of grass that *growed* up there. You had *cuckleburrs* an' *coffee weeds*, but crabgrass was the grass that *give* you the most problem in a cotton *crop* because that crabgrass would one time get aholt in the ground an' then if you had a lot of moisture in the ground, you could turn it over but you hadn' *done* anything. It'd take root over there at each joint, then you'd turn it back over an' it'd already have them *roots* there. It's not very hard to kill in the dry time, but if you got a lot of moisture in the groun', crabgrass 'll *give* you a hard time.

Finally, fieldworkers can suggest target items that do not occur in either shotgun or follow-up questions. Often suggestions can be framed in subtle ways. Although ostensibly a question, the query in Text 5 (from the same LAGS interview) is actually a suggestion. The plural of *bushel* is a LAGS target item that did not occur in conversation or responses to questions; Bailey "suggested" it in a question framed to elicit the feature.

TEXT #5

FW: How many bushels would you grow to an acre of sweet potatoes? What would be a good crop?

INF: Well now, the *potatoes* is like a lot of other things. It's got to depend on—if you get the right moisture it wouldn't be unusual for you to get *fifty bushels* of potatoes to *a acre* of good potatoes. But potato is something that's got to have water at the right time an' if it don't get it—Now last year H. R., *live down* the road *a little ways*, is into a potato business in a big way. An' they put the potatoes out an' they didn't get any water. Well, they didn't have any potatoes to gather. You wind up with *potatoes* that's great big jumbo potatoes an' they crack open hard. But the potato business, back when we was plantin' the potatoes, you didn't have anything to get your vines out with. You had to take a *pair of mules* an' pull your vines off an' then take a *middlebuster* an' throw the potatoes out. But now they've got a *machine* that *they plants these plants in the ground* an' then they got a machine to do away with the vines. They got a potato digger that digs up the *potatoes* that comes up here on a conveyor, up on top here, an' then you got a line of men sittin' there on the side selectin' potatoes, puttin' them in the crate as the machine goes along. So you do it all at one time.

The answer does generate the item that was suggested, but it generates another 10 target items as well. As Texts 3 through 5 suggest, the questioning strategy outlined in Figure 16.3 usually produced the vast majority of the grammatical and phonological target items in conversation. Once the linguistic substance of the interview has been obtained, the field-worker concludes the interview by gathering relevant demographic information.

Most questionnaires provide some contexts for generating spontaneous speech, but they generally do not provide the same flexibility that worksheets do. As a result, the structure of the questionnaire itself bears much of the responsibility for producing speech that reflects regional norms. The key lies in how questions are organized and framed. The approach used in SOD, illustrated in Figure 16.2 above, was pre-tested and modified extensively before it was actually used. Pre-tests showed that the following helped reduce the focus on attention to speech:

1. Contextualizing questions that elicit linguistic features as part of a broader investigation of the culture of an area. Our questionnaire included a number of questions about culture (e.g., “Do you consider Oklahoma a Midwestern, Western, or Southern state?”)
2. Constructing questions focused on phonological features, such as the use of rounded or unrounded vowels in *hawk*, as lexical queries (e.g., “What about those birds that sit on telephone poles and swoop down to kill mice and other small animals, what do you call them?”)
3. Framing questions focused on grammatical items such as *yall* and *fixin to* as lexical questions and embedding them with other lexical questions (see Figure 16.2 above)
4. Asking informants if they had heard a feature used in the area before asking them if they used it and offering them choices about how often they would use the term themselves. As Figure 16.2 shows, both of these mechanisms for eliciting self-reports were systematized as part of the SOD questionnaire. Bailey, Wikle and Tillery (1997) and Tillery (2000) demonstrate the success of this approach.

The questionnaire constructed for ANAE included some of these same features, as well as innovations such as the use of the semantic differential technique for eliciting pronunciation items. Labov, Ash, and Boberg (2006) provide both the complete questionnaire used in ANAE and a discussion of the logic behind the composition and ordering of questions.

16.4 The Atlas of North American English

ANAE provides an excellent illustration of how choices about types of evidence, mode of investigation, and field instrument help structure a survey, as well as an example of how a contemporary atlas can use modern technology to achieve impressive results. Begun in 1991 and published in 2006, ANAE focused primarily on on-going sound changes, including both vowel shifts and mergers. One consequence of this decision was that both observations of informants' behavior and their self-reports about whether word pairs sounded alike or different became components of the data. In addition, the decision to focus on on-going change meant that interviews needed to be done over a relatively short time frame; the mode of investigation that best enabled this objective to be met was the telephone survey. The decisions to conduct a telephone survey and to elicit self-reports as well as observations of behavior meant that a questionnaire, rather than worksheets, was the most appropriate investigative instrument and that the interview would need to be tightly scripted. Further, it meant that the questionnaire itself would bear the responsibility for mitigating the effects of attention to speech. As in most telephone surveys, informant selection in ANAE was determined according to a clearly articulated protocol. Interviewers approached informants with a script much like the one used for SOD, and their primary responsibility was to ensure that the questionnaire was fully administered. Finally, ANAE interviewers worked to get permission for follow-up interviews in which informants would read word lists and passages to allow for instrumental analysis. The results of these ANAE interviews are impressive, and the series of maps published in Labov, Ash, and Boberg (2006) provide the only continent-wide perspective on English phonology and phonological change.

16.5 Conclusion

Although technological developments in dialectology have impacted approaches to gathering data in significant ways, the construction of linguistic texts is still the central focus of the discipline, and field interviews remain the primary mechanism for the construction of linguistic

texts. Effective field interviews require a clear understanding of the type of data sought; a mode of investigation appropriate for obtaining that data; and field instruments and interview strategies that elicit the needed data while mitigating the most intrusive effects of attention to speech. The rich history of field interviews in dialectology provides clear guideposts for all of those activities. At the same time, that history provides the imperative for the continual innovation and improvement that has always undergirded best practices in the discipline.

NOTES

- 1 See Pederson (1974) for an insightful discussion of the creation of linguistic texts in dialectology. Lee Pederson's influence on many of the ideas here and on my work more generally should be apparent to anyone familiar with his research. This paper is dedicated to him.
- 2 Among Gilliéron's pioneering efforts was the use of a sample to make inferences about the behavior of a population. Although Pickford (1956), Underwood (1974), and others have criticized sampling techniques used in dialect geography, dialectology was among the first academic disciplines to use systematic sampling for gathering data, and the samples were quite sophisticated for their times. The work of Wenker, Gilliéron, Jaberg and Jud, Kurath, and others has not received the credit it deserves for methodological innovations in sample construction. For a discussion of some of the issues in sampling, see Bailey, Wikle, and Tillary (1997).
- 3 For a Neogrammarian critique of Wenker's work, see Schirmunski (1962).
- 4 The discussion that follows is based on Bailey and Tillary (1999) and Tillary (2000). Those two papers provide a detailed discussion of the elicitation of multiple modals.
- 5 See, for instance, Wood (1971).
- 6 Interestingly, the tape recording of field interviews and the separation of interviewing and transcription were not uncontroversial. See Wilson (1956), McDavid (1957), and Hedblom (1959).
- 7 Near the end of the project, LAMSAS used a few additional fieldworkers to tie up loose ends. Even with just Lowman and McDavid, however, fieldworker effects crop up in the results. See, in particular, the insightful discussion in Nerbonne and Kleiweg (2003) and the analysis in Bailey, Tillary, and Andres (2005).
- 8 See Pederson (1974) for a discussion of the implications of the use of tape recorders and the separation of interviewing and transcription.
- 9 Labov first used the telephone for a survey in 1966 but did not record the interviews. See Bailey and Bernstein (1989), Bailey and Dyer (1992), and Bailey, Tillary, and Wikle (1997) for a discussion of the construction of the tape-recorded telephone surveys in Texas and Oklahoma.
- 10 For example, the 632 tape-recorded interviews in the random sample telephone survey portion of SOD were completed in three months for \$15,000 (roughly \$25 per interview in 1991 dollars).
- 11 This paragraph summarizes data from Lavrakas *et al.* (2010).
- 12 SOD used only three fieldworkers: Sonya Davenport, Dave Groughner, and Lori Sand. Although initially inexperienced, they all proved to be superb fieldworkers.
- 13 Although in LANE and LAMSAS informant *Type* correlated with a cluster of social features (e.g., education, travel, breadth of social contacts), in LAGS informant *Type* was operationalized as education.
- 14 The use of the 1930 census reflects the historical orientation of LAGS. After 1930, the African American population declined significantly in much of the South as a result of the Great Migration. The 1930 census provides a better representation of the historical demography of the South. LAGS fieldworkers also interviewed African American Type III informants in many grids.
- 15 SOD interviewed four informants in each grid: someone about 20, about 40, about 60, and about 80—all with a high school education.
- 16 The best approach to reducing attention to speech and ameliorating the “observer's paradox” is to change the interlocutor role of the fieldworker through the use of peer group interviews or site studies (Bell 1984, 2000; Cukor-Avila and Bailey 1995, 2001). Unfortunately, the type of information needed in dialectology usually precludes these types of interviews.

REFERENCES

- Argamon, Shlomo, Moshe Koppel, James Pennebaker, and Jonathan Schler. 2007. "Mining the Blogosphere: Age, Gender, and the Varieties of Self-Expression." *First Monday* 12 (9).
- Bailey, Guy, and Cynthia Bernstein. "Methodology of a Phonological Survey of Texas." *Journal of English Linguistics* 22: 6–16.
- Bailey, Guy, and Margie Dyer. 1992. "An Approach to Sampling in Dialectology." *American Speech* 67: 1–18.
- Bailey, Guy, Jan Tillery, and Claire Andres. 2005. "Some Effects of Transcribers on Data in Dialectology." *American Speech* 80: 3–21.
- Bailey, Guy, and Jan Tillery. 1999. "The Rutledge Effect: The Impact of Interviewers on Survey Results in Linguistics." *American Speech* 74: 389–402.
- Bailey, Guy, Tom Wikle, and Jan Tillery. 1997. "The Effects of Methods on Results in Dialectology." *English World-Wide* 18: 35–63.
- Bailey, Guy, Tom Wikle, Jan Tillery, and Lori Sand. 1993. "Some Patterns of Linguistic Diffusion." *Language Variation and Change* 5: 359–390.
- Bailey, Guy, Tom Wikle, Jan Tillery, and Lori Sand. 1991. "The Apparent Time Construct." *Language Variation and Change* 3: 241–264.
- Bell, Allan. 1984. "Language Style as Audience Design." *Language in Society* 13: 145–204.
- Bell, Allan. 2000. "Back in Style: Re-working Audience Design." In Penelope Eckert and John R. Rickford, eds. *Style and Sociolinguistic Variation*. Cambridge: Cambridge University Press.
- Bernstein, Cynthia. 1993. "Measuring Social Causes of Phonological Variation in Texas." *American Speech* 68: 227–240.
- Bernstein, Cynthia, and Robert Bernstein. 1998. "Phonological Innovation in East Texas: Different Samples, Similar Explanations." *American Speech* 73: 44–56.
- Cukor-Avila, Patricia, and Guy Bailey. 2001. "The Effects of the Race of the Interviewer on Sociolinguistic Fieldwork." *Journal of Sociolinguistics* 5: 254–270.
- Cukor-Avila, Patricia, and Guy Bailey. 1995. "An Approach to Sociolinguistic Fieldwork." *English World-Wide* 16: 1–36.
- Hedblom, Folke. 1959. "Recording in Dialect Investigation in Sweden." *Phonetica* 3: 95–108.
- Kendall, Tyler. 2008. "On the History and Future of Sociolinguistic Data." *Language and Linguistic Compass* 2: 332–351.
- Kurath, Hans. 1939. *Handbook to the Linguistic Geography of New England*. With the collaboration of Marcus L. Hansen, Bernard Bloch, and Julia Bloch. Providence: Brown University Press.
- Labov, William. 1966. *The Social Stratification of English in New York City*. Washington, D.C.: Center for Applied Linguistics.
- Labov, William, Ash Sherry, and Charles Boberg. 2006. *Atlas of North American English*. Berlin: Walter de Gruyter.
- Lavrakas, Paul J. et al. 2010. "New Considerations for Survey Researchers When Planning and Conducting RDD Telephone Surveys in the U.S. with Respondents Reached via Cell Phone Numbers." Manuscript prepared for the AAPOR Council by the Cell Phone Task Force.
- McDavid, Raven I., Jr. 1957. "Tape Recording in Dialect Geography: A Cautionary Note." *Journal of Canadian Linguistics* 3: 3–8.
- McDavid, Raven, I., Jr. 1983. "The Failure of Intuition." *SECOL Review* 7: 128–139.
- Nerbonne, John and Peter Kleiweg. 2003. "Lexical Distance in LAMSAS." *Computers in the Humanities* 37: 339–357.
- Orton, Harold, and Eugene Dieth. 1952. *A Questionnaire for a Linguistic Atlas of England*. Leeds: Leeds Philosophical and Literary Society.
- Orton, Harold, et al. 1962–1971. *Survey of English Dialects*. Introduction and 4 volumes (in 3 parts each). Leeds: E.J. Arnold and Sons.
- Pederson, Lee. 1974. "Tape/Text and Analogues." *American Speech* 49: 5–23.
- Pederson, Lee, Raven I. McDavid, Jr., Charles W. Foster, and Charles E. Billiard. 1974. *A Manual for Dialect Research in the Southern States*. 2nd ed. Tuscaloosa: University of Alabama Press.
- Pickford, Glenna Ruth. 1956. "American Linguistic Geography: A Sociological Appraisal." *Word* 32: 211–233.
- Scherrer, Yves, Adrian Leeman, Marie-José Kolly, and Iwar Werlen. 2012. "Dialäkt Äpp – A Smartphone Application for Swiss German Dialects with Great Scientific Potential." 7^{ème} Congrès SIDG- Dialect 2.0, Vienne.
- Schirmunski, Victor M. 1962. *Deutsche Mundartkunde. Vergleichende Laut- und*

- Formenlehre der deutschen Mundarten.* Berlin: Akademie-Verlag.
- Thomas, Erik R. 1996. "A rural/metropolitan split in the speech of Texas Anglos." *Language Variation and Change* 9: 309–332.
- Tillery, Jan. 2000. "The Reliability and Validity of Linguistic Self Reports." *Southern Journal of Linguistics* 1: 55–69.
- Underwood, Gary N. 1974. "American English Dialectology: Alternatives for the Southwest." *International Journal of the Sociology of Language* 2: 19–40.
- Wilson, H. Rex. 1956. "The Implications of Tape Recording in the Field of Dialect Geography." *Journal of the Canadian Linguistic Association* 2: 17–21.
- Wood, Gordon R. 1971. *Vocabulary Change: A Study of Variation in Regional Words in Eight of the Southern States.* Carbondale: Southern Illinois Press.

17 Corpus-Based Approaches to Dialect Study

BENEDIKT SZMRECSANYI AND LIESELOTTE ANDERWALD

17.1 Introduction

CORPUS LINGUISTICS is a methodology that draws on (more or less) systematic collections of naturalistic, machine-readable texts to make claims about linguistic phenomena and/or linguistic variation (for textbooks see, e.g., Biber 1998; Frigina and Hardy 2014; McEnery, Xiao, and Tono 2006). Thus unlike other methodologies in linguistics—for example, those that rely on experimental data, or on elicited linguistic knowledge, or on intuitions (either the linguist’s own or somebody else’s)—corpus linguistics is the methodological outgrowth of the usage-based turn in linguistics. This is because what is of interest in corpus linguistics is what language users *do* with language (that is, their *behaviour*), not what they *know* (or *think* they know) about language. There are many different kinds of corpora: they may contain written or spoken (transcribed) language, modern or historical texts (or both), standard or non-standard language, adult language or child language, native language or learner language, and so on. It is clear that the sort of material included in a corpus will constrain the range of research questions that can be asked on the basis of the material. Here is a little taster of the sort of issues that can be tackled in a corpus-based approach: let X be some linguistic feature (a morphological marker, a lexical item, a grammatical construction, and so on) in which the researcher is interested. Corpus study may then address the following questions, among others: How—in which contexts, in which way, subject to which restrictions—is X used in a given corpus? Comparing two or more corpora sampling different registers (e.g., conversation versus academic prose), in which register is X more frequent? Comparing two or more corpora sampling historical stages (e.g. nineteenth-century English versus twentieth-century English), has X become more or less frequent over time? Comparing two or more corpora sampling different geographic language varieties, in which variety is X more frequent, that is, more widely used? Using a sociologically annotated corpus, is X more frequently used by male or by female speakers, by younger or by older speakers, by speakers from lower or from higher social classes?

A little case study may illustrate some of these points. Most native speakers of English will have intuitions about the negator *ain’t*. What can corpora tell us about the contexts in which *ain’t* occurs? A query on the *Corpus of Contemporary American English* (COCA)

(available, like all web-based corpora mentioned in this section, on Mark Davies' corpus portal at <http://corpus.byu.edu>; see Davies 2010) reveals that *ain't* occurs as negated form of *have*, as in (1); as negated form of *be*, as in (2); and as negated form of *do*, as in (3).

1. Age *ain't* got nothing to do with living or dying (COCA Bk:EdgeDarkWater)
2. If it *ain't* broke, why fix it? (COCA NPR_TalkNat)
3. Rose, you *ain't* see she's a woman? (COCA Bk:GirlGolden)

Also, COCA samples a number of different registers—spoken registers, fiction, magazine prose, newspaper prose, and academic prose. It turns out that in the COCA material, *ain't* is most popular in fiction (99 occurrences per million words [pmw]); *ain't* is least popular in academic prose, where it occurs only 3 times pmw. COCA is moreover a so-called monitor corpus—the earliest material it contains dates from the 1990s, and the corpus is being continually updated. A look at the diachronic frequency trajectory of *ain't* reveals that the form is on the decline in COCA: from 39 occurrences pmw in the 1990–1994 period to 22 occurrences pmw in the 2010–2015 period. What about regional variation? According to the *Corpus of Global Web-Based English* (GloWbE), *ain't* is most widely used in US American English (frequency: 25 occurrences pmw), and least widely used in Pakistani English (frequency: < 3 occurrences pmw). British web-based texts take the middle road: here *ain't* occurs about 13 times pmw. Thus, contrary to what some Britons may believe, *ain't* is not an Americanism. Yes, the marker is very popular in American English, but we also do find it in British varieties of English. In fact, we know from corpus analysis that *ain't* is used all over England (Anderwald 2002, 149), and especially in traditional dialects in the South of England (Szmrecsanyi 2013, 56–58). In the realm of corpus-based *dialect* study, relevant corpora typically consist of orthographically transcribed interviews with dialect speakers, similar to sociolinguistic interviews that are customary in variationist sociolinguistics. In the remainder of this contribution, our take on dialect study is restricted to the study of TRADITIONAL DIALECTS. We essentially follow Trudgill (1990, 5) in defining traditional dialects as follows: “Traditional dialects are what most people think of when they hear the term dialect, spoken by (in Western societies at least) fewer and fewer people in ‘remote and peripheral rural areas’.” This is another way of saying that we exclude from consideration corpus-based work on variation between standard varieties (e.g., British English versus American English, Netherlandic Dutch versus Belgian Dutch), global varieties (of English, Spanish, etc.), and we will also not be dealing with “urban” or “social” dialectology, which is primarily concerned with sociolinguistically conditioned variation.

Of course, corpus-based dialect study shares many methods with neighboring disciplines. SYNCHRONIC (VARIATIONIST) SOCIOLINGUISTICS, also sometimes referred to as *social* or *urban dialectology*, is methodologically essentially the same as quantitative dialectology (the difference lying in the criteria chosen for sampling). Substantially, the focus is not typically on geographic variation, even though geography is occasionally considered (e.g., Tagliamonte, Smith, and Lawrence 2005). QUANTITATIVE TEXT LINGUISTICS in the spirit of, e.g., Biber (1988) or Mair (2006), is primarily concerned with text frequencies of linguistic phenomena in usage data. As such, quantitative text linguistics overlaps methodologically with corpus-based dialect study, especially when it comes to the rigorous methodology that guides, or should guide, corpus compilation. However, quantitative text linguists have not traditionally taken an interest in dialect data. HISTORICAL CORPUS LINGUISTICS also has methods and substance in common with corpus-based dialectology. Finally, because corpus-based dialectology relies on (transcribed) interactive interviews, it is also informed by methods and interpretational frameworks developed in DISCOURSE AND CONVERSATION ANALYSIS. For example, factors like repetition and persistence are also relevant in the

analysis of dialect corpora (Szmrecsanyi 2006). Corpus-based dialect study shares with other corpus-based approaches a focus on morphology, grammar, and discourse pragmatics; corpus-based studies in phonetics and phonology are, by contrast, considerably less widespread (but see, e.g., Rácz 2012). The reason for this bias is that many corpora currently available contain written material. And even those corpora that sample spoken language more often than not digitize orthographically transcribed words, which is convenient as long as one is not interested in pronunciation. We will come back to this problem in Section 17.6, “Future Directions.”

A recent example of a corpus-based dialect study is Anderwald’s (2009) study of non-standard past tense forms (e.g., past tense *give, come, sung, drunk, or catched*) in traditional British English dialects, based on the *Freiburg Corpus of English Dialects* (FRED) (presented in more detail in Section 17.2). Because of the limits imposed by a finite corpus, attention had to be restricted to variable verbs that occur relatively frequently in the corpus material. Even so, however, a fine regional differentiation would have resulted in many empty cells for individual locales (or for individual lexemes). Rather than aggregate data across lexemes, Anderwald chose larger regional subdivisions. This is an example of the typical trade-off situations in comparative work, where breadth of coverage and depth of investigation (be it geographical, historical, linguistic, or other) cannot be simultaneously achieved.

Even given this trade-off, in doing comparative dialect studies we already make a number of important assumptions, most importantly that of equivalence. We assume that we are investigating phenomena that can in fact be compared across dialects (in Anderwald’s case: some dialects may not have a morphological category of PAST TENSE). Under such circumstances, the research question should probably be changed into an onomasiological one (e.g., “how is reference to past events expressed in dialect X?”), and researchers will have to define clearly what they regard as “alternative ways of saying the same thing.”

Comparative work across dialects also assumes that the data we work with are formally equivalent, for example, are transcribed consistently across locales, are equivalent in size and “vernacularity,” are of acceptable audio quality, and that speakers are comparable in terms of parameters such as social class, gender, age, education, or ethnicity. In fact, Wolk (2014) demonstrates that sociolinguistic imbalances like this can distort results in a corpus-based dialectology approach—even in dialect corpora whose design is overall quite (but not perfectly) homogeneous and balanced.

In summary, a number of assumptions, prerequisites, and difficulties characterize corpus-based approaches to dialect study:

- The target phenomena have to be textually relatively frequent, as the absence of features in corpus material is hard to interpret.
- The data subject to analysis must be relatively homogeneous across locales (in terms of sociolinguistic parameters, but also in terms of length, quality of recording, quality of transcription, etc.).
- Transcription must be as reliable, faithful, and internally consistent as possible. At the same time, a certain degree of regularization is indispensable, to enable the identification of non-standard forms in the material.
- When corpus-based dialectologists adopt the variationist method, it is imperative to ensure that the variants subject to study are really equivalent ways of saying the same thing (see, e.g., Cheshire 2005; Lavandera 1978).
- Forms that look similar do not necessarily have to be the same (sometimes called camouflage constructions, e.g., Spears 1982).

Some of these issues will be revisited in Section 17.5 below.

17.2 Examples of Dialect Corpora

The research community is not exactly drowning in publicly available corpora that sample traditional dialects, as per our definition in Section 17.1, and such corpora as exist mostly cover dialects of English. For example, the *Freiburg Corpus of English Dialects* (FRED) (see Hernández 2006 for a manual) is a (more or less) synchronic (but historical) corpus—most of the material was collected in the 1970s—which covers several dialect areas in Britain, concentrating on traditional dialect speakers, many of them older males (i.e., it is not designed to be representative in terms of social class, gender, or age). FRED has been orthographically transcribed. The following is a representative extract (text CON_005, county Cornwall, Southwest of England):

- Interviewer: ...recording of Wallace Jeff Baggerly of Porthmeor Farm, near Zennor, was made on the fifth of September nineteen seventy-eight. When were you born?
- Informant: in nineteen hundred and four. Seventeenth of December.
- Interviewer: And had your -- and your family had lived here...?
- Informant: Yes, my father was born here. Not in this house!
- Interviewer: No.
- Informant: But in the old house, you know. And so was my grandfather.
- Interviewer: Yeah. And that's the old house across the road, is it?
- Informant: No, no, gone. Used to be here. You know, under this, see – or, under – somewhere, like. And that 's gone. And let 's see. His father again come up from down Lower Porthmeor.
- Interviewer: Mhm.
- Informant: Now I don't know quite how long they 'd been here, but they come from St. Hilary, somewhere. To start with, if you understand what I mean. That might 've been my great-grandfather's (pause) grandfather, perhaps. Somebody come, and a – with a baby. And that was one of the oldest old men that was here 'round, you know, but I couldn't tell you exactly which generation, you know.
- Interviewer: No.
- Informant: I do know my great-grandfather was born down Lower Porthmeor, and he had uh, one, two, three brothers. I knew them by name but I didn't know them (unclear) that well Uncle Richard and Uncle Jack, and Uncle Albert, you know, they was uncles to my grandfather, see, and, and you pick up the (unclear) sayin' from years ago.
- Interviewer: Yeah.
- Informant: That old house there, I did hear a great-uncle of mine say that he could mind somebody living in en. And that, he was, he was little. Well this is Tim's house to this day.

Note, for example, how the informant uses non-standard *was* in plural contexts (... *they was uncles* ...). Based on this and the other texts in FRED, Szmrecsanyi (2010) explores the regional distribution of non-standard *was* across the major dialect areas in Great Britain, and finds that the feature does not have a significant geographic distribution overall, although it does tend to be more frequent in traditional dialects in England than in traditional dialects in Scotland (see the box plot in Figure 17.1).

The full version of FRED is available to researchers and visiting scholars at the University of Freiburg. The 1-million word sampler version of FRED (FRED-S; see Szmrecsanyi and Hernández 2007 for a manual) is publicly available and comes with part-of-speech annotation.

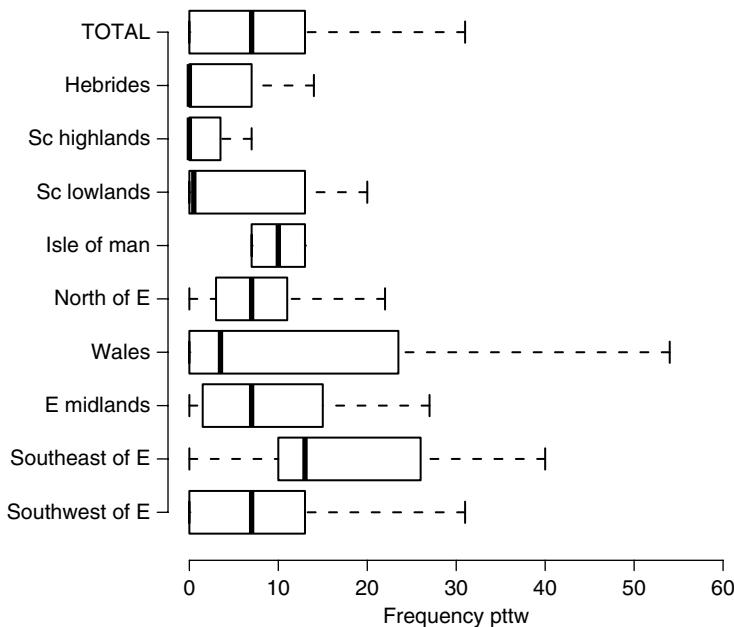


Figure 17.1 Box plot depicting frequency variance of non-standard *was* by dialect region in FRED (Szmrecsanyi 2010, 48).

In the *Diachronic Electronic Corpus of Tyneside English* (DECTE; see Corrigan, Mearns, and Moisl 2014), the locality is—by contrast to FRED—held constant: all material comes from Tyneside, but the time axis is extended, as is the social coverage. DECTE contains material from the 1960s, the 1990s (Milroy, Milroy, Hartley, and Walshaw 1994), and from the first decade of the twenty-first century. This diachronic structure allows researchers to compare dialect forms over time, and to a degree to comment on their social distribution (cf. Beal, Burbano-Elizondo, and Llamas 2012). Features that have been analyzed so far on the basis of DECTE include negation (Beal and Corrigan 2005), relativisation (Beal and Corrigan 2002; 2006), intensifiers (Barnfield and Buchstaller 2010), and phonetic variation (Corrigan, Mearns, and Moisl 2014).

A corpus that is sometimes used to investigate the regional distribution of widespread features of spoken English is the spoken material in the British National Corpus (BNC, cf. Aston and Burnard 1998), with about 5 million words of unmonitored everyday speech. However, the material was not originally intended to be regionally representative, so that both size, quality, and social make-up of the material differ significantly across regions, and the transcription was not produced by linguists. Nevertheless, studies based on the BNC have produced interesting regional results, for example, for features such as pseudopassives (Klemola 1999), the system of the English verb phrase quite generally (Sampson 2002), various features of negation (Anderwald 2002), or ditransitives (Gerwin 2013).

Important contrastive work is also carried out by sociolinguists and dialectologists on materials that are typically not made available to outsiders. Tagliamonte has collected (or supervised the collection of) corpora of York English (e.g., Tagliamonte 1998; 2001), Devon English (e.g. Godfrey and Tagliamonte 1999), Toronto English (e.g., Tagliamonte and D'Arcy 2007), or Samaná English (e.g. Poplack and Tagliamonte 2001), and typically feature analyses are compared across several of these materials (cf. Tagliamonte 2012b). Tagliamonte has also

collected materials from isolated, peripheral dialect communities in Northern Britain and Ireland in her *Roots of English* project (Tagliamonte 2012a), specifically from localities in the Northwest of England, Lowland Scotland, and Northern Ireland. Her morphosyntactic analyses include subject-verb concord, the use of adverbs without *-ly*, negative versus auxiliary attraction, relative pronouns, *that*-complementation, *for to* infinitives, future reference, the competition of past tense and present perfect, modals of obligation, possessive *have* versus *have got*, discourse marker *like* and general extenders (e.g., *and stuff like that*) (all in Tagliamonte 2012a). Rather than comparing surface structures or relative frequencies, in her “comparative sociolinguistics” approach (Tagliamonte 2012b) multivariate analyses serve to uncover the relevance of extralinguistic and intralinguistic constraints, and it is the order and strength of these constraints—rather than absolute or relative frequencies—that serve as the point of comparison across dialects (already in Poplack and Tagliamonte 2001). (Note also that many of the features she investigates are actually (variable) features of Standard English, rather than dialectal in the strict sense.)

As for dialects of languages other than English, corpus resources are still comparatively rare, but the situation is improving. Consider, for example, the *Corpus Gesproken Nederlands* (CGN; see <http://lands.let.ru.nl/cgn/ehome.htm>), which can be utilized for dialectological analysis (e.g., Hoste, Gillis, and Daelemans 2000); or the *Estonian Dialect Corpus* (see <http://www.murre.ut.ee/murdekorpus/> and Uiboaed *et al.* 2013); or the “*Phonologie du Français Contemporain*” [Phonology of contemporary French] project (see <http://www.projet-pfc.net/> and Durand 2006); there is also a project on “*Phonologischer Wandel am Beispiel der alemannischen Dialekte Südwestdeutschlands im 20.Jahrhundert*” [Phonological change in Alemannic dialects in Southwest Germany], which among other things draws on spontaneous-conversational corpus material to explore dialectal change (see Auer, Baumann, and Schwarz 2011; Streck and Auer 2012).

17.3 Research Questions

In our overview above several research questions have already been implicit. Making them explicit, we distinguish in what follows (1) geolinguistics proper, serving synchronic, diachronic or typological interests, (2) aggregate geolinguistics, (3) the discovery of constraint rankings, and (4) data mining.

The perhaps most time-honored motivation for performing comparative dialectology is the quest for areal patterns, something we may call GEOLINGUISTICS PROPER. Just as we are able to draw isoglosses of phonetic variants on maps, we can proceed in the same way for features of morphosyntax, and thus directly correlate linguistic and geographical information. The underlying motivation for discovering areal patterns might be manifold. As Kehrein (2012) has pointed out, areal patterning can be taken as indicative of different paths of language change from a (purported) original single source; geolinguistic patterns are then of interest in a structuralist, intralinguistic way, indicating the (synchronic) range of variation possible in one language. Kehrein quotes German Neogrammarian linguists of the nineteenth century as examples of this research paradigm. A more recent, but essentially very similar avenue of interpreting areal patterns is found in interpretations of a functional-typological kind (as in Kortmann 2002; Kortmann, Pietsch, Herrmann, and Wagner 2005). Here the interest is in discovering the breadth of variation synchronically, and in identifying dominant and minority patterns. Once dominant patterns are identified, they are then interpreted in terms of their (systemic or psycholinguistic) function, and often compared with patterns observable in other languages. In a cross-dialectal and cross-linguistic comparison, for example, it becomes quickly apparent that many features (such as multiple negation, the lack of marking adverbs, or the use of invariant tag questions) of non-standard English are

perfectly “normal” in a world-wide perspective, and that it is Standard English which stands out in not following these unmarked typological trends.

In a related manner of interpretation, geolinguistic patterns always held interest for historical linguistics, because they can also be read as indicating individual stages in diachrony. Geolinguistic variation is then seen as making visible the process of language change, as relic areas preserve older forms. Many maps in the Survey of English Dialects (SED) have been interpreted in this way (cf. the introductory chapter in Orton, Sanderson, and Widdowson 1978). The interest might then shift to extralinguistic factors that may have caused the observed distributions, moving into socio-cultural terrain. Thus, researchers have asked about settlement patterns, as in the Linguistic Atlas Projects of the United States (e.g., Kurath 1949), where fine-grained dialect divisions on the Eastern seaboard “fan out” toward the west into a more homogeneous area. Transport and communication networks have been invoked in the interpretation of relic areas (for the East Anglian Fens, cf. Britain 2002), or in the geographical spread of new variants (cf. Labov 2010, chapter 10). In England, the status of London and the wider Southeast as a source of leveling, and a source of innovations that spread to the rest of the country, has also been apparent through consistent patterns in dialect maps (e.g., for historical phenomena like H-dropping, the loss of postvocalic R, or L-vocalization, all in Orton *et al.* 1978). In reverse, geographical distance, but also a sociocultural sense of isolation have been invoked to explain the status of relic areas such as the Northeast of England (cf. Beal 2004). However, even in mainstream dialectology more modern conceptualizations of geography as imagined spaces (e.g., Zelinsky 1973) have not really entered mainstream publications yet.

The research cited in the foregoing discussion is primarily concerned with the geolinguistic patterning of individual phenomena. However, dialect areas are typically constituted by the bundling of isoglosses. More sophisticated methods to conduct aggregate corpus-based dialectology are now available to model variable rather than categorical features, and several rather than single ones (see, e.g., Szemrecsanyi 2013; Wolk 2014). Such methods may address new research questions, such as: To what degree does dialectal similarity or dialectal distance correlate with geographic distance? What sort of geographic distance counts—as-the-crow-flies distance, travel time, travel distance, and so on? Which features contribute to specific geographic patterns, or, in Szemrecsanyi’s words, “how do features gang up to create layered areal patterns” (Szemrecsanyi 2013, 137)?

The comparative sociolinguistics paradigm (Tagliamonte 2012a,b), which empirically relies exclusively on usage (corpus) data, has already been mentioned. Here, instead of the identification of areal patterns, dialects are compared quantitatively for their underlying CONSTRAINT RANKINGS that determine the observed surface patterns. The quantitative analysis thus yields qualitative differences, and identical rankings are taken to indicate close historical relations, typically supplemented by socio-historical evidence (for a critique, cf. Pietsch 2012).

The idea of identifying “underlying” patterns is perhaps reminiscent of the generative enterprise. However, it has to be said that a systematic investigation of dialect differences has not been at the core of Generative Grammar. While different language systems are occasionally included in generative arguments (e.g. Halle and Mohanan 1985), this is typically not based on preceding quantitative analyses, but on introspection, or access to individual informants. This is understandable, given the deep-seated scepticism in generative circles concerning the validity of corpus linguistics (Chomsky 1956; 1957; Miller and Chomsky 1963). Where dialect material is collected and compared, the interest is essentially in uncovering differences in the deep structure (e.g., Adger and Smith 2005 for a northeastern Scottish community; Henry 1995 for Belfast English; Tubau Muntañá 2008 for multiple negation in Britain), and qualitative differences are usually taken as more important than quantitative ones.

17.4 Methods

Against the backdrop of the research questions discussed in the previous section, three major methods in corpus-based dialect study may be distinguished: (1) qualitative example mining, (2) quantitative single-feature studies, and (3) quantitative multi-feature studies. Let us discuss these in turn.

QUALITATIVE EXAMPLE MINERS tap into dialect corpora to obtain evidence of the attestedness of particular linguistic features in particular dialects. Relevant examples include Henry (1995) for Belfast English constructions, or Tubau Muntaña (2008) for multiple negation in FRED.

QUANTITATIVE SINGLE-FEATURE STUDY uses quantitative methods to investigate one feature at a time (see Nerbonne 2009, 176–1177 for a critical discussion); the contribution by Anderwald (2009) mentioned above, but also those studies collected in Kortmann, Herrmann, Pietsch, and Wagner (2005) or Hernández, Kolbe, and Schulz (2011) are representative of recent work in this spirit on the grammar of traditional British English dialects. We can more specifically distinguish two variants of this approach: FREQUENCY-FOCUSED SINGLE-FEATURE STUDY, and CONSTRAINT-FOCUSED SINGLE-FEATURE STUDY. In frequency-focused single-feature study, dialectologists determine usage frequencies of particular features. In this endeavour, increased frequency is typically considered a proxy of a feature's entrenchment and/or overall importance in a particular dialect grammar. Representative examples of this approach include Anderwald (2009) discussed in Section 17.1, or Herrmann (2005) on relativisation strategies in FRED. CONSTRAINT-FOCUSED SINGLE-FEATURE STUDY is the sort of multivariate approach that characterizes variationist sociolinguistic work by Tagliamonte and collaborators mentioned above. In this line of analysis the question is "When dialect speakers have a choice between two ways of saying the same thing, which factors (language-internal or language-external) constrain their choice"? Addressing this question necessitates using multivariate analysis methods such as binary logistic regression (e.g., Varbrul). Representative work in this tradition includes Pietsch (2005), who is interested in the conditioning of verbal agreement patterns in Northern dialects of English, or Tagliamonte and Smith (2005), who study complementizer *that* retention and omission in British English dialects. Even though constraint-focused variationist (socio)linguists do not necessarily consider themselves corpus linguists, since their work is based on collections of authentic language usage (i.e., corpora), we do consider their work relevant here.

In QUANTITATIVE MULTI-FEATURE STUDY (a.k.a. CORPUS-BASED DIALECTOMETRY), analysts base claims not on the distribution of one particular feature, but of many. Quantitative multi-feature study thus adopts dialectometrical methods (see Chapter 7). The goal is to obtain a more robust geolinguistic signal; this signal can then be projected to geography in sophisticated exploratory maps, and/or correlated with language-external measures such as geographic distance. Unlike in traditional dialectometry (Séguy 1971; Goebel 1982; Nerbonne, Heeringa, and Kleiweg 1999), the primary data are not contained in dialect atlases or surveys but come from dialect corpora. Two ways of doing corpus-based dialectometry may be distinguished: TOP-DOWN CORPUS-BASED DIALECTOMETRY, and BOTTOM-UP CORPUS-BASED DIALECTOMETRY. The top-down approach first defines a feature catalogue, then establishes frequencies (Szmrecsanyi 2013) or probabilities (Wolk 2014) associated with these features, and subsequently calculates a joint measure of pairwise linguistic distances between the dialects considered. For example, Szmrecsanyi (2013) explores the extent to which grammatical variation in British English dialects is structured geographically—and thus, is sensitive to the likelihood of social contact. The study combines corpus-based variation studies with aggregative-dialectometrical analysis and visualization methods. This synthesis is desirable for two reasons. First, dialects are multidimensional, and hence call for aggregate analysis techniques. Second, compared to linguistic atlas material, corpora yield a

more radically usage-based frequency signal. Against this backdrop, Szmrecsanyi calculates an aggregate measure of dialect distance based on the discourse frequency of 57 morphosyntactic features, such as multiple negation, non-standard verbal -s (e.g., *so I says*, *What have you to do?*), or non-standard weak past tense and past participle forms (e.g., *they knowed all about these things*) in FRED (see Section 17.2). The ultimate aim is to reveal large-scale patterns of grammatical variability in traditional British English dialects. Referring back to the research questions in Section 17.3, Szmrecsanyi's study shows that it is impossible to find in England a clearly demarcated Midlands dialect area on grammatical grounds, and that travel time is a better predictor of linguistic distance than as-the-crow-flies geographic distance. In a broadly similar vein, Grieve (2009; see also 2011, 2012)¹ is interested in regional grammatical variation in American English. He defines a feature catalogue spanning 45 (standard English) grammatical variables, and examines their usage rates in a huge corpus of letters to the editor in 200 cities from across the United States. Contrary to what old-school dialectologists may have suspected, Grieve demonstrates that his rather unorthodox, written material indeed exhibits geolinguistic patterns.

In BOTTOM-UP CORPUS-BASED DIALECTOMETRY, by contrast, features are not defined *a priori*, but are allowed to emerge in a data-driven fashion. In this spirit, Wolk (2014) uses a part-of-speech-annotated version of FRED, and develops a probabilistically enhanced method (based on Nerbonne and Wiersma 2006) that draws on part-of-speech bigram frequencies (e.g., sequences of determiner-noun) to calculate an aggregate measure of dialect distance. The resulting geolinguistic signal is weaker than that yielded by top-down approaches, but it does uncover dialectologically meaningful areal patterns.

17.5 Issues and Problems

We have already implicitly hinted at potential problems and issues with the corpus-based study of non-standard materials, especially those to do with transcription and normalization regimes. Consistency of transcription, devising normalized spellings for non-standard items that enable some computer-readability without making too many theoretical assumptions, and the internal make-up of subsamples—in particular interaction between social and regional variation—are potential problem areas that researchers have to be aware of. We have also pointed out that all corpora (through being finite resources) impose a qualitative limit on what can be investigated: only features that are frequent enough can be included in comparative analyses. Arbitrary cut-off points can here lead to inter-researcher differences. What was investigated by Anderwald (2009) (past tense *drunk*, *sung*, *rung*) was excluded by Szmrecsanyi (2013) on the grounds that this feature just did not make the lower frequency threshold. The role of frequency also leads to the more general problem of how to deal with extremely rare, or even absent, features. The absence of a feature from a corpus can be a sign of its overall rarity: thus past or modal perfect progressive passive forms (e.g., *would have been being charged*) even in a huge Standard English corpus like COCA (see Section 17.1) are only attested four times, and we would thus expect that in smaller corpora these complex verb phrases, though no doubt in principle possible in English, would not occur at all. Although this problem related to size has increasingly been counteracted by building larger and larger corpora, the automation processes typically employed are not really feasible for non-standard materials, and corpora of several hundred million words like COCA are probably not a realistic goal when it comes to dialect material. This means that empirically, absence due to low text frequency is difficult to distinguish from truly ungrammatical forms (the problem of “negative evidence”). As dialectologically relevant examples, *was sat/stood* with progressive meaning, or resumptive relative pronouns (of the type *the house which he saw it*), despite being regularly cited in the dialectological literature, are not (or only very infrequently) found in FRED.

Other features are not only rare overall, but occur in such specific discourse contexts that the corpus material perhaps does not provide for their occurrence. Thus it has been observed that the English “hot-news” perfect or habitual constructions, which are both only used in pragmatically marked situations, are conspicuously absent from FRED (Anderwald and Wagner 2007). The reverse of this dependency on frequency may also be a problem, though, if a researcher only investigates what one can investigate safely, and thus lets the corpus material dictate the research question. Trivially counting for the sake of counting, possibly even without a working hypothesis, is an inherent danger in all corpus linguistics, and we only note it here for the sake of completeness, politely refraining from mentioning actual examples.

Particularly relevant for comparative dialectological work, we note the possible mistake of investigating as dialect (or even wider) universals what is trivially (namely historically) given in all varieties. Thus, in a review of comparative articles on a range of varieties of English around the world, Trudgill notes that “the fact that nonstandard dialects of English have many similarities ... is really of no great interest ... and to attempt to extrapolate universal principles out of the commonality is to credit the similarities with more importance than actually they have” (Trudgill 2013, 87), citing multiple negation, *there's* followed by plural NPs, or present participles in <-in> as examples.

Finally, we list here three more areas where we would claim that corpus-based dialect studies are probably not very useful, besides those rare and those pragmatically marked features noted above: the investigation of very local (i.e., areally skewed) features, the investigation of categorical features, and the investigation of features where surface similarities mask deeper differences (as noted above in Section 17.2). Areally skewed features, although perhaps in some respects the most interesting ones in terms of linguistic geography, do not lend themselves well to *comparative* quantitative analysis, mainly for technical reasons, because the resulting empty cells for many areas act as knock-out constraints in variable analyses. The same can be said for categorical (non-variable) features, which also do not lend themselves well to a comparative analysis of *variability*. As our final point, surface similarities that are due to underlying differences result in comparing apples and oranges, as we have noted above (although it is not always a trivial matter to find out what constitutes the apple, and what the orange ...).

17.6 Future Directions

Corpus-based approaches to dialect study are currently being refined in the following ways. For one thing, the foregoing discussion mentioned social imbalances in corpus material as a nuisance factor in corpus-based dialect study. But as a matter of fact, corpus analysts are beginning to explore the exciting opportunities that corpora offer with regard to the interface between dialectology and sociolinguistics. In this line of work, geographically conditioned patterns still take centre stage, but thanks to corpora which do not exclusively sample non-mobile old rural males (NORMs) we can increasingly explore the extent to which geographic patterns are different when our attention is restricted to male or female speakers, old or young speakers, and so on. An exemplary study highlighting the potential of interface explorations along these lines is Heeringa and Hinskens (2014), who tap into a parallel corpus database and exploit social differences between their informants to study dialect change in the Dutch language area in apparent time.

Secondly, future work is likely to advance bottom-up approaches in the spirit of Nerbonne and Wiersma (2006) and Wolk (2014). Bottom-up corpus analysis is actually quite common in, for example, phraseology and collocation research, but in corpus-based dialectology the potential afforded by bottom-up analysis is as yet underexplored. The possibility of bottom-up analysis is actually what most radically sets apart corpus-based dialectology from dialectology

based on other data sources, such as dialect atlases (although some bottom-up approaches are used here, too, as in Kretzschmar's "self-organizing" maps, e.g., Kretzschmar 2011).

Third, observe that previous corpus-based dialectology research is overwhelmingly grammar-centred (much like corpus linguistics in general has a bias towards grammar and morphology, as we noted in Section 17.1). But many of the dialect corpora that this research draws on also provide audio material, which in most cases still awaits systematic phonetic analysis, both auditory and acoustic (along the lines of Grieve 2014). Lexis is likewise a neglected domain in corpus-based dialect studies, but this neglect is primarily due to the fact that lexical research requires large corpora, and conventional dialect corpora are simply not large enough for lexical analysis. (Also, of course, corpus-based dialectology was invented to address shortcomings in traditional dialectology, and traditional dialectology has always had a strong focus on lexis, besides phonetics.)

Fourth, corpus-based dialectology has so far only marginally interpreted its results in a wider societal context, and collaborations with cultural studies for informed sociocultural analyses are a desideratum. We can imagine fruitful combinations of corpus-based dialect study with

1. perceptual dialectology, which would allow a third kind of geographical distance to be included in the analysis, that is, the "perceived" distance between locales (as in Montgomery and Beal 2011);
2. linguistic anthropology, for example, for a study of attitudes, the covert and overt prestige of features, or for aspects of identity construction, especially at the group level (perhaps through an in-depth study of the meta-reflexive enregisterment of individual features in speakers' awareness of their own variety versus the varieties of others, along the lines of Johnstone, Andrus, and Danielson 2006 for Pittsburgh English);
3. social history, to address issues such as migration and settlement patterns, urbanization patterns, the spread of standard languages through education, and so on.

And finally, we note that there is an in principle well-known overlap between the sorts of research questions asked in dialectology, on the one hand, and in crosslinguistic typology on the other hand (see, e.g., the papers in Kortmann 2004). Recent years have seen some methodological convergence, in that some crosslinguistic typologists now increasingly rely on (parallel) corpus databases, instead of decontextualized reference grammars or individual expert informants (consider the papers in Szmrecsanyi and Wälchli 2014). It will be worthwhile to further explore these methodological interfaces, for the sake of developing a more unified discipline of geolinguistics, and to contribute to a unified study of intra- and crosslinguistic variation.

NOTE

- 1 Even though Grieve (2009) is not concerned with "traditional" dialects as defined in Section 17.1, we include this work in our survey due to its methodological innovativeness, and because his results nevertheless produce clear geographical patterns

REFERENCES

Adger, David and Smith, Jennifer. 2005. Variation and the Minimalist Program, In Cornips, L. & Corrigan, K. (eds.), *Syntax and Variation:*

Reconciling the Biological and the Social, John Benjamins, Amsterdam & Philadelphia, pp. 149–178.

- Anderwald, Lieselotte. 2002. *Negation in Non-Standard British English: Gaps, Regularizations and Asymmetries*, Routledge, London & New York.
- Anderwald, Lieselotte. 2009. *The Morphology of English Dialects: Verb-Formation in Non-Standard English*, Cambridge University Press, Cambridge.
- Anderwald, Lieselotte and Wagner, Susanne. 2007. FRED - The Freiburg English Dialect corpus, In Beal, J. C., Corrigan, K. P. & Moisl, H. (eds.), *Creating and Digitizing Language Corpora*, Macmillan, London, pp. 35–53.
- Aston, Guy and Burnard, Lou. 1998. *The BNC Handbook: Exploring the British National Corpus with SARA*, Edinburgh University Press, Edinburgh.
- Auer, Peter, Baumann, Peter and Schwarz, Christian. 2011. Vertical vs. horizontal change in the traditional dialects of southwest Germany. A quantitative approach. *Taal en Tongval* 63(1).
- Barnfield, Kate & Buchstaller, Isabelle. 2010. Intensifiers on Tyneside: Longitudinal developments and new trends. *English World-Wide* 31: 252–287.
- Beal, Joan C. 2004. "Geordie Nation": Language and regional identity in the Northeast of England. *Lore & Language* 17: 33–48.
- Beal, Joan C., Burbano-Elizondo, Lourdes, and Llamas, Carmen. 2012. *Urban North-Eastern English: Tyneside to Teesside*, Edinburgh University Press, Edinburgh.
- Beal, Joan C. and Corrigan, Karen P. 2002. Relativisation in Tyneside and Northumbrian English, In Poussa, P. (ed.), *Relativisation on the North Sea Littoral*, Lincom, Munich, pp. 125–134.
- Beal, Joan C. and Corrigan, Karen P. 2005. "No, nay, never": Negation in Tyneside English, In Iyeiri, Y. (ed.), *Aspects of English Negation*, John Benjamins, Amsterdam & Philadelphia and Yushodo University Press, Tokyo, pp. 139–156.
- Beal, Joan C. and Corrigan, Karen P. 2006. A tale of two dialects: Relativization in Newcastle and Sheffield, In Filppula, M., Klemola, J., Palander, M. & Penttilä, E. (eds.), *Dialects Across Borders: Selected Papers from the 11th International Conference on Methods in Dialectology (Methods XI)*, Joensuu, August 2002, Cambridge University Press, Cambridge, pp. 211–229.
- Biber, Douglas. 1988. *Variation Across Speech and Writing*, Cambridge University Press, Cambridge.
- Biber, Douglas. 1998. *Corpus Linguistics: Investigating Language Structure and Use*, Cambridge University Press, Cambridge.
- Britain, David. 2002. Diffusion, levelling, simplification and reallocation in past tense BE in the English Fens. *Journal of Sociolinguistics* 6: 16–43.
- Cheshire, Jenny. 2005. Syntactic Variation and Beyond: Gender and Social Class Variation in the Use of Discourse-new Markers. *Journal of Sociolinguistics* 9(4): 479–508.
- Chomsky, Noam. 1956. Three models for the description of language. *Transactions on Information Theory* 2: 113–124.
- Chomsky, Noam. 1957. *Syntactic Structures*, Mouton, The Hague.
- Corrigan, Karen P., Mearns, A.J. and Moisl, Hermann. 2014. Feature-based Versus Aggregate Analyses of the DECTE Corpus: Phonological and Morphological Variability in Tyneside English, In Szemrecsanyi, B. and Wälchli, B. (eds.), *Cross-Linguistic and Language-Internal Variation in Text and Speech*, Walter de Gruyter, Berlin, pp. 113–149.
- Davies, Mark. 2010. More than a peephole: Using large and diverse online corpora. *International Journal of Corpus Linguistics* 15: 405–411.
- Durand, Jacques. 2006. Mapping French Pronunciation: The PFC project, In Montreuil, J.-P. (ed.), *New Perspectives on Romance Linguistics*. Vol. 2: Phonetics, Phonology and Dialectology. John Benjamins, Amsterdam & Philadelphia, pp. 65–82.
- Friginal, Eric and Hardy, Jack. 2014. *Corpus-Based Sociolinguistics. A Guide for Students*, Routledge, New York.
- Gerwin, Johanna. 2013. 'Give it me!': Pronominal ditransitives in English dialects. *English Language and Linguistics* 17: 445–463.
- Godfrey, Elizabeth and Tagliamonte, Sali. 1999. Another piece for the verbal -s story: Evidence from Devon in southwest England. *Language Variation and Change* 11: 87–121.
- Goebl, Hans. 1982. *Dialektometrie: Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie*, Österreichische Akademie der Wissenschaften, Wien.
- Grieve, Jack. 2009. *A Corpus-Based Regional Dialect Survey of Grammatical Variation in Written Standard American English*, PhD Dissertation, Northern Arizona University.
- Grieve, Jack. 2011. A regional analysis of contraction rate in written Standard American English, *International Journal of Corpus Linguistics* 16: 514–546.
- Grieve, Jack. 2012. A statistical analysis of regional variation in adverb position in a corpus of written Standard American English.

- Corpus Linguistics and Linguistic Theory* 8: 39–72.
- Grieve, Jack. 2014. A Comparison of Statistical Methods for the Aggregation of Regional Linguistic Variation, In Szemrecsanyi, B. & Wälchli, B. (eds.), *Cross-Linguistic and Language-Internal Variation in Text and Speech*, Walter de Gruyter, Berlin, pp. 53–88.
- Halle, Morris and Mohanan, Karuvannur P. 1985. Segmental phonology of Modern English. *Linguistic Inquiry* 16: 57–116.
- Heeringa, Wilbert and Hinskens, Frans. 2014. Convergence Between Dialect Varieties and Dialect Groups in the Dutch Language Area, In Szemrecsanyi, B. & Wälchli, B. (eds.), *Cross-Linguistic and Language-Internal Variation in Text and Speech*, Walter de Gruyter, Berlin, pp. 26–52.
- Henry, Alison. 1995. *Belfast English and Standard English: Dialect Variation and Parameter Setting*, Oxford University Press, New York & Oxford.
- Hernández, Nuria. 2006. *User's Guide to FRED*, University of Freiburg, Freiburg.
URN:nbn:de:bsz:25-opus-24895, URL: <http://www.freidok.uni-freiburg.de/volltexte/2489/>.
- Hernández, Nuria, Kolbe, Daniela and Schulz, Monika Edith. 2011. *A Comparative Grammar of British English Dialects: Modals, Pronouns and Complement Clauses*. Mouton de Gruyter, Berlin & New York.
- Herrmann, Tanja. 2005. Relative Clauses in English Dialects of the British Isles, In Kortmann, B., Herrmann, T., Pietsch, L. and Wagner, S. (eds.), *A Comparative Grammar of British English Dialects: Agreement, Gender, Relative Clauses*, Mouton de Gruyter, Berlin & New York, pp. 21–124.
- Hoste, Veronique, Gillis, Steven and Daelemans, Walter. 2000. A rule induction approach to modeling regional pronunciation variation, In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, Association for Computational Linguistics, pp. 327–333.
- Johnstone, Barbara, Andrus, Jennifer and Danielson, Andrew E. 2006. Mobility, indexicality, and the enregisterment of 'Pittsburghese'. *Journal of English Linguistics* 34: 77–104.
- Kehrein, Roland. 2012. Linguistic atlases: Empirical evidence for dialect change in the history of languages, In Hernández-Campoy, J. M. & Conde-Silvestre, J. C. (eds.), *The Handbook of Historical Sociolinguistics*, Wiley-Blackwell, Malden, MA & Oxford, pp. 480–500.
- Klemola, Juhani. 1999. *Still sat in your car? Pseudopassives with sat and stood and the history of non-standard varieties of English English*. *Sociolinguistica* 13: 129–140.
- Kortmann, Bernd. 2002. New prospects for the study of English dialect syntax: impetus from syntactic theory and language typology, In Barbiers, S., Cornips, L. and Kleij, S. v. d. (eds.), *Syntactic Microvariation*, Meertens Institute, Amsterdam, pp. 185–213.
- Kortmann, Bernd (ed.). 2004. *Dialectology Meets Typology: Dialect Grammar from a Cross-Linguistic Perspective*, Mouton de Gruyter, Berlin & New York.
- Kortmann, Bernd, Pietsch, Lukas, Herrmann, Tanja and Wagner, Susanne. 2005. *A Comparative Grammar of English Dialects: Agreement, Gender, Relative Clauses*, Mouton de Gruyter, Berlin & New York.
- Kretzschmar, William A., Jr. 2011. The beholder's eye: Using self-organizing maps to understand American dialects, In Adams, M. and Curzan, A. (eds.), *Contours of English and English Language Studies*, University of Michigan Press, Ann Arbor, Mi., pp. 53–70.
- Kurath, Hans. 1949. *A Word Geography of the Eastern United States*, University of Michigan Press, Ann Arbor, Mi.
- Labov, William. 2010. *Principles of Linguistic Change*, Wiley Blackwell, Malden, MA & Oxford.
- Lavandera, Beatriz. 1978. Where does the sociolinguistic variable stop? *Language in Society* 7: 171–182.
- Mair, Christian. 2006. *Twentieth-century English: History, Variation, and Standardization*, Cambridge University Press, Cambridge.
- McEnery, Tony, Xiao, Richard and Tono, Yukio. 2006. *Corpus-based Language Studies: An Advanced Resource Book*, Routledge, New York.
- Miller, George A. and Chomsky, Noam. 1963. Finitary models of language users, In Luce, R. D., Bush, R. R. and Galanter, E. (eds.), *Handbook of Mathematical Psychology*, Wiley, New York, pp. 419–491.
- Milroy, James, Milroy, Lesley, Hartley, Sue and Walshaw, David. 1994. Glottal stops and Tyneside glottalization: Competing patterns of variation and change in British English. *Language Variation and Change* 6: 327–357.
- Montgomery, Chris and Beal, Joan C. 2011. Perpetual dialectology, In Maguire, W. & McMahon, A. (eds.), *Analysing Variation in English*, Cambridge University Press, Cambridge, pp. 121–148.
- Nerbonne, John. 2009. Data-driven Dialectology. *Language and Linguistics Compass* 3 (1): 175–198.
- Nerbonne, John, Heeringa, Wilbert and Kleiweg, Peter. 1999. Edit Distance and Dialect Proximity, In Sankoff, David & Kruskal, Joseph (eds.), *Time*

- Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison, CSLI Press, Stanford, pp. v–xv.
- Nerbonne, John and Wiersma, Wybo. 2006. A Measure of Aggregate Syntactic Distance, In *Proceedings of the Workshop on Linguistic Distances*, pp. 82–90.
- Orton, Harold, Sanderson, Steward and Widdowson, John. 1978. *The Linguistic Atlas of England*, Croom Helm, London.
- Pietsch, Lukas. 2005. Variable Grammars: Verbal Agreement in Northern Dialects of English, Niemeyer, Tübingen.
- Pietsch, Lukas. 2012. Verbal concord, In Hickey, R. (ed.), *Areal Features of the Anglophone World*, Mouton de Gruyter, Berlin & New York, pp. 355–378.
- Poplack, Shana and Tagliamonte, Sali. 2001. *African American English in the Diaspora*, Blackwell, Oxford.
- Rácz, Péter. 2012. Operationalising salience: definite article reduction in the North of England. *English Language and Linguistics* 16: 57–79.
- Sampson, Geoffrey. 2002. Regional variation in the English verb qualifier system. *English Language and Linguistics* 6: 17–30.
- Séguy, Jean. 1971. La Relation Entre La Distance Spatiale et La Distance Lexicale. *Revue de Linguistique Romane* 35: 335–57.
- Spears, Arthur. 1982. The semi-auxiliary *come* in Black-English Vernacular. *Language* 58: 850–872.
- Streck, Tobias and Auer, Peter. 2012. Das raumbildende Signal in der Spontansprache. *Dialektometrische Untersuchungen zum Alemannischen in Deutschland*. *Zeitschrift für Dialektologie und Linguistik* 79(2): 149–188.
- Szmrecsanyi, Benedikt. 2006. *Morphosyntactic Persistence in Spoken English: a Corpus Study at the Intersection of Variationist Sociolinguistics, Psycholinguistics, and Discourse Analysis*, Mouton de Gruyter, Berlin & New York.
- Szmrecsanyi, Benedikt. 2010. *The morphosyntax of BrE dialects in a corpus-based dialectometrical perspective: feature extraction, coding protocols, projections to geography, summary statistics*. URN: urn:nbn:de:bsz:25-opus-73209, URL: <http://www.freidok.uni-freiburg.de/volltexte/7320/>. Freiburg. (64pp.)
- Szmrecsanyi, Benedikt. 2013. *Grammatical Variation in British English Dialects: A Study in Corpus-Based Dialectometry*, Cambridge University Press, Cambridge.
- Szmrecsanyi, Benedikt and Hernández, Nuria. 2007. *Manual of Information to accompany the Freiburg Corpus of English Dialects Sampler*, University of Freiburg, Freiburg.
- URN:urn:nbn:de:bsz:25-opus-28598, URL: <http://www.freidok.uni-freiburg.de/volltexte/2859/>.
- Szmrecsanyi, Benedikt and Wälchli, Bernhard (eds.). 2014. *Aggregating Dialectology, Typology, and Register Analysis: Linguistic Variation in Text and Speech*, Walter de Gruyter, Berlin.
- Tagliamonte, Sali. 1998. Was/were variation across the generations: View from the city of York. *Language Variation and Change* 10: 153–191.
- Tagliamonte, Sali. 2001. Come/came variation in English dialects. *American Speech* 76: 42–61.
- Tagliamonte, Sali. 2012a. *Roots of English: Exploring the History of Dialects*, Cambridge University Press, Cambridge.
- Tagliamonte, Sali. 2012b. *Variationist Sociolinguistics: Change, Observation, Interpretation*, Wiley Blackwell, Malden, MA & Oxford.
- Tagliamonte, Sali and D'Arcy, Alex. 2007. The modals of obligation/necessity in Canadian perspective. *English World-Wide* 28: 47–87.
- Tagliamonte, Sali and Smith, Jennifer. 2005. No Momentary Fancy! The Zero 'Complementizer' in English Dialects. *English Language and Linguistics* 9: 289–309.
- Tagliamonte, Sali, Smith, Jennifer, and Lawrence, Helen. 2005. 'No Taming the Vernacular!' Insights from the Relatives in Northern Britain. *Language Variation and Change* 17: 75–112.
- Trudgill, Peter. 1990. *The Dialects of England*, Blackwell, Cambridge, Mass.
- Trudgill, Peter. 2013. *Review of Areal Features of the Anglophone World*, Edited by Raymond Hickey. Berlin & Boston: De Gruyter Mouton, 2012. *Journal of Linguistic Geography* 1: 86–92.
- Tubau Muntaná, Susagna. 2008. *Negative Concord in English and Romance: Syntax-Morphology Interface Conditions on the Expression of Negation*, LOT Publications, Utrecht.
- Uiboaed, Kristel, Hasselblatt, Cornelius, Lindstrom, Liina, Muischnek, Kadri and Nerbonne, John. 2013. Variation of Verbal Constructions in Estonian Dialects. *Literary and Linguistic Computing* 28 (1): 42–62.
- Wolk, Christoph. 2014. *Integrating Aggregational and Probabilistic Approaches to Language Variation*. PhD Dissertation, University of Freiburg.
- Zelinsky, Wilbur. 1973. *The Cultural Geography of the United States*, Prentice-Hall, Englewood Cliffs, NJ.

18 Acoustic Phonetic Dialectology

ERIK R. THOMAS

18.1 Acoustics and Regional Dialects

Acoustic phonetic methods are becoming increasingly necessary in dialectology. They permit more precision than auditory coding, they reduce variability among practitioners, they facilitate replicability of studies, and, as Docherty and Foulkes (1999) note, they allow investigation of variables that cannot be coded by ear. Although well established in single-community sociolinguistic studies, they have only slowly penetrated studies comparing speech across geographical regions. Single-community studies, or “urban dialectology,” have utilized acoustic measurements to assess vowel quality since the 1970s (e.g., Labov 1989; Fridland 2000; Baranowski 2007). These studies have their roots in Labov, Yeager, and Steiner (1972), which demonstrated that acoustic measurements are feasible for examining dialectal variation. Nevertheless, with two massive exceptions—Labov, Ash, and Boberg’s (2006) *Atlas of North American English* (ANAE) and the SweDia project (Eriksson 2004; Leinonen 2010), dialect geographers have not wholeheartedly adopted acoustic techniques.

Even dialectal studies that utilize acoustic methods tend to be methodologically circumscribed. Most are limited to vowel quality. Several other kinds of linguistic variables could be examined acoustically as readily as vowel quality. Although consonants are traditionally studied auditorily, various parameters of consonants lend themselves to acoustic analysis. In addition, variation in prosody, especially for intonation, could be examined instrumentally now that effective techniques have been developed. As acoustic techniques become increasingly pervasive in studies of phonetic and phonological variation, dialect geographers will have to embrace them and expand their usage to non-vocalic variables.

Despite the slow adoption of acoustic methods for spatial variation studies, there is already a solid foundation of work to build upon. Vowel quality studies are well established for English and, to an extent, Dutch and Swedish. Researchers have demonstrated the usefulness of acoustic methods for consonantal variation. Moreover, a growing body of prosodic studies has shown how acoustic methods can illuminate prosodic variation. The field has reached a tipping point: one may now expect acoustic analysis to dominate future endeavors that examine dialectal diversity in speech sounds.

18.2 Vocalic Studies

The most popular use of acoustic analysis in dialectology has been for vowel quality. For such work, the frequencies of the first two (or, occasionally, three) formants are measured, as in Figure 18.1. Taking the readings involves several decisions. First, one must determine whether the measurements will depend on the onset and offset of the vowel or not. Measurements based on the onset and offset locations include using a single measurement point at the center of the vowel, taking measurements at one or more fractions of the distance between onset and offset, or taking measurements a designated number of milliseconds after the onset and/or before the offset. Measurements not dependent on the onset or offset usually involve taking readings at points where one or more formants reach an extreme position or change direction. See Thomas (2011) on the pros and cons of each method and how to determine where the onset and offset occur.

The number of measurement points within a vowel involves another decision. For some kinds of analyses, particularly with monophthongs, a single point may be sufficient. For diphthongs and triphthongs, more points are necessary to capture the character of the segment adequately. Even more measurement points may be needed to examine finer details of the vowel trajectory.

Linear predictive coding (LPC) is routinely used today for estimating formant frequencies at each timepoint. LPC makes formant estimation quick and easy. LPC can easily produce false readings, however. The LPC settings, especially the number of coefficients, should be adjusted for different vowels and different speakers to produce accurate formant measurements. Users should examine whether an LPC formant track coincides with the formants visible on the corresponding spectrogram every time measurements are obtained. If the survey contains different speakers, especially if both sexes are represented, formant

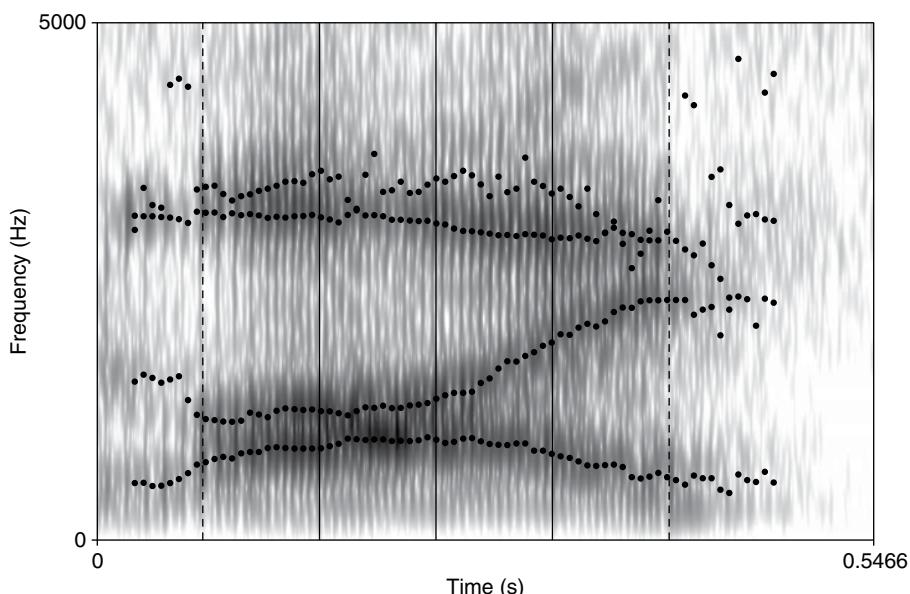


Figure 18.1 Spectrogram of a vowel with a superimposed formant track, showing possible measurement points for a token of the word *life*. Dashed vertical lines represent the onset and offset of the diphthong. Solid vertical lines represent timepoints 25%, 50%, and 75% of the distance between onset and offset.

measurements should be normalized because different speakers have oral tracts of differing lengths, leading to wide-ranging formant frequencies for vowels perceived as “the same.” Numerous normalization techniques are available: see the review in Clopper (2009). Several methods can be performed at the NORM website, <http://lingtools.uoregon.edu/norm/>. For many methods, inclusion of tokens of all the vowels or, at least, vowels filling the corners of the vowel envelope is necessary for the method to operate properly, even if the vowel being analyzed does not lie in one of the corners.

Dialectal differences in quality are assessed, for the most part, from variations in the normalized formant values. The first formant (F_1) varies inversely with vowel height and the second formant (F_2) varies directly with advancement. Lip rounding lowers formant values to varying degrees depending on the formant and the part of the vowel envelope. To examine formant trajectories, meta-analysis of normalized formant values is necessary. Such analysis might include assessing the amount of formant change from one timepoint to the next or at a particular part of the vowel.

A common way of depicting the data visually is to plot F_1 against F_2 for each speaker. Formant plots are useful for comparing features of individuals’ vowel configurations. For example, two Ohio subjects from the *Dictionary of American Regional English* (DARE) corpus are compared in Figure 18.2. In Figure 18.2, one speaker’s GOOSE and GOAT vowels are noticeably fronted (appearing more to the left) than the other’s. Dialectological work necessitates comparison of many speakers simultaneously. Formant values may be grouped into numerical ranges, which are then represented as symbols on maps. Figure 18.3 shows normalized F_2 values of the GOAT nucleus for DARE subjects and a survey of younger speakers from eastern Ohio. The horizontal line within the map indicates a division in the original settlement groups, whereas the other line indicates the area in which younger speakers were

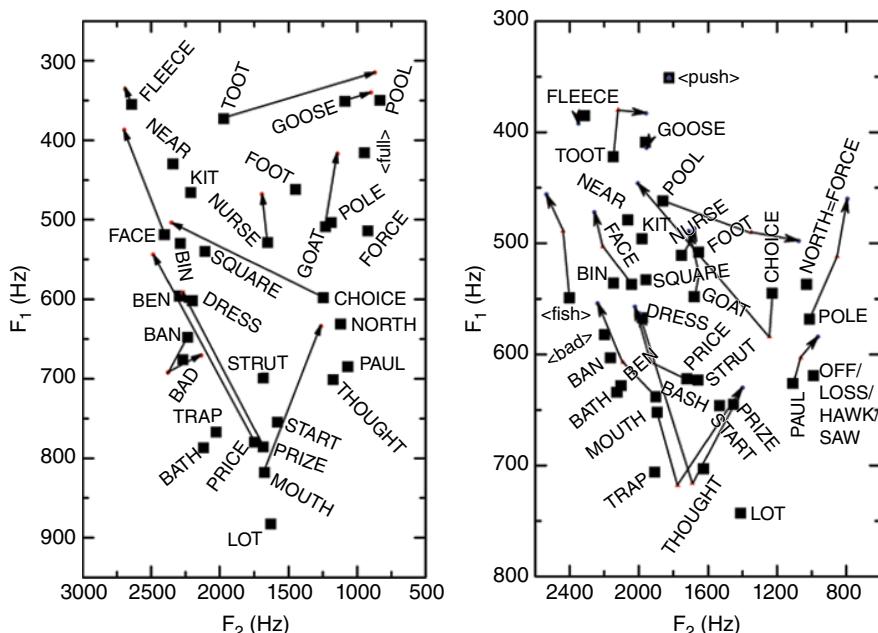


Figure 18.2 F_1/F_2 plots of mean values of the vowels of two DARE speakers, one from northern Ohio (left) and one from central Ohio (right). The relative position of the GOOSE, GOAT, and various other vowels differs. Arrows indicate the gliding of dynamic vowels. Lower-case letters represent classes with only a single token.

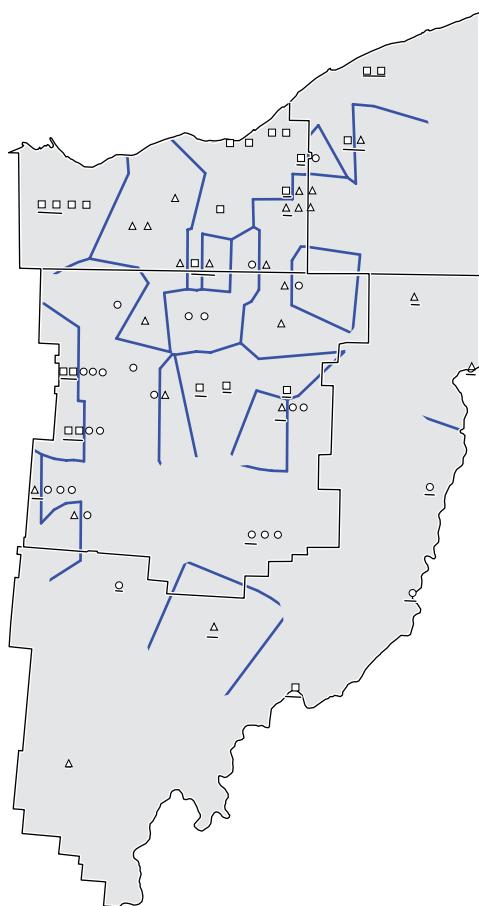


Figure 18.3 Mean normalized F_2 values of the GOAT vowel nucleus from two surveys, DARE and a sample of speakers born 1970 or later, from eastern Ohio. DARE speakers are underlined. Squares represent the most backed values, circles the most fronted, and triangles intermediate values. Younger speakers are more fronted than the DARE speakers. Circles (fronted forms) are found mostly south of the horizontal line, whereas younger speakers with backed forms (squares) are found only north of the horizontal line.

surveyed. Derivative analyses can also be plotted. The spread of a merger or allophonic differentiation could be examined by computing the difference between the sound classes undergoing merger. The spread of chain shifts could be tracked with more complex formulas.

There have been important applications of these methods to geographical variation. The use of acoustic analysis for examining geographical differences in vowel quality began with Labov *et al.* (1972). Labov *et al.* demonstrated that spectrographic measurements of the first two formants were practical for field recordings and could capture differences in vowel quality, including many that had escaped notice with auditory methods. They included subjects from several regions of the United States—mostly the Great Lakes region, the East coast, and the South—adding additional subjects from Great Britain. These data allowed them to group dialects according to similarities in vowel shifting patterns. They used the findings to formulate early versions of vowel shifting principles refined in Labov's later publications.

The most important subsequent uses of acoustic data for geographical study have been ANAE and the SweDia project. Unlike Labov *et al.* (1972), these projects are both systematic surveys, ANAE aiming to track variation in English vowel quality across the United States and Canada and SweDia accomplishing the same for Swedish. ANAE focused on major cities to make it manageable. Respondents were contacted through a random-sample telephone survey, screened to ensure that they were natives of the respective urban center, and then mailed a list of words that they were asked to read during a subsequent telephone call. The authors were then able to map the responses across North America. Some of the maps show normalized F_1 and F_2 values for each vowel. Other maps compare the extent of related vowel shifts with each other or show various mergers and splits. For example, one chapter shows maps and formant plots illustrating various elements of the Northern Cities Shift (NCS), a series of related vowel shifts found in the Great Lakes region of the United States. Many of these maps demonstrate how different vowels are geographically correlated with each other. Other chapters include similar maps for other dialect regions, such as Canada, the South, or the Midland, or examine the dialects of particular cities that have idiosyncratic vowel configurations. This study has revived the study of geographical variation in North America to some degree. It has also spawned a number of derivative studies (Boberg 2001, 2008; Labov 2007; Dinkin 2011). A handful of non-ANAE-related endeavors have used vowel formant data for geographical studies in the United States. Irons (2007) and Johnson (2010) both examined the progress of the LOT/THOUGHT merger, the former study in Kentucky and the latter along the Rhode Island/Massachusetts border, where the START class and such words as *father* and *pasta* complicated the picture. Two other studies, Ito and Preston (1999) and Gordon (2001), compared the spread of the NCS across Michigan communities. Some of my own work (Thomas 1997, 2001, 2006, 2010) has also involved acoustic analyses of geographical variation in vowels.

SweDia is a finer-meshed survey, investigating 107 rural communities across Sweden and Swedish-speaking parts of Finland. At most sites, 12 subjects were recorded, 3 each of older men, older women, younger men, and younger women. Subjects were recorded in spontaneous conversational speech and while performing tasks to elicit particular words. The recordings could then be used for a variety of acoustic analyses. Both vocalic (Leinonen 2010) and consonantal (Wretling, Strangling, and Schaeffler 2002) variables have been analyzed acoustically in studies using this corpus. Beyond the English-speaking world and Sweden, the use of acoustic analysis for vocalic variation is only incipient. Adank, van Hout, and van de Velde (2004) compared standard Netherlandic and Flemish Dutch vowels in F_1/F_2 space, finding a more rapid adoption of certain recent shifts in the Netherlands than in Flanders. Willis (2007a), using formant values, found small differences between southwest US Spanish and other dialects of Spanish. Moosmüller and Granser (2006) employed formant analysis to compare different regional dialects of Albanian and document the relationship of the new Albanian standard to the regional forms.

All of the preceding studies have examined vowels with a traditional nucleus/glide outlook, normally taking measurements at only one or two locations within each token. Trajectory approaches are relatively new. An exemplary paper illustrating the merits of exploring vowel dynamics is Jacewicz, Fox, and Salmons (2011), which used five measurement points. They examined cross-generational patterns in three widely separated communities, one each in Wisconsin, Ohio, and North Carolina, comparing how the degree of gliding changed in each location. Among the more notable trends were an increase in the gliding of the FACE vowel in Wisconsin and reduction of gliding of the KIT, DRESS, TRAP, and THOUGHT vowels in North Carolina. They also confirmed several suspected differences in gliding patterns among the communities, such as the different gliding patterns of the TRAP vowel shown in each community.

Related to vowel dynamics is vowel duration. Surprisingly little dialectal work has focused on vowel duration. Jacewicz, Salmons, and Fox (2007) conducted a study with subjects from the same communities described above for Jacewicz *et al.* (2011). By controlling for intrinsic vowel length and prosodic emphasis, they were able to show that regional differences in stressed vowel duration were quite pervasive. In a different vein, Labov and Baranowski (2006) showed that duration had changed from a subsidiary to a primary role in the phonological differentiation of the LOT and DRESS vowels in the Great Lakes region.

Acoustic aspects of vowels can figure into perception experiments on spatial differences in language as well. One type of experiment, investigating the degree of intelligibility of different dialects to listeners, is discussed by Gooskens (this volume). Another common sort of experiment is dialect identification, in which subjects are asked to identify the regional dialect of a voice (e.g., Preston 1993; Bezooijen and Gooskens 1999; Clopper and Pisoni 2004, 2006; Baker, Eddington, and Nay 2009). Results can vary depending on how many choices listeners are given and how similar the dialects being differentiated are. There are also experiments in which the abilities of listeners from different regions to understand dialectal utterances are compared (e.g., Labov and Ash 1997). Yet another kind of experiment examines differences in categorization of sounds by listeners of different dialectal backgrounds (e.g., Willis 1972; Janson 1983; Rakerd and Plichta 2010). A subtype of the latter tests whether listeners can distinguish sounds that are merged in some dialects (e.g., Labov, Karan, and Miller 1991; Rae and Warren 2002). In these experiments, the stimuli may be short excerpts of raw recordings or they can be signals that have been synthesized to represent a continuum of sounds. The stimuli may be treated in various ways as well. Treatments might include gating, in which listeners hear, in succession, increasingly longer sections of a signal; addition of noise to a signal; filtering so that certain aspects of a signal, such as segmental information or a certain vowel formant, are eliminated; or modifying formant, temporal, or pitch characteristics of a signal to disguise associated aspects of a voice.

18.3 Consonantal Studies

Although studies of consonantal variation typically rely on auditory judgments, some studies have introduced acoustic techniques. Docherty and Foulkes (1999) and Foulkes and Docherty (2006) compared the realization of /t/ between sonorants in Newcastle-upon-Tyne, England, and of word-final /t/ in Newcastle-upon-Tyne and Derby. Between vowels, as in *better*, /t/ commonly shows glottalization in vernacular British dialects. Glottalization results in slow, irregular glottal pulses just before and after the [t] occlusion. In Newcastle, the process could progress further so that the occlusion was absent—only a period of slowed vocal pulses between the two vowels remained. Indeed, this form predominated in Newcastle. For word-final position, Docherty and Foulkes termed realizations with a voiceless occlusion and a release “canonical.” Two other variants occurred in Derby and Newcastle, however. One, “continued voicing,” involved vocal pulsing throughout the stop occlusion. The other, “pre-aspiration,” occurred when the vowel became breathy instead of glottal before the occlusion. Continued voicing proved more common in Newcastle than in Derby, and pre-aspiration was increasing in Newcastle, especially among females, but not in Derby. Geographical studies of dialects in Sweden (Tronnier 2002; Wretling *et al.* 2002) have used similar acoustic methods to compare pre-aspiration there. These variants are all difficult to assess by ear but easily differentiated with acoustic equipment.

As the Newcastle and Derby evidence demonstrates, ostensibly “identical” consonants can differ subtly. This issue is magnified when dialects differ in how they contrast particular consonants. Purnell *et al.* (2005a,b) showed that contrasts between voiced and voiceless

consonants in syllable codas were once maintained in a decidedly unorthodox way in German-substrate Wisconsin English. The oldest generations, comprising German/English bilinguals, produced longer vowel durations before voiceless obstruents than before voiced obstruents. This configuration is otherwise unreported across languages. Speakers compensated by exaggerating the durations of the occlusions for voiceless consonants, making the vowel/occlusion duration ratios more like mainstream English patterns. Another study, Jacewicz, Fox, and Lyle (2009), found a geographical difference in voicing by measuring the proportions of /b/ and /d/ occlusions with vocal pulsing.

Another variable consonantal factor is Voice Onset Time, or VOT. VOT is a property of stops involving the degree of voicing and the degree of aspiration. VOT measures the amount of time between the stop burst—the popping sound created when air behind the constriction hits air in front after the release—and the onset of vocal pulsing. For a truly voiced stop, pulsing begins before the burst and VOT will be negative. For a voiceless unaspirated stop, VOT is barely above zero, whereas for a voiceless aspirated stop, VOT falls well above zero. Variation in VOT is common among L2 learners when the source and target languages differ in stop characteristics. For example, English exhibits strong aspiration for voiceless stops and inconsistent vocal pulsing for “voiced” stops, but many languages such as Spanish and French lack aspiration and show more consistent pulsing during voiced stops. L2 learners of English do not always attain native-speaker-like VOT. Some ethnic communities have undergone flux in VOT (Heselwood and McChrystal 1999; MacLagan *et al.* 2009). VOT variation has also appeared in situations without recent language shift. Docherty *et al.* (2011) examined four communities along the English/Scottish border and noted a diachronic trend toward increased aspiration of voiceless stops and less pulsing of voiced stops. Syrdal (1996) reported subtle variations across US dialect regions: for example, speakers from the Western and South Midland dialect regions showed the most pulsing for voiced stops. Takada and Tomimori (2006) found that lack of vocal pulsing for “voiced” stops was spreading from northeastern Japan to the rest of the country.

Although modern phonetic techniques await application to geographic studies of fricative variation, they have been applied to geographical variation of approximants. Many approximants are vowel-like and measurable like vowels. Exemplary of such sounds are laterals. A common variation involves whether a lateral is “light” or “clear,” without velarization, or “dark,” with velarization. F_2 is relatively high for light [l] and lower for dark [ɫ]. Recasens (2004) and Recasens and Espinosa (2005) used F_2 and F_2-F_1 values, as well as electropalatography, to show that Majorcan Catalan had darker laterals than Valencian Catalan. In England, Carter and Local (2007), using F_2 frequencies, showed that Leeds English had darker laterals than Newcastle English. Turton (2014) used ultrasound to compare /l/ realization in three British dialects and American English.

Rhotics, sounds represented as *r*, show considerable variation across and even within languages. Apical and uvular /r/ types, easily distinguished acoustically, compete within continental European languages, as do retroflex and “bunched” (pharyngeal-palatal) approximants in English and Dutch. Ultrasound work on rhotics has emerged recently (e.g., Lawson, Scobbie, and Stuart-Smith 2007; Mielke, Baker, and Archangeli 2010; Sebregts 2014). Acoustic measures are underutilized, though acoustic techniques for differentiating /r/ types exist. Labiovelar approximants in England, for example, are amenable to acoustic analysis (Foulkes & Docherty 2000). Willis (2006) and Bradley and Willis (2012) applied spectrographic analysis to the realizations in Dominican and Veracruz Mexican Spanish of /rr/, as in *perro* “dog,” and /r/, as in *pero* “but.” /rr/ could be realized as a pre-aspirated tap, a devoiced fricative, or a tap preceded or followed by an approximant, not just as a trill, the standard form. Colantoni (2006) examined assibilation of word-initial /r/ in upland and lowland varieties of Argentinian Spanish and compared it with assibilation of /j/, finding that the lowland dialect favored assibilation of /j/ and the upland dialect assibilation of /r/.

18.4 Prosodic Studies

Prosody involves variations in pitch, loudness, and timing that are not directly linked to intrinsic segmental quality or quantity. It has always attracted far less attention from dialectologists than segmental variation. Even today, investigation of spatial variation in prosody has been unsystematic. Existing studies have compared two or a few locales instead of large geographical areas as dialect surveys have. However, the dearth of systematic work constitutes an opening for future research. To date, work on prosodic variation has investigated timing differences and pitch/fundamental frequency (F_0) differences but not loudness/amplitude differences.

Timing factors are studied by measuring durations of segmental or non-segmental intervals in speech. That is, the durations of vowels or vowel sequences, consonants or consonant sequences, and pauses, measured in seconds or milliseconds, may be obtained. Durations are normally measured from spectrograms or waveforms. They are then manipulated in various ways, depending on the kind of analysis. Speech rate involves how quickly a person utters segments, syllables, or words per unit of time (e.g., syllables per second or seconds per syllable). Syllables are used more often than words or segments. There are two basic ways of approaching speech rate: speaking rate is the number linguistic units per unit of time, including pause time, while articulation rate is the same except that pause time is excluded (Robb 2004; Kendall 2013, 27).

Kendall (2013) examined speaking rate, articulation rate, and pause duration in two corpora covering different regions of the United States but based mostly on conversational speech. Some apparent regional differences, such as Northerners having shorter pauses than Southerners and Westerners showing faster speaking and articulation rates than Northerners or Southerners, emerged. The differences interacted with other factors, such as speakers' age and ethnicity. Other researchers have also investigated regional differences in rate of speech in the United States. Ray and Zahn (1990) failed to find any differences among American English dialects. However, Jacewicz *et al.* (2009) found faster articulation rates for Northerners than for Southerners. Southerners have shown longer durations than Northerners for at least some vowels (Clopper *et al.* 2005; Jacewicz *et al.* 2007). Elsewhere, Robb *et al.* (2004) found more rapid speech rates among New Zealanders than among Americans or Australians. Two studies of Dutch, Verhoeven *et al.* (2004) and Quené (2008), reported faster speech rates for natives of the Netherlands than for natives of Flanders. A comparison of the closely related languages Danish, Swedish, and Norwegian largely attributed the apparently faster rates of Danish to syllable elision in Danish (Hilton, Schüppert, and Gooskens 2011).

In contrast to speech rate assessment, measures of prosodic rhythm show the evenness of durations of intervals. They place speech samples on a continuum from syllable-timing, in which syllables have relatively even durations, to stress-timing, in which feet are regarded as having fairly even durations but individual syllables can vary widely. Different methods for assessing prosodic rhythm examine vocalic intervals, consonantal intervals, or both. Of the two most widely used methods, the nPVI (normalized Pairwise Variability Index) method (Low, Grabe, and Nolan 2000), focuses on durations of vocalic intervals, while the method of Ramus, Nespor, and Mehler (1999) uses both vocalic and consonantal intervals. nPVI divides the difference of durations of vowels in adjacent syllables by their mean. Ramus *et al.* utilize three metrics: the standard deviation of vocalic interval durations (ΔV), the standard deviation of consonantal interval durations (ΔC), and the percentage of total utterance time taken up by vowels (%V).

Prosodic rhythm, like speech rate, has garnered a smattering of studies on regional differences. Low *et al.* (2000) and Deterding (2001) compared Singapore English with British English, using different rhythm formulas, and both found that Singapore English is more syllable-timed than British English. Frota and Vigário (2001), using the metrics from Ramus

et al. (1999), presented analyses suggesting that European Portuguese lies between prototypical syllable-timed and stress-timed languages, while Brazilian Portuguese lies beyond the canonical syllable-timed languages. Gazali, Hamdi, and Barkat (2002) also used the Ramus *et al.* metrics to compare Arabic dialects from Morocco to Jordan. They found some differences, mainly because westerly areas showed reduced vowels and complex syllables while eastern dialects showed longer vowels. O'Rourke (2010) combined various methods, revealing rhythmic differences between Quechua-influenced Cuzco Spanish and Lima Spanish, which lacks apparent substrate effects. White and Mattys (2007) similarly used a combination of methods and showed that dialects in various parts of England, Scotland, and Wales differed in their rhythm characteristics.

For tone-based aspects of prosody, including lexical tone and intonation, obtaining accurate F_0 readings is crucial. Autocorrelation pitch tracking is ordinarily used to measure F_0 because it is quick, easy, and efficient to use. It works poorly with creaky or breathy phonation, though. For the former, the irregularity of the vocal pulses thwarts autocorrelation, which relies on the similarity of waveforms in successive pulse periods. For the latter, the pulses become too weak to distinguish from other sound components within the waveform. Field recordings present some challenges as well. Overly soft voices and placement of the microphone too far from a speaker create problems for autocorrelation, but these problems can often be resolved by lowering the voicing threshold for autocorrelation so that it interprets lower-amplitude sound as part of a voice. Autocorrelation is prone to doubling or halving F_0 readings. To combat this problem, researchers should watch for dramatic rises or falls in F_0 and should check the readings against their own ears. The upper and lower bounds of the allowable F_0 range for autocorrelation can be adjusted to counteract doubling and halving. Background noise can adversely affect autocorrelation, but steady noises are usually easier to handle than sudden, sharp noises. Other techniques can be used as backup methods for measuring F_0 (see Thomas 2011).

Acoustic studies of geographical variation in lexical tone are currently rare. Stanford (2008, 2009, 2012), however, investigated tonal variation in Sui, an indigenous language of the Tai-Kadai family in southern China. Sui shows a north/south dialectal division. The language has six tones, and, using F_0 analyses, Stanford showed that three of the tones exhibit significant regional differences. The society is patrilocal and exogamous, and Stanford (2008) demonstrated that married women do not accommodate tonally to the dialect of their husbands' villages.

Intonational analysis today is almost entirely based on the Tone and Break Index (ToBI) system and a few related systems. These systems recognize three kinds of intonational objects: phrases, edge tones, and pitch accents. Different languages exhibit different inventories of each of those objects. Edge tones occur at the end and rarely at the beginning of phrases. Pitch accents occur within phrases. Most languages are considered to have only two basic kinds of tones, high (H) and low (L), although they can occur in combinations, such as a H-L% edge tone or a L+H* pitch accent, thereby multiplying the tonal possibilities. ToBI requires acoustic analysis because pitch accents are used—in conjunction with auditory judgments—to identify tones. Utterances are divided into types such as yes/no questions versus declaratives or narrow-focus (i.e., with a particular word emphasized) versus broad-focus, and pitch accents are divided into nuclear (the last pitch accent in a phrase) versus pre-nuclear. However, once the tones and the type of sentence are determined, other kinds of acoustic analyses can be conducted on the tones. These analyses offer opportunities for dialectal surveys.

Acoustic research on intonational variation is off to a promising start. Thus far, all research has involved comparisons of intonational patterns of just a few locations, usually major cities. However, because intonational systems generally offer fewer variables than vowels or consonants, systematic geographical surveys might be relatively easy to conduct and less expensive than other surveys because interviews could be shorter.

Some papers on dialectal variation in intonation have examined how certain dialects differ in their inventories of tones or the functions of particular tones. Grice *et al.* (2005) determined that four dialects of Italian differ in their inventories of tones in nuclear pitch accents and allocate certain tones to different purposes, depending on the kind of sentence (broad- versus narrow-focus statements, yes/no questions, and continuation). Selting (2004) showed that Dresden German more often utilizes multiple abrupt rises in F_0 within an utterance than Berlin German. Similarly, Willis (e.g., 2004, 2007b, 2008) used quantitative analyses of F_0 to show that Dominican, Mexican, and Peninsular Spanish employ different tones for particular functions.

Beyond papers aimed at phonological differences in tone inventories, however, there are also phonetic approaches to intonational analysis. Spectrograms that allow analysts to find the onsets and offsets of vocalic and consonantal intervals are essential for analyses that gauge the position of F_0 peaks or troughs relative to the vowel that "hosts" the tone. Several studies have examined the position of F_0 peaks in dialectal comparisons. This phenomenon is often called "peak delay" because it describes how much time elapses between a reference point and the F_0 peak. The reference point may be the onset of the host vowel, the beginning of any consonant in the onset of the host syllable, or the beginning of the consonant interval before the host vowel. For these methods, the datum is simply an amount of time. Another method, however, involves computing a proportion: the duration of the host vowel is measured and then the time point of the F_0 peak relative to the vocalic interval is calculated. For example, if the vowel is 100 ms long and the peak falls 30 ms after the vowel offset, the measurement is $(100+30)/100$, or 1.3. Landmarks used for peak delay are illustrated in Figure 18.4. Barnes *et al.* (2012) suggest that the point of highest F_0 is not necessarily what listeners hear as the peak—instead, listeners perceive more broadly distributed regions of

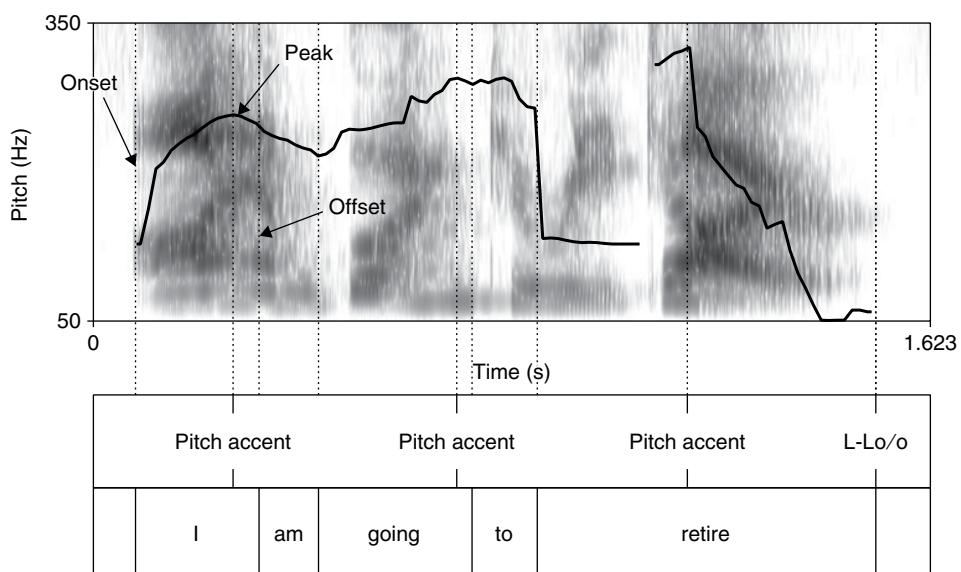


Figure 18.4 Landmarks used for the measurement of peak delay, that is, the position of the peak value of a pitch accent relative to its host syllable. The pitch track, set from 50 to 350 Hz, is superimposed on a wideband spectrogram set from 0 to 5000 Hz. The first pitch accent is hosted by the syllable *I*. The onset and offset of the *I* syllable are marked, as well as the peak of the pitch contour. The onset is at 0.082 second into the sound file, the offset at 0.271 second. The peak is 0.189 second after the onset, that is, 0.790 of the distance from onset to offset.

high or low F_0 —and that more nuanced approaches are preferable. They provide formulas for correcting peak placement to reflect human perception.

Studies on dialectal differences in peak alignment have examined several languages. Atterer and Ladd (2004) and Mücke *et al.* (2009) both compared dialects of German—Bavarian and northwestern urban German for Atterer and Ladd, Viennese and Düsseldorf German for Mücke *et al.* The former study examined only pre-nuclear pitch accents, the latter both nuclear and pre-nuclear pitch accents. In both studies, however, the more southerly accent showed, on average, later peaks than the more northerly accent. Auferbeck (2004) and Ladd *et al.* (2009) each compared a dialect from Scotland with speech from southern England. Though Auferbeck did not specify pitch accent type, Ladd *et al.* examined both nuclear and pre-nuclear pitch accents. In both studies, Scottish English showed later pitch accent peaks than southern England English. Nuclear pitch accents often differ from pre-nuclear pitch accents because, falling just before the edge tone, they may be “crowded” and thus pushed back earlier by the edge tone. Nevertheless, Mücke *et al.* (2009) and Ladd *et al.* (2009) each demonstrated that dialectal differences occurred in both nuclear and pre-nuclear contexts. O’Rourke (2004) compared pre-nuclear peaks in Quechua-substrate Spanish from Cuzco, Peru, with those in the Spanish of Lima. Pitch accents exhibited earlier peaks in Cuzco than in Lima. Language contact situations such as Cuzco can engender intonational variation: see Queen (2012) regarding Turkish/German contact.

A different acoustic analysis of pitch accents was conducted by Grabe *et al.* (2000). They examined the preference of four urban dialects in the British Isles for compression or truncation of pitch contours when the contour’s duration was shortened. For compression, the pitch excursion (the difference between the highest and lowest F_0 values in the contour) remains roughly the same as for a phonologically equivalent contour with a longer duration. For truncation, shortening the duration reduces the degree of pitch excursion. The rate of F_0 change—that is, excursion divided by time—can also be useful. Grabe *et al.* found that the four dialects differed in their tendency toward compression or truncation. Grabe *et al.* used sets of words read by 12 speakers (six of each sex) from each of the four cities that were included. For less tightly controlled samples, pitch excursion values should be normalized by first converting Hertz values to ERB units (Greenwood 1961) and then subtracting.

18.5 Outlook

What is clearest by now is that research on regional dialects has many avenues for exploiting acoustic methods. The only line of research that has taken full advantage of acoustic techniques is analysis of vowel quality, and then only for a few languages. One hardly knows where to begin in naming additional areas for which application of acoustic methods is a priority. Consonantal analysis of various sorts, intonation, applications to more languages: all sorely need insights supplied by acoustic techniques. However, the sparseness of past work results in a wide open field. Moreover, the fresh perspectives that acoustic methods provide can keep dialect geography viable and vigorous far into the future.

Nearly every technique described here continues to be refined. Kendall (2013), for example, has developed several new methods for analyzing speech rate, and Zhou *et al.* (2008) discovered an acoustic method of distinguishing two common /r/ types in English. The use of ultrasound will permit analysis of articulation to accompany acoustic analysis, particularly for consonantal investigations. Yet another ongoing development is the creation of automatic alignment programs such as FAVE (Forced Alignment & Vowel Extraction, <http://fave.ling.upenn.edu/>). These programs align transcripts with recordings using acoustic markers, after which acoustic analysis can be performed on segments. Having a machine find segments instead of requiring an analyst to find them speeds up the measurement process immensely.

Automatic alignment introduces errors that necessitate checking and works poorly with low-quality recordings, but otherwise is remarkably efficient.

Analysis of archival recordings is another direction for future work. Such recordings afford researchers windows into the history of dialects, but scholars sometimes hesitate to apply acoustic analysis to them because of various problems with the recordings: noise present in the original environment, outmoded equipment that introduced more noise and possibly distortion, and deterioration of the recordings. Nevertheless, many older recordings have excellent sound quality, and even when they do not, skilled analysts can still procure a great deal of data from them, as Purnell (2013) and Thomas (2017) show. The road is open, in both future dialect surveys and further analysis of past surveys, for acoustic studies.

REFERENCES

- Adank, Patti, Roeland van Hout, and Hans van de Velde. 2007. An acoustic description of the vowels of northern and southern standard Dutch II: Regional varieties. *Journal of the Acoustical Society of America* 121: 1130–1141.
- Atterer, Michaela, and D. Robert Ladd. 2004. On the phonetics and phonology of “segmental anchoring” of F0: Evidence from German. *Journal of Phonetics* 32: 177–197.
- Auferbeck, Margit. 2004. Identifying sociolinguistic variables in intonation: The onset onglide in Anstruther Scottish English. In Gilles and Peters (eds.), 33–48.
- Baker, Wendy, David Eddington, and Lyndsey Nay. Dialect identification: The effects of region of origin. *American Speech* 84: 48–71.
- Baranowski, Maciej. 2007. *Phonological Variation and Change in the Dialect of Charleston, South Carolina*. Publication of the American Dialect Society 92. Durham: Duke University Press.
- Barnes, Jonathan, Nanette Veilleux, Alejna Brugos, and Stefanie Shattuck-Hufnagel. 2012. Tonal center of gravity: A global approach to tonal implementation in a level-based intonational phonology. *Laboratory Phonology* 3: 337–383.
- Bezooijen, Renée van, and Charlotte Gooskens. 1999. Identification of language varieties: The contribution of different linguistic levels. *Journal of Language and Social Psychology* 18: 31–48.
- Boberg, Charles. 2001. The phonological status of western New England. *American Speech* 76: 3–29.
- Boberg, Charles. 2008. Regional phonetic differentiation in standard Canadian speech. *Journal of English Linguistics* 36: 129–154.
- Bradley, Travis G., and Erik W. Willis. 2012. Rhotic variation and contrast in Veracruz Mexican Spanish. *Estudios de Fonética Experimental* 21: 43–74.
- Carter, Paul, and John Local. 2007. F2 variation in Newcastle and Leeds English liquid systems. *Journal of the International Phonetic Association* 37: 183–199.
- Clopper, Cynthia G. 2009. Computational methods for normalizing acoustic vowel data for talker differences. *Language and Linguistics Compass* 3: 1430–1442.
- Clopper, Cynthia G., David B. Pisoni. 2004. Homebodies and army brats: Some effects of early linguistic experience and residential history on dialect categorization. *Language Variation and Change* 16: 31–48.
- Clopper, Cynthia G., David B. Pisoni. 2006. Effects of region of origin and geographic mobility on perceptual dialect categorization. *Language Variation and Change* 18: 193–221.
- Colantoni, Laura. 2006. Micro and macro sound variation and change in Argentine Spanish. In Nuria Sagarría and Almeida Jacqueline Toribio (eds.), *Selected Proceedings of the 9th Hispanic Linguistics Symposium*, 91–102. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #1369.
- Deterding, David. 2001. The measurement of rhythm: A comparison of Singapore and British English. *Journal of Phonetics* 29: 217–230.
- Dinkin, Aaron J. 2011. Weakening resistance: Progress toward the low back merger in New York State. *Language Variation and Change* 23: 315–345.
- Docherty, Gerard J., and Paul Foulkes. 1999. Derby and Newcastle: Instrumental phonetics and variationist studies. In Paul Foulkes and Gerard J. Docherty (eds.), *Urban Voices: Accent*

- Studies in the British Isles*, 47–71. London: Arnold.
- Docherty Gerard J, Dominic Watt, Carmen Llamas, Damien Hall, and Jennifer Nyocz. 2011. Variation in voice onset time along the Scottish-English border. In *Proceedings of ICPHS XVII, Hong Kong, 17–21 August 2011*, pp. 591–594.
- Eriksson, Anders. 2004. Swedia-Projektet: Dialektforskning i ett jämförande perspektiv. *Folkmålsstudier* 43: 11–31.
- Foulkes Paul, and Gerard J. Docherty. 2000. Another chapter in the story of /r/: “Labiodental” variants in British English. *Journal of Sociolinguistics* 4: 30–59.
- Foulkes, Paul, and Gerard J. Docherty. 2006. The social life of phonetics and phonology. *Journal of Phonetics* 34: 409–38.
- Fridland, Valerie. 2000. The Southern Shift in Memphis, Tennessee. *Language Variation and Change* 11: 267–285.
- Frota, Sónia, and Marina Vigário. 2001. On the correlates of rhythmic distinctions: The European/Brazilian Portuguese case. *Probus* 13: 247–275.
- Ghazali, Salem, Rym Hamdi, and Melissa Barkat. 2002. Speech rhythm variation in Arabic dialects. In *Proceedings of Speech Prosody 2002*, 331–334. Aix-en-Provence: Laboratoire Parole et Langage. <http://aune.lpl.univ-aix.fr/sp2002/>
- Gilles, Peter, and Jörg Peters (eds.). 2004. *Regional Variation in Intonation*. Linguistische Arbeiten 492. Tübingen: Max Niemeyer Verlag.
- Gordon, Matthew J. 2001. *Small-Town Values and Big-City Vowels: A Study of the Northern Cities Shift in Michigan*. Publication of the American Dialect Society 84. Durham: Duke University Press.
- Grabe, Esther, Brechtje Post, Francis Nolan, and Kimberley Farrar. 2000. Pitch accent realization in four varieties of British English. *Journal of Phonetics* 28: 161–185.
- Greenwood, Donald D. 1961. Critical bandwidth and the frequency coordinates of the basilar membrane. *Journal of the Acoustical Society of America* 33: 1344–1356.
- Grice, Martine, Mariapaola D’Imperio, Michelina Savino, and Cinzia Avesani. 2005. Strategies for intonation labelling across varieties of Italian. In Sun-Ah Jun (ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford, UK: Oxford University Press. 362–389.
- Heselwood, Barry, and Louise McChrystal. 1999. The effect of age-group and place of L1 acquisition on the realisation of Panjabi stop consonants in Bradford: An acoustic sociophonetic study. *Leeds Working Papers in Linguistics and Phonetics* 7: 49–69.
- Hilton, Nanna Haug, Anja Schüppert, and Charlotte Gooskens. 2011. Syllable reduction and articulation rates in Danish, Norwegian and Swedish. *Nordic Journal of Linguistics* 34: 215–237.
- Irons, Terry Lynn. 2007. On the status of low back vowels in Kentucky English: More evidence of merger. *Language Variation and Change* 19: 137–180.
- Ito, Rika, and Dennis R. Preston. 1998. Identity, discourse, and language variation. *Journal of Language and Social Psychology* 17: 465–483.
- Jacewicz Ewa, Robert A. Fox, and Samantha Lyle. 2009. Variation in stop consonant voicing in two regional varieties of American English. *Journal of the International Phonetic Association* 39: 313–334.
- Jacewicz, Ewa, Robert A. Fox, Caitlin O’Neill, and Joseph Salmons. 2009. Articulation rate across dialect, age, and gender. *Language Variation and Change* 21: 233–256.
- Jacewicz, Ewa, Robert Allen Fox, and Joseph Salmons. 2011. Cross-generational vowel change in American English. *Language Variation and Change* 23: 45–86.
- Jacewicz, Ewa, Joseph Salmons, and Robert A. Fox. 2007. Vowel duration in three American English dialects. *American Speech* 82: 367–385.
- Janson, Tore. 1983. Sound change in perception and production. *Language* 59: 18–34.
- Johnson, Daniel Ezra. 2010. *Stability and Change along a Dialect Boundary: The Low Vowels of Southeastern New England*. Publication of the American Dialect Society 95. Durham, N.C.: Duke Univ. Press.
- Kendall, Tyler. 2013. *Speech Rate, Pause, and Sociolinguistic Variation*. Basingstoke, U.K./New York: Palgrave Macmillan.
- Labov, William. 1980. The social origins of sound change. In William Labov (ed.), *Locating Language in Time and Space*. New York: Academic. 251–265.
- Labov, William. 2007. Transmission and diffusion. *Language* 83: 344–387.
- Labov, William, and Sharon Ash. 1997. Understanding Birmingham. In Cynthia Bernstein, Thomas Nunnally, and Robin Sabino (eds.), *Language Variety in the South Revisited*, 508–573. Tuscaloosa/London: University of Alabama Press.

- Labov, William, Sharon Ash, and Charles Boberg. 2006. *The Atlas of North American English: Phonetics, Phonology and Sound Change*. Berlin: Mouton de Gruyter.
- Labov, William, and Maciej Baranowski. 2006. 50 msec. *Language Variation and Change* 18: 223–240.
- Labov, William, Mark Karan, and Corey Miller. 1991. Near-mergers and the suspension of phonemic contrast. *Language Variation and Change* 3: 33–74.
- Labov, William, Malcah Yaeger, and Richard Steiner. 1972. *A Quantitative Study of Sound Change in Progress*. Philadelphia: U.S. Regional Survey.
- Ladd, D. Robert, Astrid Schepman, Laurence White, Louise May Quarmby, and Rebekah Stackhouse. 2009. Structural and dialectal effects on pitch peak alignment in two varieties of British English. *Journal of Phonetics* 37: 145–161.
- Lawson, Eleanor, James M. Scobbie, and Jane Stuart-Smith. 2007. The social stratification of tongue shape for postvocalic /r/ in Scottish English. *Journal of Sociolinguistics* 15: 256–268.
- Leinonen, Therese N. 2010. An acoustic analysis of vowel pronunciation in Swedish dialects. Ph.D. dissertation, Rijksuniversiteit Groningen.
- Low, Ee Ling, Esther Grabe, and Francis Nolan. 2000. Quantitative characterizations of speech rhythm: Syllable-timing in Singapore English. *Language and Speech* 43: 377–401.
- MacLagan, Margaret, Catherine I. Watson, Ray Harlow, Jeanette King, and Peter Keegan. 2009. /u/ fronting and /t/ aspiration in Māori and New Zealand English. *Language Variation and Change* 21: 175–192.
- Mielke, Jeff, Adam Baker, and Diana Archangeli. 2010. Variability and homogeneity in American English /ɹ/ allophony and /s/ retraction. In Cécile Fougeron, Barbara Kühnert, Mariapaola D’Imperio, and Nathalie Vallée, eds., *Variation, Detail, and Representation (LabPhon 10)*, 699–719. Berlin: Mouton de Gruyter.
- Moosmüller, Sylvia, and Theodor Granser. 2006. The spread of standard Albanian: An illustration based on an analysis of vowels. *Language Variation and Change* 18: 121–140.
- Mücke, Doris, Martine Grice, Johannes Becker, and Anne Hermes. 2009. Sources of variation in tonal alignment: Evidence from acoustic and kinematic data. *Journal of Phonetics* 37: 321–338.
- O'Rourke, Erin. 2004. Peak placement in two regional varieties of Peruvian Spanish intonation. In Julie Auger, J. Clancy Clements, and Barbara Vance (eds.), *Contemporary Approaches to Romance Linguistics: Selected Papers from the 33rd Linguistic Symposium on Romance Languages (LSRL), Bloomington, Indiana, April 2003*, 321–341. Amsterdam studies in the theory and history of linguistic science, Series IV: Current issues in linguistic theory 238. Amsterdam/Philadelphia: John Benjamins.
- O'Rourke, Erin. 2010. Speech rhythm variation in dialects of Spanish: Applying the Pairwise Variability Index and Variation Coefficients to Peruvian Spanish. In *Proceedings of Speech Prosody 2008: Fourth Conference on Speech Prosody, Campinas, Brazil, May 6–9, 2008*, 431–434. <http://aune.lpl.univ-aix.fr/~sprosig/sp2008/papers/id173.pdf>
- Preston, Dennis R. 1993. Folk dialectology. In Dennis R. Preston (ed.), *American Dialect Research*, 333–377. Amsterdam/Philadelphia: John Benjamins.
- Purnell, Thomas. 2013. Hearing the American language change: The state of DARE recordings. *American Speech* 88: 275–301.
- Purnell, Thomas, Joseph Salmons, and Dilara Tepeli. 2005a. German substrate effects in Wisconsin English: Evidence for final fortition. *American Speech* 80: 135–164.
- Purnell, Thomas, Joseph Salmons, Dilara Tepeli, and Jennifer Mercer. 2005b. Structured heterogeneity and change in laryngeal phonetics: Upper Midwestern final obstruents. *Journal of English Linguistics* 33: 307–338.
- Queen, Robin. 2012. Turkish-German bilinguals and their intonation: Triangulating evidence about contact-induced language change. *Language* 88: 791–816.
- Quené, Hugo. 2008. Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. *Journal of the Acoustical Society of America* 123: 1104–13.
- Rae, Megan, and Paul Warren. 2002. Goldilocks and the three beers: Sound merger and word recognition in NZE. *New Zealand English Journal* 16: 33–41.
- Rakerd, Brad, and Bartłomiej Plichta. 2010. More on Michigan listeners' perceptions of /a/-fronting. *American Speech* 85: 431–449.
- Ramus, Franck, Marina Nespor, and Jacques Mehler. 1999. Correlates of linguistic rhythm in the speech signal. *Cognition* 73: 265–292.
- Ray, George B., and Christopher J. Zahn. 1990. Regional speech rates in the United States: A preliminary analysis. *Communication Speech Reports* 7: 34–37.

- Recasens, Daniel. 2004. Darkness in /l/ as a scalar phonetic property: Implications for phonology and articulatory control. *Clinical Linguistics & Phonetics* 18: 593–603.
- Recasens, Daniel, and Aina Espinosa. 2005. Articulatory, positional and coarticulatory characteristics for clear /l/ and dark /l/: Evidence from two Catalan dialects. *Journal of the International Phonetic Association* 35: 1–25.
- Robb, Michael P., Margaret A. MacLagan, and Yang Chen. 2004. Speaking rates of American and New Zealand varieties of English. *Clinical Linguistics & Phonetics* 18: 1–15.
- Sebregts, Koen. 2014. *The Sociophonetics and Phonology of Dutch r*. Utrecht: Landelijke Onderzoekschool Taalwetenschap.
- Selting, Margret. 2004. Regionalized intonation in its conversational context. In Gilles and Peters (eds.), 49–73.
- Stanford, James N. 2008. A sociotonetic analysis of Sui dialect contact. *Language Variation and Change* 20: 409–450.
- Stanford, James N. 2009. “Eating the food of our place:” Socio-linguistic loyalties in multi-dialectal Sui villages. *Language in Society* 38: 287–309.
- Stanford, James N. 2012. One size fits all? Dialectometry in a small clan-based indigenous society. *Language Variation and Change* 24: 247–278.
- Syrdal, Ann K. 1996. Acoustic variability in spontaneous conversational speech of American English talkers. In *ICSLP 96 (Fourth International Conference on Spoken Language Processing)*, Philadelphia, Oct. 3–6. <http://www.asel.udel.edu/icslp/cdrom/vol1/582/a582.pdf>
- Takada, Mieko, and Nobuo Tomimori. 2006. The relationship between VOT in initial voiced plosives and the phenomenon of word-medial plosives in Nigata and Shikoku. In Yuji Kawaguchi, Susumu Zaima, and Toshihiro Takagaki (eds.), *Spoken Language Corpus and Linguistic Informatics*, 365–379. Usage-based linguistic informatics 5. Amsterdam/Philadelphia: John Benjamins.
- Thomas, Erik R. 1997. A Rural/Metropolitan Split in the Speech of Texas Anglos. *Language Variation and Change* 9: 309–332.
- Thomas, Erik R. 2001. *An Acoustic Analysis of Vowel Variation in New World English*. Publication of the American Dialect Society 85. Durham, NC: Duke University Press.
- Thomas, Erik R. 2006. Evidence from Ohio on the Evolution of /æ/. In *Language Variation and Change in the American Midland: A New Look at “Heartland” English*, ed. Thomas E. Murray and Beth Lee Simon. Amsterdam/Philadelphia: John Benjamins, 69–89.
- Thomas, Erik R. 2010. A longitudinal analysis of the durability of the Northern-Midland dialect boundary in Ohio. *American Speech* 85: 375–430.
- Thomas, Erik R. 2011. *Sociophonetics: An Introduction*. Basingstoke, U.K./New York: Palgrave Macmillan.
- Thomas, Erik R. 2017. Analysis of the ex-slave recordings. In Raymond Hickey, ed., *Listening to the Past: Audio Records of Accents of English*. Cambridge University Press. 350–374.
- Tronnier, Mechtild. 2002. Preaspiration in southern Swedish dialects. *Proceedings of Fonetik*, TMH-QPSR 44: 33–36. <http://www.speech.kth.se/qpsr>
- Turton, Danielle. 2014. Some /l/ s are darker than others: Accounting for variation in English /l/ with ultrasound tongue imaging. *University of Pennsylvania Working Papers in Linguistics* 20: 189–198.
- Verhoeven, Jo, Guy De Pauw, and Hanne Kloots. 2004. Speech rate in a pluricentric language: A comparison between Dutch in Belgium and the Netherlands. *Language and Speech* 47: 297–308.
- White, Laurence, and Sven L. Mattys. 2007. Rhythmic typology and variation in first and second languages. In Pilar Prieto, Joan Mascaró, and Maria-Josep Solé (eds.), *Segmental and Prosodic Issues in Romance Phonology*, 237–257. Current issues in linguistic theory series. Amsterdam/Philadelphia: John Benjamins.
- Willis, Clodius. 1972. Perception of vowel phonemes in Fort Erie, Ontario, Canada, and Buffalo, New York: An application of synthetic vowel categorization tests to dialectology. *Journal of Speech and Hearing Research* 15: 246–255.
- Willis, Erik W. 2004. Dominican Spanish absolute interrogatives in broad focus. In Timothy L. Face, ed., *Laboratory Approaches to Spanish Phonology*, 61–91. Berlin: Mouton de Gruyter.
- Willis, Erik W. 2006. Trill variation in Dominican Spanish: An acoustic examination and comparative analysis. In Nuria Sagarra and Almeida Jaqueline Toribio, eds., *Selected Proceedings of the 9th Hispanic Linguistics Symposium*, 121–131. Somerville: Cascadilla Proceedings Project.
- Willis, Erik W. 2007a. An initial examination of Southwest Spanish vowels. *Southwest Journal of Linguistics* 24: 185–1198.

- Willis, Erik W. 2007b. Utterance signaling and tonal levels in Dominican Spanish declaratives and interrogatives. *Journal of Portuguese Linguistics* 5/6: 179–202.
- Willis, Erik W. 2008. Tonal characteristics of pronominal interrogatives in Puebla Mexico Spanish. In Esther Herrera Z. and Pedro Martín Butragueño, eds., *Fonología Instrumental: Patrones Fónicos y Variación*, 357–376. Mexico City: El Colegio de México.
- Wretling, Pär, E. Strangert, and F. Schaeffler. 2002. Quantity and preaspiration in northern Sweden. In *Proceedings of speech prosody 2002, Aix-en-Provence, France* 11–13 April 2002, not paginated. www.ling.gu.se/~anders/SWEDIA/papers/wretling_sp2002.pdf
- Zhou, Xinhui, Carol Y. Espy-Wilson, Suzanne Boyce, Mark Tiede, Christy Holland, and Ann Choe. 2008. A magnetic resonance imaging-based articulatory and acoustic study of “retroflex” and “bunched” American English /r/. *Journal of the Acoustical Society of America* 123: 4466–4481.

19 Computational Dialectology

WILBERT HEERINGA AND JELENA PROKIĆ

19.1 Introduction

In Chapter 7, Hans Goebel introduced *dialectometry*. In this chapter we give an overview of the main computational methods that can be used to operationalize dialectometric analyses. We distinguish between categorical comparisons (Section 19.2), frequency-based methods (Section 19.3), and string edit distance (Section 19.4). In Section 19.5 we present some approaches for validating dialectometric methods, and in Section 19.6 three freely available dialectometric software packages are presented. In Section 19.7, we make suggestions for future work.

19.2 Categorical Comparisons

19.2.1 *Relative Identity Value*

Séguy (1973) and his research team measured the linguistic distance between two local neighboring dialects as the number of items on which they disagreed. Goebel (1982) measured similarities instead of differences, and referred to this as *relative identity value* (see also Goebel (2010a), 439–440).

19.2.2 *Weighted Identity Value*

The *relative identity value* (RIV) is usually applied to the measurement of similarities between a pair of sites. Goebel (1982, 1984) also introduced *weighted identity value* (WIV)

to underscore rare features rather than those linguistic types (taxates) which are found elsewhere. This view corresponds with many competent linguistics who think that rare and therefore “more important” language features should be privileged over frequent ones, as they might be considered “trivial” (Goebel 2010b).

Assume 100 dialects are compared to each other by just considering a lexical variable that has lexeme A for 90 dialects and lexeme B for 10 dialects. Using RIV the similarity of a dialect pair is 100% if both dialects have lexeme A or when both dialects have lexeme B. When one dialect has lexeme A and another has lexeme B, the similarity is 0%. When using WIV the similarity will be equal to $1 - ((90-1)/100) = 11\%$ when both dialects have lexeme A and

$1 - ((10-1)/100) = 91\%$ when both dialects have lexeme *B*. WIV varies between 0% and 100%. For a detailed description of the *weighted identity value* (in German *Gewichteter Identitätswert*) see Goebel (1984: I, 83–86).

Similarly, a weighted difference weighting can be measured, as has been done by Nerbonne and Kleiweg (2007), for lexical distances between American English varieties and by Spruit (2008) and Spruit *et al.* (2009) for lexical and syntactic distances between Dutch dialects. They calculated precisely the inverse of WIV: 1-WIV. In the approach of Spruit (2008) and Spruit *et al.* (2009) lexeme *A* versus lexeme *A* would give $(10/100) = 90\%$, and lexeme *B* versus lexeme *B* would give $(90/100) = 10\%$.

Nerbonne and Kleiweg (2007) compared various weighting schemes using *local incoherence* (see Section 19.5.4 below) and concluded that “Goebel’s disproportionate weighting of the overlap of similar items is more sensitive in uncovering the linguistic affinities between sites.”

19.3 Frequency-Based Methods

The approaches presented in the previous section are atlas based and draw on categorical data. They are sometimes criticized as mediated and reductionist (cf. Wälchli 2009) with respect to input data. The methods presented in this section do not rely on atlas data, but on frequency information derived from a careful analysis of language use in authentic, naturalistic texts.

19.3.1 Phone and Phonetic Feature Frequency Methods

Hoppenbrouwers and Hoppenbrouwers developed the *phone frequency method* (PFM) and the *feature frequency method* (FFM) to measure dialect distances on the basis of pronunciation. The methods were introduced in 1988 and also described by Hoppenbrouwers and Hoppenbrouwers (2001).

19.3.1.1 Phone Frequency Method

Hoppenbrouwers and Hoppenbrouwers (2001, 1) suggested comparing varieties based on samples of speech in phonetic transcription. They counted the frequencies of phones (segments) in each sample. Since samples differ in size, relative frequencies were used. The distance between varieties is the sum of the absolute values of the differences between their (relative) phone frequencies.

19.3.1.2 Feature Frequency Method

To accommodate the insight that phonetic similarity is gradual, Hoppenbrouwers and Hoppenbrouwers (1988) also developed a *feature frequency method*. Hoppenbrouwers and Hoppenbrouwers then counted the number of sounds in the sample as [ADVANCEMENT FRONT], [HEIGHT LOW], and so on, to obtain feature frequencies. From a speech sample, a histogram of the relative frequencies of different feature values was obtained (see Figure 19.1). Hoppenbrouwers and Hoppenbrouwers (1988, 2001) calculated histogram similarity using the Pearson’s correlation coefficient (a measure of how well two variable properties agree, e.g., the height and weight of adults).

19.3.1.3 Applications

Hoppenbrouwers and Hoppenbrouwers (2001), being the most extensive application of the *feature frequency method*, considers 156 local Dutch dialects. See Heeringa, Nerbonne, and Osenova (2010) for an application of Hoppenbrouwers’ techniques in contact linguistics that

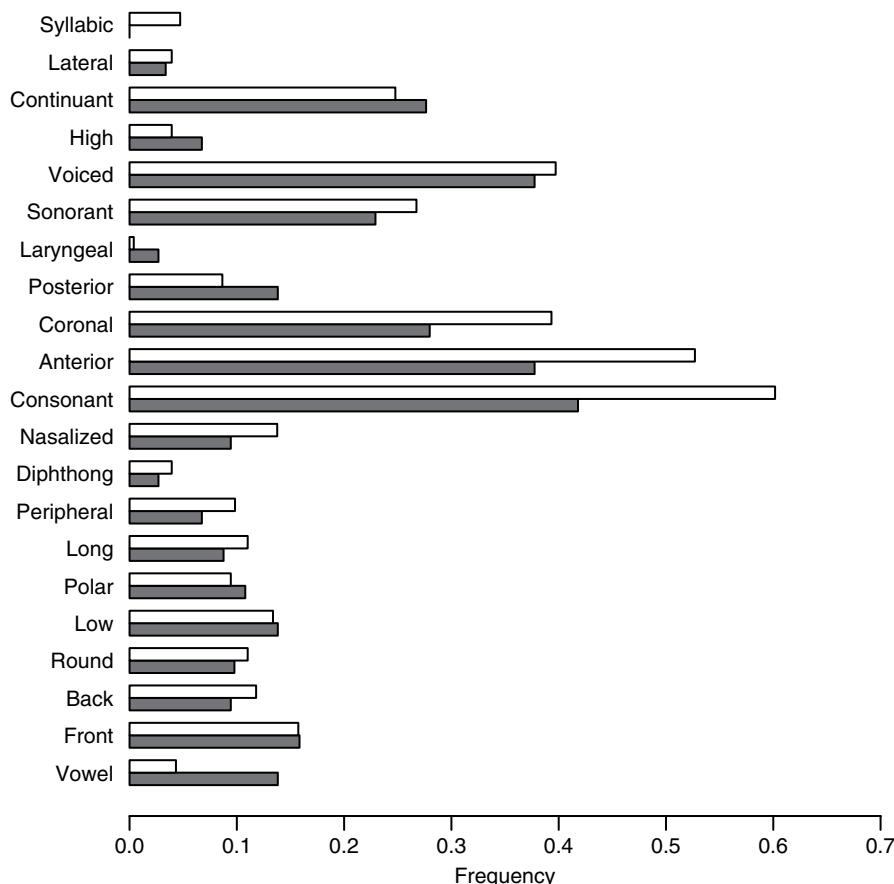


Figure 19.1 Relative feature frequencies measured for 21 features by Hoppenbrouwers and Hoppenbrouwers (2001). Gray bars represent frequencies for the northern Dutch dialect of Wagenborgen, and black bars represent frequencies for the southern Dutch dialect of Kerkrade. Frequencies are divided by the total number of segments in a dialect text. The correlation between the frequencies of the two dialect locations is $r=0.95$. Picture after Hoppenbrouwers and Hoppenbrouwers (2001, 9–10).

exploit the advantage of these techniques that they do not require the same words to be sampled in different sites.

19.3.2 Syntactical Feature Frequency Method

Szembricsanyi (2008) investigated variation in British English dialects by using the Freiburg English Dialect Corpus (FRED), a naturalistic speech corpus sampling interview material from 162 different locations in 38 different counties all over the British Isles, excluding Ireland. The corpus was analyzed to obtain text frequencies of 62 morphosyntactic features, yielding a structured database that provides a 62-dimensional frequency vector per locality. The feature frequencies were subsequently normalized to frequency per 10,000 words (because textual coverage in FRED varies across localities) and log-transformed to de-emphasize large frequency differences and to alleviate the effect of frequency outliers. The resulting 38×62 table (on the county level—i.e., 38 counties characterized by 62 feature frequencies

each for the full dataset) was converted into a 38×38 distance matrix using Euclidean distance—the square root of the sum of all squared frequency differentials—as an interval measure. This distance matrix was subsequently analyzed dialectometrically. See also Szmrecsanyi (2010), Szmrecsanyi and Wolk (2011), and Szmrecsanyi (2013).

19.3.3 Frequencies of Trigrams of Part-of-Speech Tags

Nerbonne and Wiersma (2006), Lauttamus, Nerbonne, and Wiersma (2007), Wiersma, Nerbonne, and Lauttamus (2010), and Nerbonne *et al.* (2010b) measured the total impact of L1 on L2 syntax in second language acquisition (SLA) on the basis of corpora of English of Finnish Australians. They presented an application of a technique from language technology to tag a corpus automatically and to detect syntactic differences between two varieties of Finnish Australian English, one spoken by the first generation and the other by the second generation. The idea of the technique is to utilize frequency profiles of trigrams of POS categories as indicators of syntactic distance between the groups and then examine potential effects of language contact and language (“vernacular”) universals in SLA. The frequency vectors are compared and analyzed by using a permutation test, which results in both a general measure of difference and a list with the n -grams that are most responsible for the difference. The findings showed syntactic “contamination” from Finnish in the English of the adult first-generation speakers of Finnish ethnic origin. The results demonstrated that we can attribute some interlanguage features in the first generation to Finnish substratum transfer.

Sanders (2007) extended the method by using leaf-path ancestors presented in Sampson (2000) instead of trigrams. Di Buccio, Di Nunzio, and Silvello (2014) introduced a variant of this syntactic measure in which frequencies of POS trigrams are gathered into vectors and compared using cosine.

19.4 String Edit Distance

19.4.1 Levenshtein

In 1995, Kessler introduced the use of the Levenshtein distance, also known as string edit distance, as a tool for measuring dialect distances (Kessler 1995). The Levenshtein distance—named after Vladimir Levenshtein who presented this distance in 1965 in a Russian paper and in 1966 in an English paper (Levenshtein 1966)—is a numerical value of the cost of the least expensive set of insertions, deletions or substitutions that would be needed to transform one string into another (Kruskal 1999). Assuming *hart* is pronounced as [hart] in one dialect and as [ærtə] in another, the algorithm will align the realizations as follows:

	1	2	3	4	5
h	a	r	t		
	æ	r	t	ə	
	1	1			1

We find three edit operations: the [h] is deleted, the [a] is replaced by [æ], and [ə] is inserted. In this example, each of these operations has a cost of 1, so the total distance is $1+1+1=3$.

Kessler applied Levenshtein distance successfully to data from *The Linguistic Atlas and Survey of Irish Dialects*. Kessler used 95 varieties and selected 51 concepts for each. Using the Levenshtein distance, he calculated the distances between the dialects. On the basis of these distances, cluster analysis was performed. The resulting dialect areas were continuous, aligned with traditional provincial boundaries and agreed with commonly accepted taxonomies.

Levenshtein distance has several advantages compared to categorical measures (Séguy, Goebl), especially when our data consists of transcriptions of realizations of words. Nerbonne *et al.* (2010a) writes that the use of Levenshtein distance (or: edit-distance)

"provides a broader view of the variation in typical atlas data because it incorporates entire pronunciations in its measurements instead of relying on the analyst's choice of variables to extract. This means that dialectometrists using edit-distance measures of pronunciation are less likely to fall prey to choosing their variables in a way that biases results."

and:

"The fact that genuine measurements are made instead of predication of identity vs. non-identity has a consequence that less data is required for reliable assessment of the relations among sites in a dialect landscape."

19.4.2 Variants

19.4.2.1 Phone String Comparison Versus Feature String Comparison

The simplest technique is *phone string comparison*. In this approach all operations bear the same "cost" of, for instance, 1. In our example, the substitution of the [a] by [æ] has a cost of 1, and the substitution of the [a] by [y] would cost the same, since the phonetic affinity between phones is ignored. A more sensitive technique might use gradual distances between segments as operation weights, perhaps calculating the weights on the basis of segmental features. When, for example, vowels are described by three features (height, backness, and roundness), distances between vowels can be calculated in three-dimensional space and then used as operation weights in the Levenshtein distance, so that the substitution of [a] by [æ] will cost less than the substitution of [a] by [y]. Kessler called this metric *feature string comparison*. Heeringa (2004) measured segment distances on the basis of several feature systems, including the system implied by the IPA table organizing sounds by place and manner of articulation and voicing (see also Heeringa and Braun 2003, who provide details). He also measured distances between segments on the basis of their spectrograms. Since a spectrogram is the curve plotting intensity against time and frequency, the curve distances between the spectrograms reflect acoustic differences. The samples that the spectrograms are based on were pronounced by John Wells and Jill House (Wells and House 1995). Heeringa (2004) focused especially on the more perceptually oriented models, which emphasize the differences that a speaker can hear in someone else's speech.

The choice of operation weights depends on one's research goal. If the goal is to approximate how differences are perceived by dialect speakers, then the use of binary costs (0/1) outperforms that of gradual costs (Heeringa 2004), which suggests that the fact that segments differ is more important than the degree to which they differ, perhaps reflecting the categorical basis of speech distinction. All segmental weighting schemes to date have led to only small differences in measurements at the aggregate varietal level (Heeringa 2004), giving only slightly different correlations to distances as perceived by the dialect speakers themselves in a perception experiment.

19.4.2.2 Free Alignment Versus Forced Alignment

To deal with syllabicity, the Levenshtein algorithm may be adapted so that only vowels may match with vowels, and consonants with consonants, with several exceptions: [j] and [w] may match with both consonants and vowels, [i] and [u] with both vowels and consonants, and central vowels (which may boil down to schwa) with both vowels and sonorant consonants. So the [i], [u], [j], and [w] align with anything, central vowels with syllabic (sonorant) consonants, but otherwise vowels align with vowels and consonants with consonants. In this way unlikely matches (e.g., a [p] with an [a]) are prevented. This approach was first applied to Sardinian dialects (Bolognesi and Heeringa 2002), and then to Dutch (Heeringa and Braun 2003). In a validation study, Heeringa *et al.* (2006) found that forced alignments perform better than free alignments, that is, they approach dialect perception more closely (see Section 19.5.3).

19.4.2.3 Relative Distances Versus Absolute Distances

Nerbonne *et al.* (1996) normalized Levenshtein distances by dividing the absolute distance by the length of the longer word, calling this *relative edit distance*. The idea behind this is to emphasize the perception of words as crucial linguistic units. Sometimes the distance is also divided by the sum of the two variants. The different ways of normalizing are discussed by Heeringa (2004), and he advocates normalizing by the length of the alignment. When several alignments give the same absolute distance, normalization is done by dividing this distance by the length of the longest alignment, since the longest alignment has the greatest number of matches. In the example with the two variants of the word for "heart," we found a Levenshtein distance of 3. The alignment length is 5, so the normalized distance is $3/5=0.6$ or 60%.

The choice between relative distances and absolute distances may depend on one's scientific goal. Heeringa *et al.* (2006) showed that absolute Levenshtein distances approximate dialect differences as perceived by the dialect speakers better than results based on relative Levenshtein distances, that is, normalized by alignment length. This suggests that the weight of the substitution of, for example, the [u] in a word realization of dialect A by the [y] in the realization of the same word in dialect B is independent of the length of the realization in the perception of the speakers. On the other hand, Beijering, Gooskens, and Heeringa (2008) found that intelligibility between languages correlates better with relative distances than with absolute distances. In both cases the differences were not large.

19.4.2.4 All-Word Comparison Versus Same-Word Comparison

Kessler calculated edit distances not only for words that are phonetic variants of each other, the so-called *same-word* approach, but also for lexical variants, calling this the *all-word* approach. Nerbonne and Kleiweg (2003, 2007) applied this approach to the LAMSAS data (see Section 19.2.1) and refer to this as measuring *related lexical items*.

Both approaches have been applied to the same set of 360 Dutch dialects (see also Section 19.3.2). The *all-word* approach is found in Heeringa (2004), and the *same-word* approach is found in Heeringa and Nerbonne (2006). Although the correlation between both measures was not published, it was in fact extremely high ($r=0.99$).

19.4.2.5 The Use of n -Gram Weights

The Levenshtein distance as presented in the previous sections is based on unigrams: the operation weights are calculated on the basis of the comparison of individual speech segments, where neighboring segments are ignored.

Kondrak (2005) proposed to use n -grams (bigrams and trigrams) rather than unigrams. Inkpen, Frunza, and Kondrak (2005) also considered “xbigrams,” which are trigrams without the middle element. Returning to our example above, the use of trigrams would result in the following alignment:

1	2	3	4	5	6	7
--h	-ha	har	art	rt-	t--	
	--æ	-ær	ært	rtə	tə-	ə--
1	1	1	1	1	1	1

Using binary weights, we do not find any pair of trigrams to be equal. The distance is equal to $7*1=7$. Kondrak proposed two more refined measures: comprehensive n -gram similarity, which is to compute the unigram similarity between n -grams, and positional n -gram similarity, which is to simply count identical unigrams in corresponding positions within the n -grams. Note that Kondrak measures similarity rather than distance. Using the latter measure as a distance measure we find:

1	2	3	4	5	6	7
--h	-ha	har	art	rt-	t--	
	--æ	-ær	ært	rtə	tə-	ə--
1	1	0.67	0.33	0.33	0.67	1

$1+1+0.67+0.33+0.33+0.67+1=5$, which is larger than the distance obtained on the basis of unigrams: 3.

Kondrak evaluates the measures on three different word-comparison tasks: the identification of genetic cognates, translational cognates, and confusable drug names. The results of his experiments suggest that the n -gram measures outperform their unigram equivalents.

19.4.3 PMI Levenshtein

Next to phone string and feature string methods for segment comparison (Section 19.4.2), another, data-driven, solution was proposed by Wieling, Prokić, and Nerbonne (2009). They applied pointwise mutual information (PMI) in order to automatically acquire segment distances from the phonetic transcriptions and obtain more precise alignments. PMI is an association measure that estimates the amount of information one event tells us about the other. Applied on the phonetic transcriptions of words, PMI can help us estimate how strongly associated two phones are; the more often they are aligned, the stronger the association between them. Given two phones in the aligned transcription, pointwise mutual information I is calculated as:

$$I(x, y) = \log_2 \frac{P(x, y)}{P(x) \times P(y)}$$

The numerator $P(x, y)$ tells us how often two phones are observed aligned in an entire set of aligned transcriptions, whereas denominator $P(x) \times P(y)$ tells us how often we can expect

these two phones to be aligned together by chance. The procedure of calculating the association strength between the phones and improving the alignments at the same time is iterative and consists of the following steps:

1. Align all word transcriptions using standard Levenshtein algorithm, where distances between the phones are not given.
2. From the obtained alignments, for all pairs of phone segments calculate PMI values using the above formula.
3. Transform PMI values into distances.
4. For all pairs of phones that never align in the data, set the distance to an arbitrary large value.
5. Align all word transcriptions once more using Levenshtein algorithm, but based on the phone distances generated in the previous steps.
6. Repeat steps 2-5 until there are no changes in phone distances and alignments.

The final result of this procedure are distances between each two phone segments in the data as well as the alignments that show improvement when compared to the alignments obtained using the standard Levenshtein algorithm. Wieling, Prokić, and Nerbonne (2009) tested Levenshtein PMI on the Bulgarian dialect data and the number of correct alignments is 1.02% higher than when using slightly modified version of the standard Levenshtein algorithm where vowels are allowed to align only with vowels and consonants only with consonants. Prokić (2010) showed that at the aggregate level distances between language varieties calculated using two versions of the Levenshtein distance correlate highly—98%, but that automatically induced distances between the phones provided much better model of regular sound correspondences than the binary model. Wieling, Margaretha, and Nerbonne (2012) used PMI Levenshtein on the most frequent phones from six dialect data sets and found that PMI induced segment distances correlate reasonably well with acoustic distances in formant space ($.61 < r < .76$). PMI has also been used in computational historical linguistics by Jäger (2013) to infer phylogenies from 5644 word lists taken from the Automated Similarity Judgment Project database (Wichmann *et al.* 2012). The results have shown that the PMI weighted alignment improves the accuracy of the phylogenetic inference in comparison to plain Levenshtein-based alignments 1 to 3% based on the evaluation method used.

19.4.4 Three- and Five-Dimensional Levenshtein

The Levenshtein distance in its original form is *two*-dimensional. The algorithm compares *two* strings with each other and finds the least costly set of operations that map the one string onto the other. In a study of Heeringa and Hinskens (2015) three- and five-dimensional implementations of the algorithm are used. The authors recorded older male speakers and younger female speakers of 86 local dialects of Dutch. Using these data they analyzed and visualized the influence of standard Dutch on apparent time changes in these dialects. Focusing for the most part on variation in the sound components, they tested (I) whether dialect change is mainly the result of convergence to standard Dutch; (II) whether sound changes in two dialects that make them converge to standard Dutch make them also become more similar; and (III) whether sound changes in two dialects, which make them diverge from standard Dutch, make them also become less similar.

In order to test the first hypothesis a three-dimensional Levenshtein distance was used. Per word under consideration three realizations needed to be aligned to each other: older male versus younger female versus standard Dutch. These three-dimensional alignments enabled the authors to distinguish between sound changes that cause a dialect to

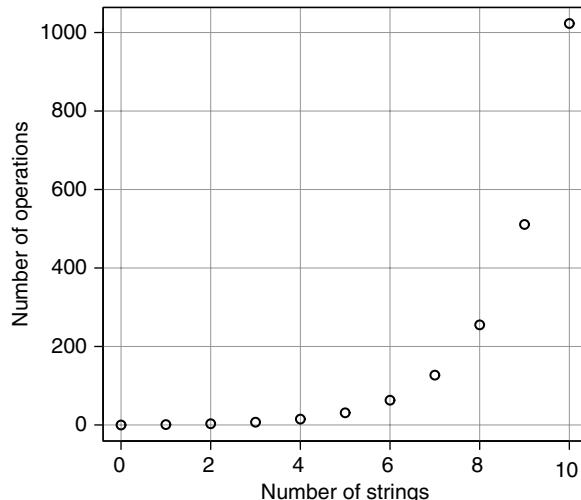


Figure 19.2 The number of operations in a multidimensional Levenshtein distance increases exponentially with the number of dimensions ($2^n - 1$), that is, the number of strings that is simultaneously considered by the algorithm. Picture taken from Heeringa and Hinskens (2015).

convergence to standard Dutch and sound changes that cause a dialect to diverge from standard Dutch.

In order to test the second and third hypothesis, a five-dimensional Levenshtein distance was used since five realizations needed to be aligned to each other. When considering dialect pair *A/B*, per word under consideration the realizations of the older male speaker of dialect *A*, the younger female speaker of dialect *A*, the older male speaker of dialect *B*, the younger female speaker of dialect *B*, and standard Dutch needed to be aligned. This enabled the authors to determine whether sound changes that make dialects converging to standard Dutch also cause them converging to each other, and whether sound changes that make dialects diverging from standard Dutch also cause them to diverge from each other. The findings corroborated all three hypotheses.

When considering multidimensional Levenshtein distance we find that the number of operations increases exponentially, as appears from Figure 19.2. Four-dimensional Levenshtein distance has 15 operations and five-dimensional Levenshtein distance has 31 operations. It would be difficult to implement a 10-dimensional Levenshtein algorithm; therefore, we need other techniques that are presented in Section 19.4.5.

19.4.5 Multiple Sequence Comparison

Unlike pairwise string alignment algorithms, where only two strings are aligned and compared at the time, like the Levenshtein algorithm, multiple string alignment algorithms align and compare more strings at the same time. In recent years, several algorithms for automatic multiple sequence alignment were developed to deal with the linguistic data (Alonso *et al.* 2004; Bhargava and Kondrak 2009; Steiner, Stadler, and Cysouw 2011; List and Moran 2013). Automatic multiple string alignment approach was for the first time applied in dialectology by Prokić, Wieling, and Nerbonne (2009), who relied on the ALPHAMALIG algorithm (Alonso *et al.* 2004) to align and analyze Bulgarian dialect data. Below we

illustrate the results of automatically aligning six pronunciations of word *a3* "I" (example taken from Prokić 2010):

j	a	-	-	-	-
-	a	s	-	-	-
j	a	z	e	-	-
j	ɛ	-	-	-	-
j	a	z	e	k	a
-	d	s	-	-	-

Simultaneous comparison of several strings allows the researcher to easily detect sound correspondences, like [a], [ɛ], and [ɒ] in the second column in the above example. At the same time, alignments and distances between each two strings can be calculated more precisely than in the pairwise comparison approach, since multi-aligned strings preserve the information on the word positions where the loss of a sound occurred in both strings.

In Prokić, Wieling, and Nerbonne (2009) evaluation of the quality of the automatically multi-aligned strings was done by comparing it to the manually corrected multi-aligned alignments. The comparison of the two sets of alignments has shown that the automatically produced alignments correspond 93% to the manually corrected alignments. However, this kind of direct evaluation of the phonetic alignment algorithms, both for pairwise and multiple string alignment, is very rarely performed due to lack of publicly available benchmarks of manually edited alignments. Benchmark Database of Phonetic Alignments in Historical Linguistics and Dialectology (BDPA)¹ (List and Prokić 2014) is the first benchmark database of phonetic alignments in historical linguistics and dialectology that can be used to directly evaluate the performance of alignment algorithms. It contains manually edited multiple alignments taken from 12 different sources, including 8 data sets that contain phonetic dialect data.

Research based on the multiple string alignments in dialectology include work by Prokić and Nerbonne (2013) and Prokić and Cysouw (2013). In order to find concrete phonetic features that make one dialect group distinct from other dialect areas, Prokić and Nerbonne (2013) applied a method a bit like Fisher's linear discriminant (Prokić, Çöltekin, and Nerbonne 2012) to the multi-aligned Bulgarian dialect data. The method seeks the features that differ very little within the group in question and a great deal outside that group. By proceeding from the multi-aligned data they assured that every position within a word is treated as a separate feature. As a result, for each dialect group the most important determinants, that is, phonetic features, are found. Prokić and Cysouw (2013) relied on the multi-aligned Bulgarian data to simultaneously track the regularity of sound correspondences through the lexicon and through the space, that is, through the population of speakers. This approach combines the co-occurrence of the phones extracted from the multi-aligned strings and network-based approach to the spread of linguistic innovations through space. It allows researchers to detect areas of intensive language contact and to infer *focal*, *transitional*, and *relic* areas of the spread of sound change.

19.4.6 Naive Discriminative Learning

Wieling *et al.* (2014) measured pronunciation distances based on naive discriminative learning (NDL). The intuition behind their approach is that a speaker of a certain dialect or language variety is predominantly exposed to speakers who speak similarly, and this input shapes the network of association strengths between cues (in this case, trigrams of phonetic segments) and outcomes (in this case, the meaning of the pronounced word) for the speaker.

The learner tries to optimize word discrimination only, and the resulting function matches how different foreign pronunciations sound. This suggests that dialect perception need not be learned separately. Validation indicated that the results are comparable with results, which are obtained with Levenshtein distance (see Section 19.5.3 for more details).

19.5 Consistency and Validity

From the previous sections it is clear that a great number of alternative methods is available for comparing dialects. Many of the alternatives are refinements of one another, leading to the question, “Which methods are most suitable in general?” In this regard, we distinguish between checking reliability (Section 19.5.1) and validity. We distinguish three types of validation: comparison to expert consensus (Section 19.5.2), correlation with perceptual measurements (Section 19.5.3), and geographic incoherence (Section 19.5.4).

19.5.1 Cronbach’s α

This measure applies only for item-based methods, that is, methods where variation of items is represented by categorical variables (Section 19.2) or phonetic transcriptions (Section 19.4). Cronbach’s α is a popular method to measure consistency or reliability. Cronbach (1951) proposed the coefficient as a lower bound to the reliability coefficient in classical test theory. The value of α indicates the extent to which the items measure the same concept.

When n dialects are compared to each other on the basis of m items, for each item we obtain a matrix that contains $(n \times (n-1))/2$ distances. The matrices are correlated with each other, and the sum of the correlations is divided by the number of matrix pairs, which is $(m \times (m-1))/2$. This is called the average inter-item correlation \bar{r} . Cronbach’s α can be written as a function of the number of items and the average inter-item correlation among the words:

$$\alpha = \frac{m\bar{r}}{1 + (m-1)\bar{r}}$$

Cronbach’s α normally ranges between 0 and 1. The higher the α , the more reliable the method. A widely accepted threshold in social science is that α should be 0.70 or higher for a set of items to be considered a reliable scale (Nunnally 1978).

Cronbach’s alpha is commonly used in two different ways, first in testing to evaluate how reliable a test is. Following this, Cronbach’s α has been used in numerical dialectometric studies (Nerbonne and Heeringa 1997; Heeringa 2004; Szemrećsanij 2008; Spruit *et al.* 2009) in order to evaluate the reliability of item sets with respect to their size. Heeringa *et al.* (2006) wrote that “Cronbach’s measure rises with the sample size, and it is therefore normally used to determine whether samples are large enough to provide reliable signals.” It implicitly assumes that there is a single signal in the data and checks to see how strong it is.

Second, it’s also used in factor analysis (FA) to check whether the variables collected into a factor define it really. Grieve (2013) wrote that a high Cronbach’s α “does not indicate that a set of linguistic variables follows a similar regional pattern.” We assume that Grieve has the FA use in mind, and he is right that one might separate the variables into different factors, each with a high Cronbach’s α , even while the whole set of variables doesn’t. But if a large set of variables has a high Cronbach’s α , then those variables are definitely signaling consistently. Moreover, when examining large sets, high Cronbach’s α values are usually encountered.

19.5.2 Expert Consensus

Heeringa *et al.* (2002) and Prokić and Nerbonne (2008) examined several techniques for validating dialectometric measures with respect to a “gold standard.” A gold standard provides a classification of language varieties with which (nearly) all experts agree. Dialect distances are converted to partitions and compared to the gold standard using the Rand index (Rand 1971), Fowlkes and Mallows Index (Fowlkes and Mallows 1983) and entropy and purity measures (Zhao and Karypis 2001).

19.5.3 Perception

Gooskens and Heeringa (2004) proposed that one examines the correlations of dialectometric measurements with the results of psychoacoustic judgments of similarity. For each of 15 varieties a recording of the fable “The North Wind and the Sun” was presented to 15 groups of Norwegian high school pupils, one group from each of the 15 dialect sites represented in the material. All pupils were familiar with their own dialect and had lived most of their lives in the place in question. The 15 dialects were presented in a randomized order. While listening to the dialects, the listeners were asked to judge each of the 15 dialects on a scale from 1 (similar to native dialect) to 10 (not similar to native dialect). This means that each group of listeners judged the linguistic distances between their own dialect and the 15 dialects, including their own dialect. In this way we get a matrix with 15×15 perceived linguistic distances. In order to use this material to calibrate the different computational measurements, the correlations between the 15×15 computational matrices with the 15×15 perceptual matrix were examined. While calculating correlations, the distances of dialects with respect to themselves were excluded. In Table 19.1 we give an overview of the main results from validation studies that are based on this material.

Wieling *et al.* (2014a) validated the PMI-based Levenshtein distance applied to over 800 foreign pronunciations (of 69 words each) in a collection of several hundred native speakers’ judgments of “nativelikeness,” obtaining very strong correlation ($r = -0.81$).

19.5.4 Geography

It is fundamental to dialectology that the linguistic distances between geographically closer varieties are, in general, smaller. In order to test the extent to which dialectometric distances agree with this fundamental dialectological postulate, Nerbonne and Kleiweg (2007) introduced a measure of *local incoherence*.

The basic idea is that we begin with each measurement site s , and inspect the n linguistically most similar sites in order of decreasing linguistic similarity to s . We then measure how far away these linguistically most similar sites are geographically, for example, in kilometers. Unlike poor measurements, good ones show that linguistically similar sites are also geographically close. Nerbonne and Kleiweg (2007) restricted the attention to the eight most similar linguistic varieties in calculating local incoherence. The lower the local incoherence value, the better the measurement technique. The advantage of this method is that it does not require expert consensus (Section 19.5.2) or perceptually measured dialect distances (Section 19.5.3).

19.6 Available Tools

VDM is short for “Visual DialectoMetry,” and refers to a powerful computer program created by Edgar Haimerl (Blaustein, Germany) between 1997 and 2000 in Salzburg in cooperation with Hans Goebel. The program offers the possibility to calculate and visualize

Table 19.1 Correlations of linguistic measures with Norwegian perceptual distances.
The Levenshtein variants given here generate absolute distances with forced alignment (see Section 19.4.2).

Method	Reference	Correlation			
		All-word		Same-word	
		Comprehensive	Positional	Comprehensive	Positional
lexical	RIW	Gooskens and Heeringa (2006)	0.29 ²		
	GIW		0.37 ²		
frequency method	phones	Heeringa (2004)	0.66		
	features		0.47		
Levenshtein distance	unigrams	Heeringa <i>et al.</i> (2006)	0.66 ²	0.71 ²	
	bigrams		0.72 ²	0.71 ²	0.67 ²
	trigrams		0.73 ²	0.72 ²	0.69 ²
	xbigrams		0.73 ²	0.72 ²	0.69 ²
	PMI		0.63 ²	0.71 ²	0.68 ²
	NDL	Wieling <i>et al.</i> (2014)		0.72	

similarities between local dialects on the basis of large numbers of categorical variables, and between local dialects and one particular reference point, for example, a standard language. The program offers the possibility of classifying dialects by means of cluster analysis. VDM is a desktop application. More information can be found at <http://dialectometry.com/dmdocs/index.html>.

Gabmap has initially been created at the University of Groningen by Peter Kleiweg under supervision of John Nerbonne (see Nerbonne *et al.* 2011). In addition to VDM, Gabmap offers the possibility to use Levenshtein distance, noisy clustering, and multidimensional scaling. Gabmap is a web application, available at <http://www.gabmap.nl/>.

A group of researchers at the University of the Basque Country (UPV-EHU) and the Basque Summer University (UEU) created “DiaTech,” a freely available tool for doing dialectometric analyses (see Aurrekoetxea *et al.* 2013). This program has incorporated features of previous programs, and especially of the VDM program created under the direction of Goebel.

19.7 Future Challenges

We note several areas where improvements ought to be within reach. With respect to lexis, we like to see experimentation with weightings, other than RIW, for example, weightings of lexemes by their frequency of use. In pronunciation, we note that, whereas aggregate Levenshtein distances have been validated, validation might be made more fine-grained by considering users’ reactions to individual words. In the approach in Section 19.5.3,

perceptual distances are also influenced by lexical and syntactical variation. Perceptual distances between words might be obtained on the basis of purely phonic variation. It would be interesting to see more work on the regularity of sound correspondences between dialects. This can, for example, be measured with Shannon's entropy. Prosody has not been studied much in dialectometry; in fact, it has been studied only to quantify (lexical) tone differences (Gooskens and Heeringa 2006, Yang and Castro 2008). An acoustically based measure might compare pitch patterns over larger stretches of speech.

Morphological variation has been measured categorically, but there is not yet an accepted way of evaluating the case where morphological categories are mismatched, for example, where one variety has a three-gender system and the other two. Heeringa *et al.* (2014) used Levenshtein distance for measuring affixal distances between languages, which can also be applied to dialect data. This might provide a fresh perspective on morphological variation. In syntax, the measure introduced by Nerbonne and Wiersma (2006) has been applied to foreigners' speech. We would like to see the method applied to (large) dialect data sets. Additionally, Heeringa *et al.* (2017) compared sentence structures of languages by means of a Levenshtein distance-like approach. This method can also be used for quantifying syntactical dialect differences.

Finally, it would be valuable to have a measure that combines the linguistic levels. The challenge will be to find the right weight of each level. We suggest that these can be found by comparing the levels to perceptual distances in a multiple linear regression model.

NOTES

- 1 <http://alignments.lingpy.org>.
- 2 The correlations shown here have been calculated by the first author of this chapter and may differ from the corresponding ones published in literature due to corrections in the data which have been made afterward. Due to a bug in the software the results in Heeringa *et al.* (2006) are not reliable. The results presented here are correct. The correlations presented in the column 'Same-word' have not been published before.

REFERENCES

- Alonso, Laura, Irene Castellón, Jordi Escribano, Xavier Messeguer, and Lluís Padró. 2004. "Multiple sequence alignment for characterizing the linear structure of revision." In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, 403–406.
- Aurrekoetxea, Gotzon, Karmele Fernandez-Aguirre, Jesús Rubio, Borja Ruiz, and Jon Sánchez. 2013. "'DiaTech': A new tool for dialectology." *Literary and Linguistic Computing* 28(1): 23–30.
- Beijering, Karin, Charlotte Gooskens, and Wilbert Heeringa. 2008. "Modelling intelligibility and perceived linguistic distances by means of the Levenshtein algorithm." In *Linguistics in the Netherlands 2008* edited by M. van Koppen and B. Botma, 13–24. Amsterdam: John Benjamins Publishing Company.
- Bhargava, Aditya, and Grzegorz Kondrak. 2009. "Multiple word alignment with Profile Hidden Markov Models." In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Student Research Workshop and Doctoral Consortium*, 43–48.
- Bolognesi, Roberto, and Wilbert Heeringa. 2002. "De invloed van dominante talen op het lexicon en de fonologie van Sardische

- dialecten." *Gramma/TTT: tijdschrift voor taalwetenschap*, 9(1): 45–84.
- Cronbach, Lee Joseph. 1951. "Coefficient alpha and the internal structure of tests." *Psychometrika* 16–3: 297–334.
- Di Buccio, Emanuele, Giorgio Maria Di Nunzio, and Gianmario Silvello. 2014. "A vector space model for syntactic distances between dialects." *Language Resources and Evaluation Conference*. Paris: ELRA. 2486–2489.
- Fowlkes, Edward B., and Colin L. Mallows. 1983. "A method for comparing two hierarchical clusterings." *Journal of the American Statistical Association* 78: 553–569.
- Goebl, Hans. 1982. *Dialektometrie; Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie*. Wien: Verlag der Öst. Akademie der Wissenschaften.
- Goebl, Hans. 1984. *Dialektometrische Studien anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*. Volume 1. (Volumes 2 and 3 contain maps and tables). Tübingen: Max Niemeyer.
- Goebl, Hans. 2010a. "Dialectometry and quantitative mapping." In *Language and Space. An International Hand of Linguistic Variation. Volume 2: Language Mapping. Handbücher zur Sprach- und Kommunikationswissenschaft [HSK]*, edited by Alfred Lameli, Roland Kehrein and Stefan Rabanus, 30.2, 433–457, 2201–2212. Berlin: de Gruyter Mouton.
- Goebl, Hans. 2010b. "Dialectometry: Theoretical prerequisites, practical problems, and concrete applications (mainly with examples drawn from the *Atlas Linguistique de La France*, 1902–1910)." *Dialectologia*. Special Issue, I(2010): 63–77.
- Gooskens, Charlotte, and Wilbert Heeringa. 2004. "Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data." *Language Variation and Change*, 16–3: 189–207.
- Gooskens, Charlotte, and Wilbert Heeringa. 2006. "The relative contribution of pronunciation, lexical and prosodic differences to the perceived distances between Norwegian dialects." *Literary and Linguistic Computing* 21(4): 477–492. Special issue on *Progress in Dialectometry: Toward Explanation*, edited by John Nerbonne and William Kretschmar, Jr.
- Grieve, Jack. 2014. "A comparison of statistical methods for the aggregation of regional linguistic variation." In *Aggregating dialectology, typology, and register analysis: Linguistic variation in text and speech* edited by Benedikt Szmrecsanyi and Bernhard Wälchli. Berlin/New York: Walter de Gruyter.
- Heeringa, Wilbert. 2004. *Measuring dialect pronunciation differences using Levenshtein distance*. Ph.D. dissertation University of Groningen.
- Heeringa, Wilbert, and Angelika Braun. 2003. "The use of the Almeida-Braun system in the measurement of Dutch dialect distances." *Computers and the Humanities* 37(3): 257–271.
- Heeringa, Wilbert, and John Nerbonne. 2006. "De analyse van taalvariatie in het Nederlandse dialectgebied: Methoden en resultaten op basis van lexicon en uitspraak." *Nederlandse Taalkunde* 11(3): 218–257.
- Heeringa, Wilbert, and Frans Hinskens. 2015. "Dialect change and its consequences for the Dutch dialect landscape. How much is due to the standard variety and how much is not?" *Journal of Linguistic Geography* 3(1): 20–23.
- Heeringa, Wilbert, John Nerbonne, and Peter Kleiweg. 2002. "Validating dialect comparison methods." In *Proceedings of the 24th Annual Meeting of the Gesellschaft für Klassifikation*, edited by W. Gaul & G. Ritter, 445–452. Heidelberg: Springer.
- Heeringa, Wilbert, Peter Kleiweg, Charlotte Gooskens, and John Nerbonne. 2006. "Evaluation of string distance algorithms for dialectology." In *Linguistic Distances Workshop at the joint conference of International Committee on Computational Linguistics and the Association for Computational Linguistics, Sydney, July, 2006*, edited by John Nerbonne and Erhard Hinrichs, 51–62. Stroudsburg PA: The Association for Computational Linguistics (ACL).
- Heeringa, Wilbert, Femke Swarte, Anja Schüppert, and Charlotte Gooskens. 2014. "Modeling intelligibility of written Germanic languages: do we need to distinguish between orthographic stem and affix variation?" *Journal of Germanic Linguistics* 26(4): 361–394.
- Heeringa, Wilbert, Femke Swarte, Anja Schüppert, and Charlotte Gooskens. 2017. "Measuring Syntactical Variation in Germanic Texts." To appear in: *Digital Scholarship Humanities*.
- Heeringa, Wilbert, John Nerbonne, and Petya Osenova. 2010. "Detecting contact effects in pronunciation." In *Language Contact. New Perspectives*, edited by Muriel Norde, Bob de Jonge and Cornelius Hasselblatt, 131–153. Series IMPACT: Studies in Language and Society. Amsterdam: Benjamins.
- Hoppenbrouwers, Cor, and Geer Hoppenbrouwers. 1988. "De feature frequentie methode en de classificatie van Nederlandse

- dialecten." *TABU, Bulletin voor taalwetenschap*, 18: 51–92.
- Hoppenbrouwers, Cor and Geer Hoppenbrouwers. 2001. *De indeling van de Nederlands streektalen: dialecten van 156 steden en dorpen geklasseerd volgens de FFM*. Assen: Koninklijke Van Gorcum.
- Inkpen, Diana, Oana Frunza, and Grzegorz Kondrak. 2005. "Automatic identification of cognates and false friends in French and English. In *International Conference Recent Advances in Natural Language Processing*, edited by G. Angelova, K. Bontcheva, R. Mitkov, N. Nicolov, and N. Nicolov, 251–257. Borovets.
- Jäger, Gerhard. 2013. "Phylogenetic inference from word lists using weighted alignment with empirically determined weights." In *Language Dynamics and Change*, 3(2): 245–291.
- Kessler, Brett. (1995). "Computational dialectology in Irish Gaelic." In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, pages 60–67, Dublin. EACL.
- Kondrak, Grzegorz. 2005. "N-gram similarity and distance." *Proceedings of the 12th International Conference, SPIRE 2005, Buenos Aires, Argentina, November 2-4, 2005*, 115–126.
- Kruskal, Joseph B. 1999. "An overview of sequence comparison." In *Time Warps, String edits, and Macromolecules. The Theory and Practice of Sequence Comparison* edited by D. Sankoff, and J. Kruskal, 2nd ed., 1–44. Stanford: Center for the Study of Language and Information. 1st edition appeared in 1983.
- Lauttamus, Timo, John Nerbonne, and Wybo Wiersma. 2007. "Detecting syntactic contamination in emigrants: The English of Finnish Australians." *SKY Journal of Linguistics* 21(2007): 273–307.
- Levenshtein, Vladimir I. 1966. "Binary codes capable of correcting deletions, insertions, and reversals." *Cybernetics and Control Theory*, 10(8): 707–710.
- List, Johann-Mattis, and Jelena Prokić. 2014. "A benchmark database of phonetic alignments in historical linguistics and dialectology." In *Proceedings of the International Conference on Language Resources and Evaluation (LREC) 2014, 26 May-31 May 2014, Reykjavik, Iceland*.
- List, Johann-Mattis, and Steven Moran. 2013. "An open source toolkit for quantitative historical linguistics." In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, August 4-9, Sofia, Bulgaria*, 13–18.
- Nerbonne, John, Rinke Colen, Charlotte Gooskens, Peter Kleiweg, and Therese Leinonen. 2011. "Gabmap - a web application for dialectology." *Dialectologia: revista electrônica*, special issue on Production, Perception and Attitude edited by John Nerbonne, Stef Grondelaers, Dirk Speelman & Maria-Pilar Perea, 65–89.
- Nerbonne, John, and Wilbert Heeringa. 1997. "Measuring dialect distance phonetically." In *Workshop on Computational Phonology, Special Interest Group of the Association for Computational Linguistics*, edited by John Coleman, 11–18.
- Nerbonne, John, Wilbert Heeringa, Erik van den Hout, Peter van der Kooi, Simone Otten, and Willem van de Vis. 1996. "Phonetic distance between Dutch dialects." In *CLIN VI, Papers from the sixth CLIN meeting*, edited by G. Durieux, W. Daelemans, and S. Gillis, 185–202. Antwerp: University of Antwerp, Center for Dutch language and speech.
- Nerbonne, John, and Peter Kleiweg. 2003. "Lexical distance in LAMSAS." In *Computational Methods in Dialectometry*, special issue of *Computers and the Humanities*, edited by John Nerbonne and William Kretschmar, 37(3), 339–357.
- Nerbonne, John, and Peter Kleiweg. 2007. "Toward a dialectological yardstick." *Journal of Quantitative Linguistics* 14(2), 148–167.
- Nerbonne, John, Jelena Prokić, Martijn Wieling, and Charlotte Gooskens. 2010a. "Some further dialectometrical steps." In *Tools for Linguistic Variation Bilbao: Supplements of the Anuario de Filología Vasca "Julio Urquijo"*, edited by G. Aurrekoetxea and J.L. Ormaetxea, XIII. 41–56.
- Nerbonne, John and Wybo Wiersma. 2006. "A measure of aggregate syntactic distance." In *Linguistic Distances Workshop at the joint conference of International Committee on Computational Linguistics and the Association for Computational Linguistics, Sydney, July, 2006*, edited by J. Nerbonne & E. Hinrichs, 82–90.
- Nerbonne, John, Timo Lauttamus, Wybo Wiersma, and Lisa Lena Opas-Hänninen. 2010b. "Applying language technology to detect shift effects." In *Language Contact. New Perspectives*, edited by Muriel Norde, Bob de Jonge and Cornelius Hasselblatt, 27–44. Series IMPACT: Studies in Language and Society. Amsterdam: Benjamins.
- Nunnally, Jum C. 1978. *Psychometric Theory*. McGraw-Hill, New York.
- Prokić, Jelena, and John Nerbonne. 2008. "Recognizing groups among dialects."

- International Journal of Humanities and Arts Computing* 2(1-2), Special Issue on Language Variation, 153–172.
- Prokić, Jelena, Martijn Wieling, and John Nerbonne. 2009. “Multiple string alignments in linguistics.” In *Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH – SHELT&R 2009) EACL Workshop* edited by Lars Borin and Piroska Landvai, 18–25.
- Prokić, Jelena. 2010. *Families and Resemblances*. PhD thesis. University of Groningen.
- Prokić, Jelena, Çağrı Çöltekin, and John Nerbonne. 2012. “Detecting shibboleths.” In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, 72–80. Avignon: Association for Computational Linguistics.
- Prokić, Jelena, and John Nerbonne. 2013. “Analyzing dialects biologically.” In *Classification and Evolution in Biology, Linguistics and the History of Science. Concepts, Methods, Visualization*, 147–161. Stuttgart: Steiner Verlag.
- Prokić, Jelena, and Michael Cysouw. 2013. “Combining regular sound correspondences and geographic spread.” *Language Dynamics and Change*, 3(2), 147–168.
- Rand, William M. 1971. “Objective criterion for evaluation of clustering methods.” *Journal of the American Statistical Association*, 66: 846–850.
- Sampson, Geoffrey. 2000. “A proposal for improving the measurement of parse accuracy.” *International Journal of Corpus Linguistics*, 5(1): 53–68.
- Sanders, Nathan C. 2007. “Measuring syntactic difference in British English.” In *Proceedings of the ACL 2007 Student Research Workshop*, 1–6. Madison: Omnipress.
- Seguy, Jean. 1973. “La dialectometrie dans l’Atlas linguistique de la Gascogne.” *Revue de linguistique romane*, 37: 1–24.
- Spruit, Marco. 2008. *Quantitative perspectives on syntactic variation in Dutch dialects*. PhD thesis, University of Amsterdam, LOT Dissertation Series 174, Utrecht: Utrecht Institute of Linguistics.
- Spruit, Marco René, Wilbert Heeringa, and John Nerbonne. 2009. “Associations among linguistic levels.” In *Lingua*, special issue on *The Forests behind the Trees*, edited by John Nerbonne and Franz Manni, 119(11): 1624–1642.
- Steiner, Lydia, Peter F. Stadler, and Michael Cysouw. 2011. “A pipeline for computational historical linguistics.” *Language Dynamics and Change*, 1(1), 89–127.
- Wichmann, Søren, André Müller, Viveka Velupillai, Annkathrin Wett, Cecil H. Brown, Zarina Molochieva, Julia Bishoffberger, Eric W. Holman, Sebastian Sauppe, Pamela Brown, Dik Bakker, Johann-Mattis List, Dmitry Egorov, Oleg Belyaev, Matthias Urban, Harald Hammarström, Agustina Carrizo, Robert Mailhammer, Helen Geyer, David Beck, Evgenia Korovina, Pattie Epps, Pilar Valenzuela, and Anthony Grant. 2012. *The ASJP Database* (version 15). <http://email.eva.mpg.de/~wichmann/ASJPHomePage.htm>
- Szmrecsanyi, Benedikt. 2008. “Corpus-based dialectometry: aggregate morphosyntactic variability in British English dialects.” In *International Journal of Humanities and Arts Computing, special issue on Language Variation*, edited by John Nerbonne, Charlotte Gooskens, Sebastian Kürschner and Renée van Bezooijen, 2(1–2): 279–296.
- Szmrecsanyi, Benedikt. 2010. “Geography is overrated.” In *Dialectological and Folk Dialectological Concepts of Space*, edited by S. Hansen, C. Schwarz, P. Stoeckle, and T. Streck. Berlin: Walter de Gruyter.
- Szmrecsanyi, Benedikt, and Christoph Wolk. 2011. “Holistic corpus-based dialectology.” In *Brazilian Journal of Applied Linguistics/Revista Brasileira de Linguística Aplicada*, special issue on *Corpus studies: future directions*, edited by Stefan Th. Gries, 11(2): 561–592.
- Szmrecsanyi, Benedikt. 2013. *Grammatical Variation in British English Dialects: A Study in Corpus-Based Dialectometry*. Series: Studies in English Language. Cambridge: Cambridge University Press.
- Wälchli, Bernhard. 2009. “Data reduction typology and the bimodal distribution bias.” *Linguistic Typology* 13: 77–94.
- Wells, John, and Jill House. 1995. *The sounds of the international phonetic alphabet*. London: University College, Department of Phonetics and Linguistics.
- Wieling, Martijn, Jelke Bloem, Kaitlin Mignella, Mona Timmermeister, and John Nerbonne. 2014a. “Measuring foreign accent strength in English. Validating Levenshtein distance as a measure.” *Language Dynamics and Change* 4(2): 253–269.
- Wieling, Martijn, Eliza Margaretha, and John Nerbonne. 2012. “Inducing a measure of phonetic similarity from dialect variation.” *Journal of Phonetics*, 40(2), 307–314.
- Martijn Wieling, John Nerbonne, Jelke Bloem, Charlotte Gooskens, Wilbert Heeringa, and R. Harald Baayen. 2014. A cognitively

- grounded measure of pronunciation distance.
PLOS ONE, 9(1): 1–7.
- Wieling, Martijn, Jelena Prokić, and John Nerbonne. 2009. “Evaluating the pairwise alignment of pronunciations.” In *Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities and Education*, edited by L. Borin and P. Lendvai, 26–34. Athens: EACL.
- Wybo Wiersma, John Nerbonne, and Timo Lauttamus. 2010. “Automatically extracting typical syntactic differences from corpora.”
- Literary and Linguistic Computing* 26(1), 107–124.
- Yang, Cathryn, and Andy Castro. 2008. “Representing tone in Levenshtein distance.” *International Journal of Humanities and Arts Computing* 2(1–2): 205–219.
- Zhao, Ying, and George Karypis. 2001. *Criterion functions for document clustering: Experiments and analysis*. Technical report 01–40, Department of Computer Science, University of Minnesota, Minneapolis, MN.

20 Dialect Maps

STEFAN RABANUS

20.1 Introduction

Dialect maps are visualizations of the spatial distribution of linguistic features or, more generally, of feature-based areal structures. In order to address the issue of dialect maps properly we have to clarify, first, the key notions “dialect,” “linguistic feature,” and “feature-based areal structure” (in this introduction) before proceeding to the introduction of the mapping techniques (Section 20.2), the purposes of dialects maps and key examples of good practice (Section 20.3), future perspectives, challenges, and risks in the field of dialect maps (Section 20.4). This article offers a typology of mapping techniques and the purposes of dialect maps but does not provide a systematic survey of the history of linguistic cartography. For this issue, refer to, for example, the articles by Lameli (2010), Veith (2006a, 2006b), and with respect to language-specific or national traditions, the chapters of part II of the *Language Mapping* handbook edited by Lameli, Kehrein, and Rabanus (2010).

The notion of “dialect” has quite different meanings, according to the scientific tradition in which it is used (see also in the introduction to this handbook). In the Anglo-Saxon tradition “dialect” is an equivalent of “variety,” so there can be something like a “standard dialect,” whereas in the Continental European, especially in the German perspective “dialect” is used only for geographically delimited varieties, opposed to “standard variety” (cf. Britain 2004, p. 268). In this chapter, “dialects” are regarded as “geolects,” which may include varieties such as British and American English—which can be delimited geographically—but exclude pure “sociolects” (Durrell 2004, p. 201: “social dialects”; Trudgill 1975a, p. 34: “social-class dialects”). We will see below that this does not exclude sociolinguistic issues from our topic. We take the position that all branches of linguistics with a clear reference to geographic space can be summarized under the notion of “areal linguistics.”¹

The typical dialect map visualizes the areal extension of single “linguistic features,” with regard to the variation of sounds, forms and words of a language. In Figure 20.1, the distributional areas of the variants for “hungry” in the dialects of England are shown (AED, map 53, here reproduced from Upton 2010, map 0705). Dialect maps require a reference system. The variable whose variants are depicted on the map may be represented by a form of the respective standard language (standard English *hungry* in Figure 20.1) or by a form from a historical reference system (see subsection 20.3.4). “Feature-based areal structures” are the outcome of aggregate analyses in which the comparison of the areal distribution of many single features leads to the delimitation of dialect areas (Trudgill 1975b, p. 233: “focal areas”). The very first dialect maps were visualizations of feature-based areal structures, and this



Figure 20.1 Isogloss map (variants for “hungry” in the dialects of England, AED, map 53).

type of map developed from maps with only intuitively delimited dialect areas as Bernhardi’s *Sprachkarte von Deutschland* (Bernhardi 1843)² through maps like Ellis’ subdivision of English dialects into districts in *Existing Phonology of English Dialects* (1889), which was already based on a quite accurate analysis of the areal distribution of major phonological feature, to today’s statistically calculated similarity and dissimilarity maps (cf. Nerbonne 2010, Goebel 2010).

20.2 Mapping Techniques

We consider “maps” *sensu stricto* to be geographical models (Ormelinc 2010, p. 22): “Maps are two-dimensional graphic models of (parts of) the Earth’s surface—or of geospatial phenomena related to that surface—produced to scale for decision-making purposes.” In the following subsections (20.2.1–20.2.4) we discuss the various types of dialect maps in this strict sense. For atypical visualizations that are maps only in a broader sense because the locations are not positioned according to their geographic coordinates (e.g., MDS plots, dendograms, neighbor

nets) please refer to Rabanus (2011, pp. 44–45), Szmrecsanyi and Anderwald (in this volume, Chapter 17) and Nerbonne and Wieling (in this volume, Chapter 23). An easily readable English introduction on how to map language data is Upton (2010).

A dialect map in the strict sense consists of two layers. The basic layer is derived from a topographical map of the survey area that provides the minimum of information, which is necessary in order to enable the map reader to locate the linguistic data in the geographic space. Most maps use rivers, coast lines, major cities, and/or political or administrative borderlines for this purpose. Usually, the basic layer is graphically coded as background information using weak colors (e.g., light gray for county boundaries in Figure 20.1). The main layer contains the linguistic information, represented by graphic elements that are more prominent than the geographical background information. These graphic elements may be point-related, line-related, area-related or surface-related (for the following subsections cf. Rabanus 2012 and Bollmann 2010, pp. 51–54). Bear in mind, however, that the ultimate goal of most dialect mapping is the delimitation of areal structure (core areas, transition areas), that is, even point-related maps (Fig. 20.2) and line-related maps (Fig. 20.1) usually serve that purpose. Furthermore, combinations of graphic means in dialect maps are more common than pure representatives of the types.

20.2.1 Point-Related Maps

In point-related maps dots or other symbols (triangles, circles, or other geometrical signs) are drawn at the points where the data is located. “Point,” ideally, means the exact geographical position (geographic coordinates: latitude and longitude) of the data informant’s location. Since graphic symbols on the map sheet have a graphic extension, their position can be shifted with respect to the location’s exact geographic coordinates because of the scale, the density of symbols, or for aesthetic reasons. The first and most frequent subtype of point-related maps is the *qualitative point-symbol map*. In these maps the linguistic data are grouped and represented by symbols, which denote the same quantity but different quality by different shape or color. Figure 20.2 is a good example from the *Carpathian Dialect Atlas* (OKDA, vol. 5, map 73) because relatively few different symbols denote the linguistic differences. Less-effective examples of this type are maps in which—although they perfectly correspond to the definition of point-symbol maps—the large number of different symbols makes the areal interpretation difficult, as for example in map 1 of the *Linguistic Atlas of the Northern and Southern Netherlands* (TNZN). Qualitative point-symbol maps usually display one linguistic variant per location, regardless of the number of attestations, informants, or linguistic variation at the location (Fig. 20.2). The number of attestations is symbolized in *proportional point-symbol maps*: in the *Historischer Südwestdeutscher Sprachatlas* (HSS) dot size corresponds to the number of attestations in historic documents (see also Fig. 20.6, in subsection 20.2.3). Variation at the location can be visualized by means of *point-diagram maps*. Figure 20.3 is a phonological map from the *Dialect Atlas of the Armenian Language* (Sargsyan 2008, map 10 [originally in color]) in which the variation at the locations is shown by bipartite symbols (map legend: “-/+” symbols below the feature numbers 38–42).

In the *Atlas of North American English* (ANAE) the linguistic data are related to single informants giving rise to a kind of *dot-density map* that is intended to mirror the differences in population density across the territory. However, with respect to geographic standards the ANAE representation is problematic for two reasons. First, while the overall picture may mirror the population density, in many cases the number of dots per location does not correspond to the population (e.g., in the maps of chapter 9 [ANAE, pp. 58–63, North American mergers in progress], San Diego, California, urbanized area population 2,348,106, is represented by two dots; Aberdeen, South Dakota, population 24,927, by six dots [cf. ANAE,

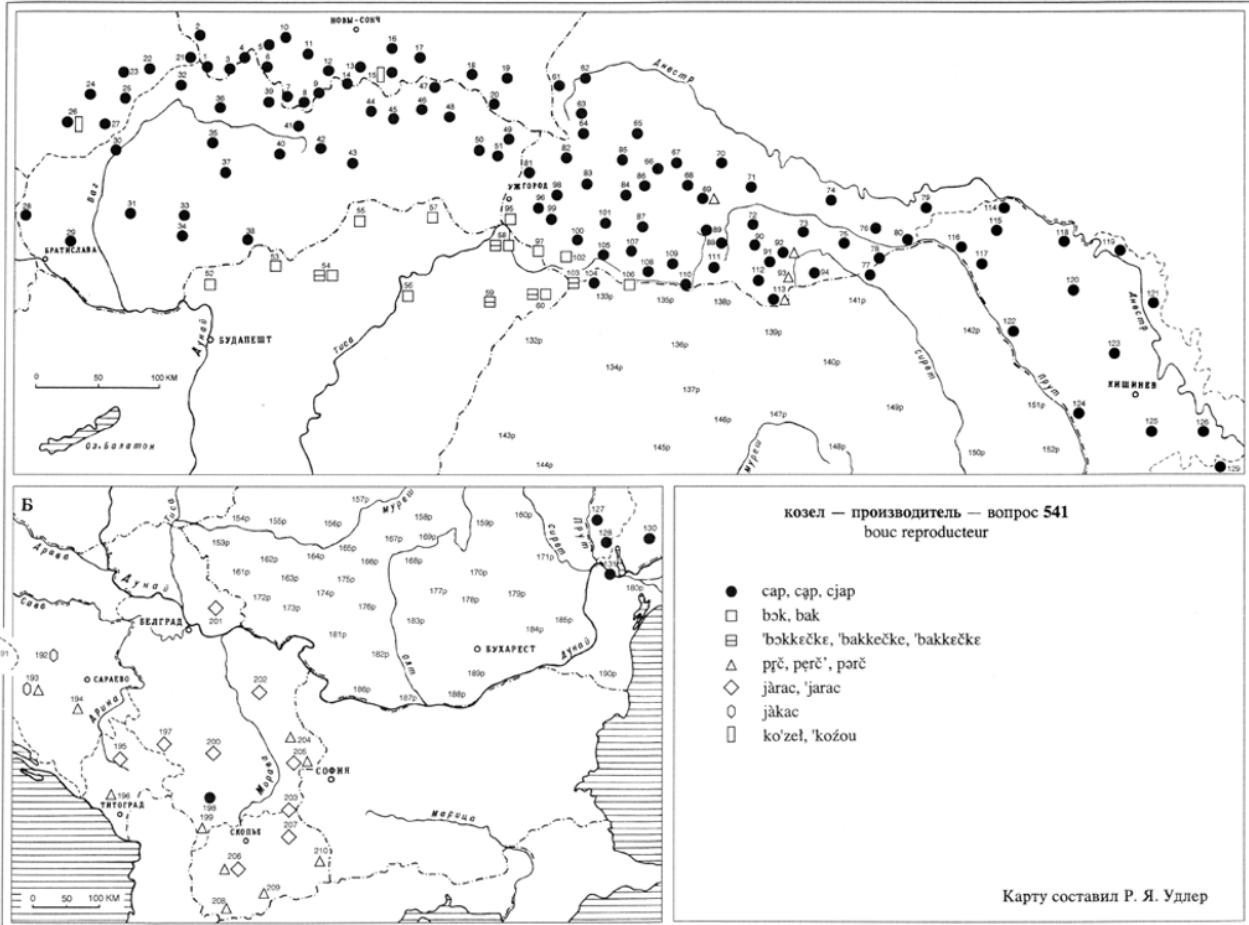


Figure 20.2 Point-symbol map (lexemes for “billy goat” in the dialects of the Carpathian Mountains, OKDA, vol. 5, map 73).



Figure 20.3 Point diagram map (several vowel features for the historically [1914] Armenian-speaking territories in the Middle East, Sargsyan 2008, map 10 [detail]).

appendix 4.1, pp. 30–31]). Second, with several dots per location the geographically faithful positioning of the dots gets lost.

The second subtype of point-related maps is the *point-text map*. In point-text maps written forms (words or phonetic transcriptions) are superimposed directly at the locations, without any further symbolization and without the linguistic classification that is necessary to group the data in order to draw point-symbol maps. The method was first used in the *Atlas linguistique de la France* (ALF) and then widely applied for all Romance languages. Figure 20.4 shows the variants for “oil lamp” in the Southern Italian dialects (*Sprach- und Sachatlas Italiens und der Südschweiz* [AIS], vol. 5, map 915, detail; see also subsection 20.3.2). Point-text maps are geographically coded databases rather than maps in the above defined way since it is almost impossible to get an immediate picture of the data’s geographical distribution. On the other hand, scholars of Romance languages often emphasize that the point-text maps were only one aspect of the French method that necessarily included the drawing of dialect areas by means of isoglosses on so-called “mute maps”, which were provided together with the ALF maps (cf. Goebel 2004, p. 529).

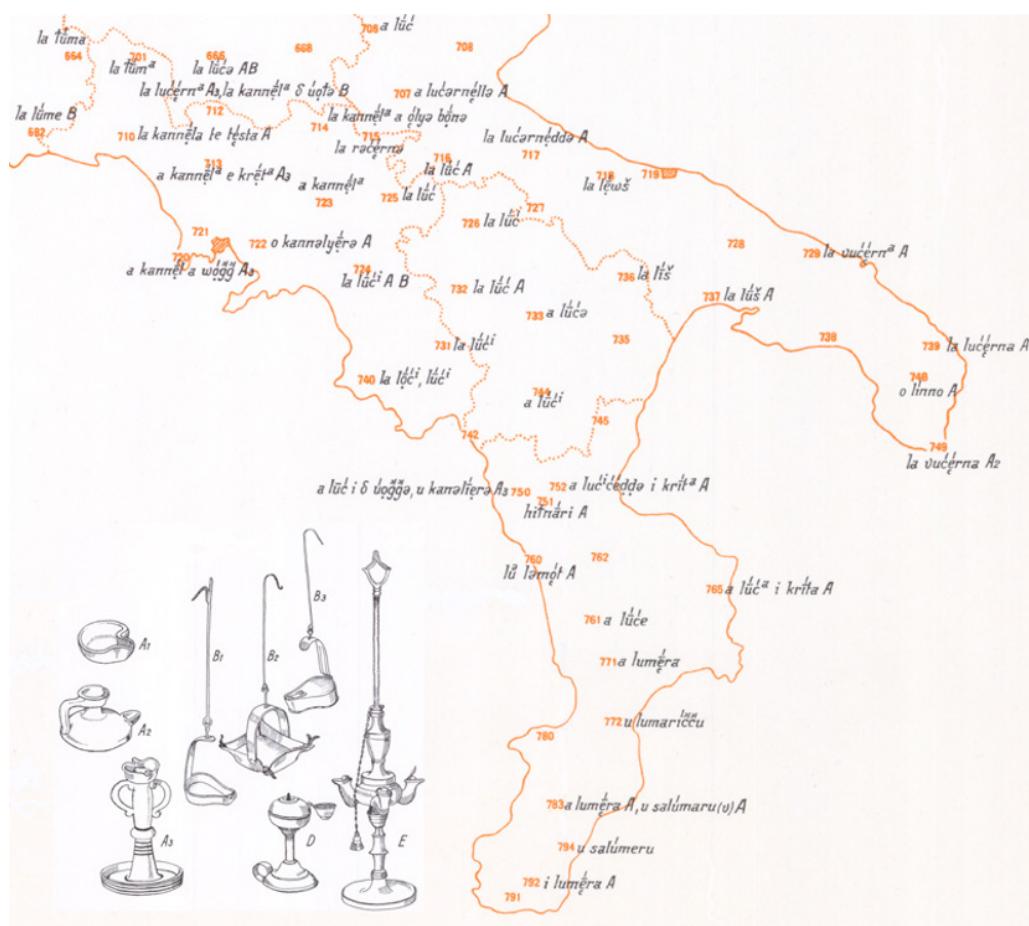


Figure 20.4 Point-text map (variants for “oil lamp” in the Southern Italian dialects, AIS, vol. 5, map 915 [detail]). The numbers (e.g., “791” etc.) refer to collection sites.

20.2.2 Line-Related Maps

Line-related maps comprise two different types. In the first type, lines are drawn between locations with different linguistic data in order to separate them from each other. These lines, which in most cases occur in form of isoglosses (see subsection 20.4.1), constitute areal structures, thus, the maps are area maps rather than line maps (see Fig. 20.1). In the second type, lines connect locations, thus, lines are used in their proper sense. In perceptual dialectology, line maps are used to display the degree of perceived similarities between local dialects (cf. Preston 2010 and Preston, in this volume, Chapter 10). In historical linguistics directed line (arrow) maps are often used to visualize spread of features and directions in language change: Figure 20.5 (Schirmunski 1956, p. 374, map 15) shows the movements of the variants of the intervocalic consonant cluster *-hs/ss/ks-* in Germany with arrows. In linguistic typology, the spread of varieties (carried by migrating populations) may be shown by lines (arrows) on maps (e.g., Wurm *et al.* 1996, map 49, "Formation and spread of Tok Pisin 1880–1920"). Network maps (beam maps) are undirected line maps in which the lines, connecting locations, differ in color and/or darkness revealing areal structure. In "Salzburg-style" beam maps, only neighboring locations are connected (e.g., Goebel 2010, p. 449, map 2210, Interpoint similarity values for dialects of France, based on ALF data), in "Groningen-style" network maps every location may be linked to any other location of the considered area (e.g., Nerbonne 2010, p. 482, map 2402, "Aggregate pronunciation distances in Germany").³ In both, directed and undirected line maps, the thickness of the lines may reflect the degree of similarity of locations or the impact of movements.



Figure 20.5 Directed line map (movements of the variants of the intervocalic consonant cluster *-hs/ss/ks-* in Germany, Schirmunski 1956, map 15).

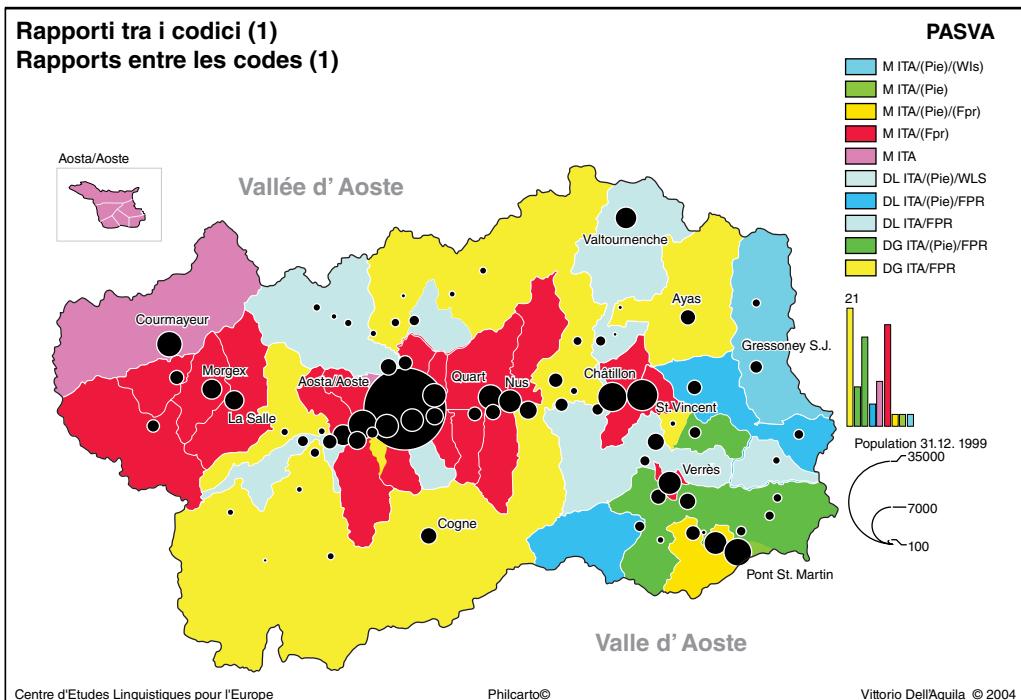


Figure 20.6 Qualitative area map/chorochromatic map (multilingualism in the Aosta Valley, Dell'Aquila 2010, map 2304 [redrawn in grayscale]).

20.2.3 Area-Related Maps

Since the main goal of linguistic cartography is the visualization of feature-based areal structures, area-related maps seem to be the best choice for that purpose. Hence, area maps are common in all dialect-mapping schools, and also used in linguistic typology. However, the concept of “area” is problematic in linguistic cartography because “most of the Earth’s surface [is] uninhabited. Hence, linguistic areas always enclose points to which linguistic values are not applicable” (Rabanus 2012, subsection 2.3) Additionally, linguistic surveys are never complete: area maps always include territories for which no linguistic data are available.

The visualization of areas on maps differs in shape and contiguity. In most cases, lines are used to delimit areas. However, lines are not constitutive for areas, and area marking without lines is possible, by means of color (e.g., the macro-areal profile of Africa by Lameli, Kehrein, and Rabanus 2010, vol. 2, map 2911, discussed by Güldemann 2010, pp. 575–577) or by means of shading or hatching (e.g., Chambers and Trudgill 1998, pp. 133–134, figs. 8–3 and 8–4, maps of “mixed” and “fudged lects” in the south of England). Area maps can be subdivided in qualitative area maps and quantitative area maps.

Qualitative area maps are drawn by considering localities with identical data as forming continuous areas. The area often is marked by coloring, shading, or hatching, where each color, shading, or hatching corresponds to a different qualitative value. This makes a *chorochromatic map*, which is common for representing language areas (e.g., many *Ethnologue* maps of the world’s languages, cf. www.ethnologue.com). In Figure 20.6 (Dell’Aquila 2010, map 2304) the multilingualism in the Aosta Valley in northern Italy is visualized (the colors in the original map correspond to the types of multilingualism, e.g., yellow [redrawn as hatch pattern with horizontal lines] stands for diglossia [DG] Italian [ITA] and Francoprovençal

[FPR]; last row of the legend, peripheral areas in the south and the north-east).⁴ In area-text maps the linguistic data of the area is made explicit by text, written inside the area on the map sheet. Sometimes there are no colors or shading but only isoglosses and text, as in Figure 20.1. Alternatively, the data of the area can be represented by symbols or diagrams, thus yielding so-called area-symbol and area-diagram maps. The difference to the point-symbol and point-diagram maps introduced above is that the graphic symbols are area related.

Quantitative area maps: Whereas chorochromatic maps can show any kind of qualitative data, in *choropleth maps* the data has to be numerical and the different degrees of coloring or shading are calculated by specific algorithms, and not arbitrarily assigned by the map drawer. An important impetus for choropleth maps in dialectology came from the so-called “Salzburg school of dialectometry,” which is well known for similarity maps in which degrees of coloring (usually ranging from “hot red” to “cold blue”) show how linguistically similar (red) or dissimilar (blue) an area is with regard to a fixed reference point (see examples for maps on Romance languages on the website, www.dialectometry.com and cf. Goebel 2010 and Goebel in this volume, Chapter 7). Besides the classical similarity maps, Lameli (2013) uses choropleth maps in order to show the degree of standard deviation with respect to a reference point (centrally displayed maps in the map series in Lameli 2013, pp. 129–180).

We point out that the areal patterns depicted by choropleth maps are, in fact, based on point-related data. Each polygon is a mathematical projection of one location. While in the densely populated middle European territories like France and Germany these projections constitute areal patterns with a complete coverage of the territory (cf. Goebel 2010, map 2205; Nerbonne 2010, map 2405), in the MDS-based maps of linguistic distances between Swedish dialects (Leinonen 2010, pp. 129–150, maps 7.2, 7.3 etc.) the punctual character of the data becomes evident, where the polygons are replaced by colored circles around few inhabited locations in northern Sweden.

Quantitative area maps seem to be—together with other quantitatively based maps such as network maps (see above, subsection 20.2.2)—the most fertile area for the development of dialect maps at the moment. This approach is characterized by the application of statistics and modern GIS technology on dialect data. It enables the semi-automatic identification of dialect areas or, at least, the reliable recognition of overall differences and similarities among the dialects of a given area. Notable is the application of these technologies on the massive dialect data collected for huge national-atlas surveys whose results can hardly be rivaled with respect to either quality (e.g., base dialects that today have died out) or quantity (number of survey locations, up to more than 40,000 in Wenker’s *Sprachatlas des deutschen Reichs*, cf. DiWA). These new mapping techniques have been successfully applied to old data of German (Lameli 2013, data from Wenker’s *Sprachatlas des deutschen Reichs* [collected 1876–1887], Nerbonne 2010, data from the *Phonetischer Atlas Deutschlands* [collected 1965–1991]), French (cf. Goebel 2010, data from the ALF [collected 1897–1900]), English (cf. Nerbonne 2015, data from the LAMSAS [collected 1933–1974]), and so on.

20.2.4 Surface Maps

In cartography “surfaces” are different from “areas” in that surface maps are used for the representation of three-dimensional elements, originally those of the Earth’s surface including altitude differences (cf. Unwin 1981, pp. 21–23 for the distinction between area and surface maps). Figure 20.7 (Wattel and van Reenen 2010, map 2502) shows the spelling variants in fourteenth-century Middle Dutch charters. The map shows a complete surface extrapolated from very unevenly distributed empirical data (e.g., many charters around Maastricht and only a few around Antwerp; cf. Wattel and van Reenen 2010, map 2501). Wattel and van Reenen’s algorithm copes with the uneven distribution of the data and infers transition areas between the core areas of the *o* and *a* variants (different colors in the original represent

different percentages of *o* and *a* spellings). “By making a distinction between the value of a dialect observation [...] and the weight of the dialect observation [...], the extrapolation procedure calculates the influence of the value on its surroundings by diminishing the weight over distance” (Wattel and van Reenen 2010, p. 499). Wieling (2012) uses surface maps for the visualization of pronunciation differences between standard Dutch and Dutch dialects (p. 90, fig. 6.1), standard Catalan and Catalan dialects (p. 119, fig. 7.2), and lexical differences between standard Italian and Tuscan dialects (p. 133, fig. 8.2).

20.3 Purposes of Dialect Maps

Dialect maps can be used, in principle, in five different disciplines that shall be illustrated in this section:

- dialectology in a narrow sense, that is, visualization of dialectological facts (20.3.1);
- cultural history, that is, investigation of historical and cultural facts (20.3.2);
- sociolinguistics (20.3.3);
- historical linguistics (20.3.4);
- language theory (20.3.5).

20.3.1 *Dialectology*

The most basic objective of dialect maps is the visualization of the spatial distribution of linguistic features or feature-based areal structures. Maps that pursue this objective might be elaborate but they simply show the areal picture and leave it up to the map reader to draw further conclusions. Note that dialect-area maps, the result of aggregate mappings (including MDS plots and neighbor nets), also belong in this field.

20.3.2 *Cultural History*

Dialect maps can be used to investigate and present historical and cultural facts, for example, historical territories, ecclesiastical areas, migration movements, trading relations (e.g., Hanseatic league) and other communication contacts that leave linguistic traces. This kind of dialect geography, aimed at the identification of historic-cultural areas, has particularly been developed in Germany in the first half of the twentieth century in the so-called “Rhenish school” (cf. Aubin, Frings and Müller 1926) with its opus magnum *Atlas der deutschen Volkskunde* (ADV) (cf. Cox and Zender 1998), but was also common in Romance, American, and Scandinavian dialectology (e.g., AIS [see Fig. 20.4], Kurath 1940, Bandle 2002). Using a similarity map, Lameli (2013, pp. 106–108, fig. 5–10) shows the historical east-middle German origins of mine-worker settlements in northern Germany (the Goslar district is a language island whose linguistic features are most similar to some distant east-middle German districts, even a century after the mine-worker families migrated). In the *Atlas of Languages of Intercultural Communication in the Pacific, Asia and the Americas* several directed line maps show the development of Pidgin and Creole English varieties connected to the plantation network in the Pacific (Wurm *et al.* 1996, maps 47, 59, etc.).

20.3.3 *Sociolinguistics*

In the introduction we promised that the term “dialect” shall be used here only for areal variation. However, in order to properly understand the nature of language variation, both the geographical and the sociological perspectives are necessary. Consequently, social

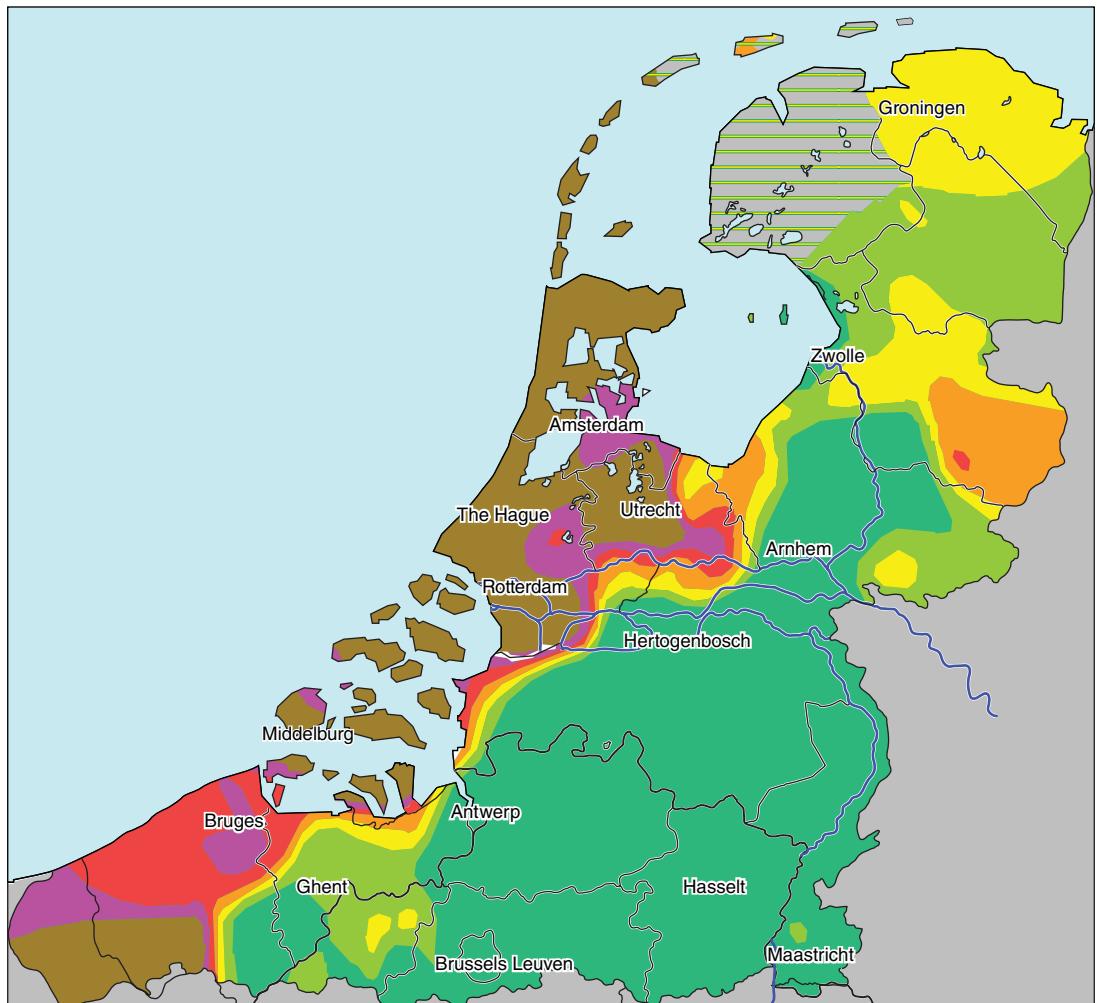


Figure 20.7 Surface map/shaded isarithms (spelling variants *o* [along the north-western coastline including Middelburg, Rotterdam, The Hague, brown in the original color version of the map] and *a* [in the triangle roughly between Arnhem, Brussels, and Maastricht, green in the original] in *af* “off” in fourteenth-century Middle Dutch charters (Wattel and van Reenen 2010, map 2502 [redrawn in grayscale]).

variation is recorded in many dialect maps and dialect atlases. One example is the *Mittelrheinischer Sprachatlas* (MRhSA) in which data was collected and mapped for two age groups: one map for the NORMs (“non-mobile older rural males”) and another for the younger informants, with age-related differences shown by different (red) colors of the symbols at the same locations.⁵ Another example is the *Atlas lingüístico Diatópico y Diastrático del Uruguay* (ADDU), which even acknowledges sociolinguistics in its title (*diastrático* “related to social class”) and depicts the data for four social groups (according to age and formal education). In some cases, diaphasic or stylistic variation within social groups is also depicted (e.g., ADDU, maps 16/17, cf. Thun 2010, p. 519), giving rise to what is called “Pluridimensional Cartography” (Thun 2010).

20.3.4 Historical Linguistics

Dialect maps are an important tool for historical linguistics and the study of language change. The importance of dialects for historical linguistics arises from two important properties:

- a. Different dialects preserve different stages in the development of the language.
- b. Dialects change rather freely as they are less inhibited by normative grammar. They thus enable the detection of “natural” directions and regularities of language change.

Note that the use of dialect maps in historical linguistics is different from the function of a historical reference system as *tertium comparationis* for the map. The use of Proto-Slavic roots as variables in the “Pan-Slavic Linguistic Atlas” (OLA) or of Middle High German vowels in Wenker’s *Sprachatlas des deutschen Reichs* (cf. DiWA) does not necessarily mean that all mapped variants have been derived from the (sometimes hypothetical, unattested) forms of the reference system. The reference system has a purely methodical function. But dialect maps can be used in many ways in historical linguistics. We discuss the possibilities, arranged in four categories.

- i. Single-dialect maps depict single features in “frozen images” of language history, that is, “snapshots” of the past including what has disappeared in the varieties now considered proper or standard. For American English, Kurath (1938, p. 6, quoted from Kehrein 2012, p. 482) claims that the dialect documentation in the *Linguistic Atlas of New England* will “make possible the reconstruction, in its main features, of the linguistic structure of New England before the industrial era [...], so that ultimately the British sources of New England speech can be determined.”⁶

Single-dialect maps may also visualize the course of language-change in the past. A famous example is the High German consonant shift, that is, the transformations of the Germanic plosives /p, t, k/ into the German fricatives /f, s, ch/ in certain phonetic environments (e.g., Gmc. *water* > High German *Wasser*) and into the affricates /pf, ts, kch/ in other positions (Gmc. *tid* “time” > High German *Zeit*, /tsait/). Aggregate dialect maps of the phenomenon (e.g., König 2011, pp. 230–231) show that the shift of the complete set of plosives in (almost) all phonetic environments occurred only in the very south of the German-speaking area: the more northern a location, the fewer the number of shifted plosive types at the location. This synchronic areal structure can be diachronically interpreted, in the sense that the consonant shift originated in the south and then proceeded toward north, losing force on the way up and coming to a final stop at the famous Benrather line in the north-west of Germany.

- ii. The configuration of feature-based areas gives hints about the relative age of current forms and the probable direction of language change. This idea was elaborated first by the Italian “neolinguists” whose major contribution to dialect geography were “areal norms” (Bertoni and Bartoli 1928). These norms predict, for example, that, given two linguistic forms, the form found in isolated or peripheral areas is older than the form found in central areas or areas more accessible for communication. While this is undoubtedly true for Cimbrian dialects, where isolated and peripheral dialects preserve Old High German forms, Trudgill (1975b) pointed out in the 1970s that these norms quickly became discredited because there were too many exceptions to the “laws” of the neolinguists. However, Trudgill (1975b, p. 236) also concedes that “the principles have some validity if they are regarded as guidelines rather than as ‘laws.’” And although there is skepticism toward the characterization of (geo)graphical configurations as “diffusion rings”, “diffusion fans”, “tubes”, and so on (cf. Girnth 2010, pp. 112–116) in terms of language-change directions and types—based solely on synchronic data—a certain tradition of this kind continues.
- iii. More possibilities are opened up by going beyond single dialect maps and considering series of maps. This study can be done within one and the same linguistic atlas. Kehrein (2012, p. 485) quotes several examples in which “dialect differences between speakers from two or more generations [are interpreted] as dialect change in apparent time.” In the MRhSA (Section 20.3.3) the intergenerational differences between NORM speakers and younger-generation speakers displayed by map pairs are projected on the time line and interpreted as language change.
- iv. Obviously, such “apparent-time” analyses (see iii above) have to be cautiously interpreted, since the differences may be less temporal than social: when speakers age they might recover their parents’ speech which they had only temporally lost because of the requirements in their professional life. Real-time analyses with dialect maps that display data collected at different points in time are preferable. They are possible for well documented and repeatedly mapped dialect areas. The *Atlas of North American English* (ANAE) features maps in which the extension of dialect areas is depicted: in map 9.4 (ANAE, p. 66, *The Development of the Low Back Merger From the 1930s to the End of the Century*) three different isoglosses compare data from the 1930s/1940s (Kurath and McDavid 1961), data from a telephone survey carried out by Labov in 1966 and ANAE data from the 1990s. In the ANAE, a small selection of the real-time analyses are shown on static maps.

In contrast, the *Digitaler Wenker-Atlas* (DiWA) is designed to enable the reader to freely compare all maps available in the database according to his/her individual research interest. The main purpose of the DiWA, with regard to language-change studies, is the creation of a collection of digital editions of maps of German dialects from the middle of the nineteenth century to the present, which is freely accessible via Internet and in which the comparison of maps is facilitated by GIS technology. The maps are digitized (scanned, transformed into an image file) and georeferenced, that is, geographic coordinates are assigned to each pixel of the map image, so that two map images get positioned over each other correctly if they are opened in the same browser window. Then, one or more maps can be superimposed, rendered transparent, and thus, compared quite easily. For details on the DiWA project, which has a wide range of uses, refer to, Rabanus, Kehrein, and Lameli (2010). Using DiWA technology, Rabanus (2008) investigated the development of morphological subsystems of 2,427 High German local dialects over a span of 100 years in order to detect principles of morphological change—a number of dialects that would have been impossible to compare without DiWA technology.⁷

20.3.5 Language Theory

The utility of dialect maps for language theory is emerging in recent years, as theoretical linguists become more interested in language variation and, hence, also in dialect data. Kortmann (2010, p. 838) points out that the inspiration is reciprocal; he notes the “syntactic turn” in dialectology on the one hand, and the “dialect turn” in modern linguistic theorizing on the other. To better understand general principles of language the areal picture is interpreted as a frame that delimits the structural possibilities of the language in question, allowing, thus, the definition of general properties of the language the dialect belongs to or, sometimes, of language in general. In syntactic theory dialect variation is considered under the headings of “Syntactic Microvariation” (cf. Barbiers, Cornips, and van der Kleij [eds.] 2002) or “micro-comparative syntax” (Kayne 2013, p. 137). Some very recent papers also argue with dialect maps. For example, Weiß (2013, pp. 189–190) derives a development cycle for the evolution of null pronouns from weak pronouns that emerge from strong pronouns (modified version of the “null subject cycle” postulated by Fuß and Wratislav [2013]) from observations of syntactic variation in German dialects. The important point here is that in Weiß’s article dialect maps are used to refine a general principle of language change. Another example is the above (subsection 20.3.4) quoted study by Rabanus (2008) in which, on the basis of a comparison of dialect maps, a minimal threshold of necessary morphological distinctions in German is identified (“Morphologisches Minimum”), a threshold that resists the general tendency to increase (nominal) case and (verbal) number/person syncretisms (see a short English summary of main results in Schmidt 2010, pp. 210–211).

20.4 Future Perspectives, Challenges, and Risks

In the previous sections the field of dialect maps has been proven to have both a long history and a very vivid present. The blurring of traditional established boundaries between dialectology on the one hand and typology, sociolinguistics, language theory and cognitive linguistics, discourse analysis, and pragmatics on the other opens up many new applications for dialect and language maps. Computerization also gives important impulses to the future development of the field. There are, however, some old questions and problems that continue to be relevant, and some new risks that are tightly connected to today’s enhanced mapping possibilities. Some of the most important issues shall be discussed in the final subsections.

20.4.1 Isogloss

One of the key concepts in classical dialect mapping is the “isogloss.” Isoglosses—the term “isogloss” was coined by Bielenstein (1892)⁸ (cf. Händler and Wiegand 1982, pp. 502–507)—are “lines that separate linguistic phenomena with different features and connect linguistic phenomena with the same features” (Girneth 2010, p. 112).⁹ While isoglosses are a good means for visualizing areal structure (see subsections 20.2.2, 20.2.3), they are problematic with respect to both conceptual content and practical application. Conceptually, they suggest clear borders where, in most cases, continuous transitions exist. In traditional dialectology, the map draftsman sketched the isogloss according the data, but also using his/her intuition and, probably, overall knowledge of the dialect areas. The course of the isogloss often cannot be based solely on the data that is depicted on the map in question only, as illustrated in Figure 20.8, which is an extract of the lexical map *pail* (symbolized by open circles) versus *bucket* (black dots) from *American Regional Dialects* (Carver 1987, p. 11). Note the locations with both forms, and note how the isogloss runs south of the locations marked as “1” and “2” but north of “3.”

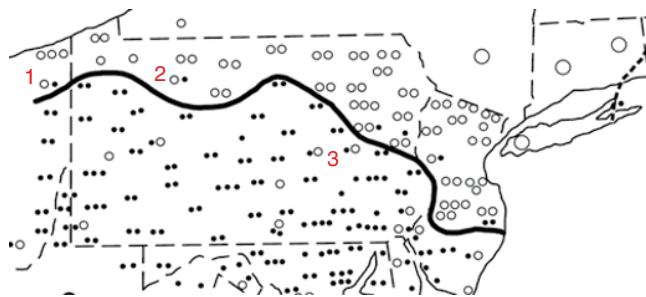


Figure 20.8 Point-symbol map with isogloss (*pail* [open circles] versus *bucket* [dots] in Pennsylvania (Carver 1987, p. 11 [detail; numbers not in the original]).

20.4.2 *Absence of Knowledge and Absence of Phenomenon*

Another problem, related to the drawing of isoglosses and the identification of areas, can be subsumed under the headings “absence of knowledge” and “absence of phenomenon” (cf. Ormeling 2010, p. 24). In subsection 20.2.3 we noted that while a value for soil quality can be assigned to every point of the Earth’s surface, there are many points without linguistic values—either because the data haven’t been collected (absence of knowledge) or because the points are uninhabited and, hence, there is no data to collect (absence of phenomenon). This is not a problem for point-related maps in which only attested values are signed. But for area-related maps there is the question of where draw isoglosses with respect to, for example, sporadically inhabited or completely uninhabited territories or, in other words, to which area a territory without speakers (absence of phenomenon) belongs to. Sometimes draftsmen try to explicitly mark large sporadically inhabited territories, as is the case, for example, in the maps of the languages of the Russian Federation on *Ethnologue*. However, while the maps probably show correctly the sporadically inhabited areas of, for example, the Siberian Republic of Yakutia, they fail to depict the same for the only slightly more densely populated Republic of Karelia, situated along the Finnish border. Uninhabited areas also exist in central Europe. Additionally, dialect surveys never include all relevant locations (absence of knowledge). The issue is particularly problematic in mapping earlier language history for which missing records cannot be recovered. Therefore, it is not a coincidence that probabilistic maps are used primarily in historical dialectology. In a *probabilistic map* a linguistic value is assigned to every mathematical point in the area by an extrapolation procedure, whether or not there are real empirical data. Probabilistic maps usually come in form of *surface maps*, see subsection 20.2.4 and Figure 20.7.

20.4.3 *Pitfalls of Symbolization*

Dialect maps are thematic maps where linguistic information is visualized by graphic symbols that can be grouped into the four classes points, lines, areas, and surfaces introduced in Section 20.2. These symbols have the attributes of size, brightness, form, and color (cf. Bollmann 2010, p. 52), which are chosen and combined by the cartographer in order to visualize the relevant linguistic distinctions as clearly as possible (for the first theoretical approach to the impact of graphic variables on the perception of information, cf. Bertin 1967). However, the proper choice of the attributes is one of the most complex problems in cartography since the design of the symbols may have undesired side effects. The map *auch* “also” from the *Sprachatlas der deutschen Schweiz* (SDS, vol. IV, map 151)¹⁰ is, in principle, a good example for a properly constructed point-symbol map. However, the

symbols used in this map differ in prominence because of their different degrees of darkness or “blackness.” For example, the small number of thickly printed triangles in the north-western angle of Switzerland (south of Basel), surrounded by other dark and prominent symbols, stand out and may lead readers to misconstrue the geographical center of the triangle-symbolized feature in that area. In actuality, there are many more triangles in central Switzerland and in the canton of Valais. But they are less visible since they are printed in light gray.

The map designer’s choices are also crucial when it comes to the colors in choropleth maps. Whereas, on the one hand, the different degrees of coloring are calculated by program (e.g., by GIS programs such as ArcGIS), the choice of the colors at both ends of the color scale is up to the map designer. This choice has an effect on the interpretation of the map. Colors are associated with human perceptions, for example, red is normally associated with warmth, blue with cold. These associations may be consciously used. In the Salzburg-style similarity maps, red color is applied for expressing the maximum of similarity, blue color stands for the maximum of difference (cf. Goebl 2010, p. 448). However, in other maps these associations may lead to interpretations that were not intended by the map designer. The pureness and brightness of color, enabled by modern print and digital reproduction technologies, is another pitfall. Nerbonne (2010, p. 491, referring to Imhof 1965, p. 83) points out that pure colors may have confusing effects and proposes shaded colors in order to better visualize distinctions.¹¹

A general problem is that color differences, carefully chosen and realized on the computer, disappear when the map is printed or photocopied on ordinary photocopying machines (or in this handbook), making the symbols indistinguishable. Full-color print and photocopy are not yet standard. Hence, it is always a good choice to design maps with black-and-white or gray differences, as long as the map type allows this. However, some map types work only with colors.

20.4.4 Internet

The internet offers a wide range of possibilities for the publication of dialect maps. The internet publication of dialect maps can also enhance maps by providing access to additional data such as audio/video files, written records of the survey data, or bibliographic references. In the georeferenced maps of DiWA (see Section 20.3.4, [iv]) a mouse click on the location provides a menu that lists the data that are available for the location in question. The internet also enables the publication of map series in which one map is superimposed upon another producing a sequence of images without any immediately obvious breaks (like a video clip). This possibility is especially useful for the representation of processes in historical linguistics, as illustrated by Hanewinkel and Losang (2010, pp. 427–431) on the example of the spread of the Austronesian languages.¹²

The internet’s speed and flexibility enhance the possibilities for map drawing. However, they may also have a disadvantage whose impact cannot be estimated at the moment: the long-term storage problem. Printed dialect maps are usually distributed over many different libraries and archives. Usually there are—in spite of neglect or even destruction of archive material—still some readable copies even hundreds of years after publication, whereas many of us have already experienced personally the loss of electronic data from a mere decade ago because they were saved on media that are no longer compatible with current standards. Long-term storage of electronic data requires long-term strategies, financing, and continuous care. Hence, it is one of the greatest challenges for information technology to enable a fast and flexible access to today’s data without losing yesterday’s knowledge.

NOTES

- 1 This position is exemplified by the fact that in De Gruyter's HSK series the two-volume handbook *Dialektologie* (1982/1983) was not updated but substituted by the new HSK subseries *Language and Space: An International Handbook of Linguistic Variation*, which includes, besides a core formed by topics from traditional dialectology, also sociolinguistic and typological issues (first vols. edited 2010 by Auer and Schmidt [Theories and Methods] and Lameli, Kehrein and Rabanus [Language Mapping]).
- 2 For some of the maps quoted in this article, digital versions are available in the REDE database (www.regionalsprache.de) and direct links to these digital versions will be indicated. The direct link to Bernhardi's map is: www.regionalsprache.de/Map/t8zx4UOx.
- 3 Actually, the Groningen software supports different sorts of restrictions on the lengths of "beams" in network maps, see "statistics and difference maps" at www.gabmap.nl (demos).
- 4 The black circles are proportional symbols that represent the population of the locations in the various districts.
- 5 See, for example, the digitally re-elaborated version of MRhSA, vol. 3, map 220, *Hunde* "dogs" (accented vowel) at www.regionalsprache.de/Map/xW0QkjXT. Age-related contrasts are marked with red color.
- 6 Shackleton (2010) is a detailed quantitative survey of the sources of American speech in the English dialects.
- 7 DiWA has since been integrated in the new Geolinguistic Information System "Regionalsprache.de" (REDE: www.regionalsprache.de), cf. Kehrein (2012, pp. 493–494).
- 8 Bielenstein's isogloss map of Latvia is reproduced in Händler and Wiegand (1982, p. 505) and Veith (2006a, p. 520).
- 9 Goebel (2004, p. 526) notes the double meaning of "isogloss" as cartographic means according to the quoted definition on the one hand, and as features shared by two or more languages.
- 10 Direct link to the map: www.regionalsprache.de/Map/VQwkLAVV.
- 11 Colors that are derived form a three-dimensional color space, for example, in MDS-based maps, are usually shaded, cf. maps 2405 and 2406 in Nerbonne (2010) and fig. 7–5 in Lameli (2013, p. 197).
- 12 The animated map is not available any more at the URL indicated by Hanewinkel and Losang (2010: www.map-service.de/austronesian) by today (20 november 2016): a good example for the long-term storage problem discussed in the final paragraph of subsection 20.4.4.

REFERENCES

- ADDU = Thun, Harald and Elizaincín, Adolfo 2000. *Atlas lingüístico Diatópico y Diastrático del Uruguay*. Vol. I. Kiel: Westensee-Verlag.
- ADV = Harmjanz, Heinrich and Röhr, Erich. (eds.) 1937–1939. *Atlas der deutschen Volkskunde*. Berlin: De Gruyter/Leipzig: Hirzel.
- AED = Upton, Clive and Widdowson, John 2006. *An Atlas of English Dialects*. 2nd ed. Oxford: Oxford University Press.
- AIS = Jaberg, Karl and Jud, Jakob 1928–1940. *Sprach- und Sachatlas Italiens und der Südschweiz*. 8 vols. Zofingen: Ringier.
- ALF = Gilliéron, Jules and Edmont, Edmond 1902–1910. *Atlas linguistique de la France*. 9 vols. Paris: Champion.
- Ammon, Ulrich et al. (eds.) 2010. *Sociolinguistics*. 2nd ed. Vol. 1. (HSK 3.1.) Berlin/New York: De Gruyter.
- ANAE = Labov, William, Ash, Sharon, and Boberg, Charles. 2006. *The Atlas of North American English: Phonetics, Phonology and Sound Change*. Berlin/New York: Mouton de Gruyter.
- Aubin, Hermann, Frings, Theodor, and Müller, Josef 1926. *Kulturströmungen und Kulturprovinzen in den Rheinlanden. Geschichte, Sprache, Volkskunde*. Bonn: Röhrscheid.
- Auer, Peter and Schmidt, Jürgen Erich (eds.) 2010. *Language and Space*. Vol. 1: *Theories and Methods*. (HSK 30.1.) Berlin/New York: De Gruyter Mouton.

- Bandle, Oscar 2002. Nordic language history and cultural geography. In: Bandle, Oscar *et al.* (eds.): *The Nordic languages*. Vol. 1. (HSK 22.1.). Berlin/New York: De Gruyter, 338–343.
- Barbiers, Sjef, Cornips, Leonie, and van der Kleij, Susanne (eds.) 2008. *Syntactic Microvariation*. Amsterdam: Meertens Institute Electronic Publications in Linguistics (MIEPiL). Vol. II. URL: www.meertens.knaw.nl/books/synmic/.
- Bernhardi, Karl 1843. *Sprachkarte von Deutschland*. Kassel: Bohné.
- Bertin, Jacques 1967. *Semiologie graphique. Les diagrammes, les réseaux, les cartes*. Paris: Gauthier-Villars.
- Bertoni, Giulio and Bartoli, Matteo G. 1928. *Breviario di neolinguistica*. Modena: Società Tipografica modenese.
- Bielenstein, Johann 1892. *Die Grenzen des lettischen Volksstammes und der lettischen Sprache in der Gegenwart und im 13. Jahrhundert. Ein Beitrag zur ethnologischen Geographie und Geschichte Rußlands*. Saint Petersburg: Eggers.
- Bollmann, Jürgen 2010. Maps and cognition. In: Lameli, Kehrein and Rabanus (eds.), 40–69.
- Britain, David 2004. Dialect and Accent. In: Ammon *et al.* (eds.), 267–273.
- Brown, Keith *et al.* (eds.) 2006. *Encyclopedia of Language and Linguistics*. 2nd ed. Vol. 3. Oxford: Elsevier.
- Carver, Craig M. 1987. *American Regional Dialects. A Word Geography*. Ann Arbor, MI: University of Michigan Press.
- Chambers, J.K. and Trudgill, Peter 1998. *Dialectology*. 2nd ed. Cambridge: Cambridge University Press.
- Cox, Heinrich L. and Zender, Matthias 1998. Sprachgeschichte, Kulturräumforschung und Volkskunde. In: Besch, Werner *et al.* (eds.): *Sprachgeschichte*. 2nd ed. Vol. 1. (HSK 2.1.). Berlin/New York: De Gruyter, 160–172.
- Dell'Aquila, Vittorio 2010. GIS and sociolinguistics. In: Lameli, Kehrein and Rabanus (eds.), 458–476.
- DiWA = Schmidt, Jürgen Erich and Herrgen, Joachim (eds.) 2001–2009. *Digitaler Wenker-Atlas*. Compiled and prepared by Alfred Lameli, Tanja Giessler, Roland Kehrein, Alexandra Lenz, Karl-Heinz Müller, Jost Nickel, Christoph Purschke and Stefan Rabanus. Marburg: Forschungszentrum Deutscher Sprachatlas. URL: www.diwa.info.
- Durrell, Martin 2004. Sociolect. In: Ammon *et al.* (eds.), 200–205.
- Ellis, Alexander J. 1889. *Existing Phonology of English Dialects. Compared with that of West Saxon Speech: Forming Part V of "Early English Pronunciation"*. London: Trübner and Co.
- Fuß, Eric and Wratil, Melani 2013. Der Nullsubjektzyklus: Etablierung und Verlust von Nullargumenten. In: Fleischer, Jürg and Simon, Horst (eds.): *Sprachwandelvergleich – Comparing Diachronies*. Tübingen: Niemeyer, 163–196.
- Girnth, Heiko 2010. Mapping language data. In: Lameli, Kehrein and Rabanus (eds.), 98–121.
- Goebl, Hans 2010. Dialectometry and quantitative mapping. In: Lameli, Kehrein and Rabanus (eds.), 433–457.
- Goebl, Hans 2004. Eine Glosse zur Isoglosse. In: Krisch, Thomas *et al.* (eds.): *Analecta homini universalis dicata. Festschrift für Oswald Panagl zum 65. Geburtstag*. Stuttgart: Verlag Hans-Dieter Heinz, 527–537.
- Güldemann, Tom 2010. Sprachraum and geography: Linguistic macro-areas in Africa. In: Lameli, Kehrein and Rabanus (eds.), 561–585.
- Händler, Harald and Wiegand, Herbert Ernst 1982. Das Konzept der Isoglosse: methodische und terminologische Probleme. In: Besch, Werner *et al.* (eds.): *Dialektologie*. Vol. 1. (HSK 1.1.). Berlin/New York: De Gruyter, 501–527.
- Hanewinkel, Christian and Losang, Erik 2010. Animated maps. In: Lameli, Kehrein and Rabanus (eds.), 415–433.
- HSS = Kleiber, Wolfgang, Kunze, Konrad, and Löffler, Heinrich 1979. *Historischer Südwestdeutscher Sprachatlas. Aufgrund von Urbaren des 13. bis 15. Jahrhunderts*. 2 vols. Bern/Munich: Francke.
- Imhof, Eduard. 1965. *Kartographische Geländedarstellung*. The Hague: Walter de Gruyter.
- Kayne, Richard S. 2013. Comparative syntax. In: *Lingua* 130, 132–115.
- Kehrein, Roland 2012. Linguistic Atlases: Empirical Evidence for Dialect Change in the History of Languages. In: Hernández-Campoy, Juan Manuel and Conde Silvestre, Juan Camilo (eds.): *The Handbook of Historical Sociolinguistics*. Oxford: Wiley-Blackwell, 480–500.
- König, Werner 2011. *dtv-Atlas Deutsche Sprache*. 17th ed. Munich: Deutscher Taschenbuch Verlag.
- Kortmann, Bernd 2010. Areal variation in syntax. In: Auer and Schmidt (eds.), 837–864.
- Kurath, Hans 1938. *Linguistic Atlas of New England. Prospect*. Providence: Brown

- University for the American Council of Learned Societies.
- Kurath, Hans 1940. Dialect areas, settlement areas, and cultural areas in the United States. In: Ware, Caroline F. (ed.): *The cultural approach to history*. New York: Macmillan, 331–345.
- Kurath, Hans and McDavid, Raven I. 1961. *The Pronunciation of English in the Atlantic States*. Ann Arbor: University of Michigan Press.
- Lameli, Alfred 2010. Linguistic atlases – traditional and modern. In: Auer and Schmidt (eds.), 567–592.
- Lameli, Alfred 2013. *Strukturen im Sprachraum. Analysen zur arealtypologischen Komplexität der Dialekte in Deutschland*. Berlin/Boston: De Gruyter.
- Lameli, Alfred, Kehrein, Roland, and Rabanus, Stefan (eds.) 2010. *Language and Space*. Vol. 2: *Language Mapping*. (HSK 30.2.) Berlin/New York: De Gruyter Mouton.
- LAMSAS = McDavid, Raven I. et al. 1980. *Linguistic Atlas of the Middle and South Atlantic States*. Chicago: University of Chicago Press.
- Leinonen, Therese 2010. *An Acoustic Analysis of Vowel Pronunciation in Swedish Dialects*. PhD thesis, University of Groningen. Available at: www.rug.nl/research/portal/.
- MRhSA = Bellmann, Günter, Herrgen, Joachim, and Schmidt, Jürgen Erich 1994–2002. *Mittelrheinischer Sprachatlas*. 5 vols. Tübingen: Niemeyer.
- Nerbonne, John 2010. Mapping aggregate variation. In: Lameli, Kehrein and Rabanus (eds.), 476–495.
- Nerbonne, John 2015. Various Variation Aggregates in the LAMSAS South. In: Picone, Michael D. and Evans Davies, Catherine (eds.): *Language Variety in the South: Historical and Contemporary Perspectives*. Tuscaloosa: University of Alabama Press, 723–753.
- OKDA = *Общекарпатский диалектологический атлас* [Carpathian Dialect Atlas] 1987–2003. 8 vols., various editors and publishing houses, see Lameli, Kehrein and Rabanus 2010, vol. 2, 426–427 for a complete reference list.
- OLA = *Общеславянский лингвистический атлас* [Pan-Slavic Linguistic Atlas] 1965–2008. Many vols., eleven national committees, various editors and publishing houses; see Lameli, Kehrein and Rabanus 2010, vol. 2, 427–429 for a complete reference list.
- Ormeling, Ferjan 2010. Visualizing geographic space: The nature of maps. In: Lameli, Kehrein and Rabanus (eds.), 21–40.
- Preston, Dennis R. 2010. Mapping the geolinguistic spaces of the brain. In: Lameli, Kehrein and Rabanus (eds.), 121–141.
- Rabanus, Stefan 2008. *Morphologisches Minimum. Distinktionen und Synkretismen im Minimalatz hochdeutscher Dialekte*. Stuttgart: Steiner.
- Rabanus, Stefan 2011. The State of the Art in Linguistic Cartography. In: Reinhämmar, Maj, Edlund, Lars-Erik, and Elmevik, Lennart (eds.): *Studier i dialektologi och sociolinguistik*. Föredrag vid den Nionde nordiska dialektologkonferensen, Uppsala 18–20 augusti 2010. Uppsala: Kungl. Gustav Adolfs Akademien för svensk folkkultur, 31–52.
- Rabanus, Stefan 2012. Mapping techniques. In: Kortmann, Bernd (ed.): *Theories and Methods in Linguistics*. Berlin/New York: De Gruyter. DOI: 10.1515/wsk.35.0.mappingtechniques.
- Rabanus, Stefan, Kehrein, Roland, and Lameli, Alfred 2010. Creating digital editions of historical maps. In: Lameli, Kehrein and Rabanus (eds.), 375–385.
- Sargsyan 2008 = Սարգսյան, Ա. [Sargsyan, A.] 2008. Հայոց լեզվի բարբառային ասլաւ: Դասընթաց բանափրական ֆակուլտետների ուսանողության և ասպիրանտների համար: Պրակտ. Ա [Dialect Atlas of the Armenian Language. A Course for Students and PhD Students of the Humanities. Part A]. Երևան (հեղինակային հրատարակություն) [Yerevan (self published)].
- SDS = Hotzenköcherle, Rudolf et al. (1962–1997): *Sprachatlas der deutschen Schweiz*. 8 vols. Bern/Basel: Francke.
- Schirmunski 1956 = Жирмунский, В.М. [Жирмунский, В.М.] 1956. *Немецкая диалектология* [German Dialectology]. Москва / Ленинград: Издательство академии наук СССР [Moscow / Leningrad: Publisher of the Academy of Sciences of the USSR].
- Schmidt, Jürgen Erich 2010. Language and space: The linguistic dynamics approach. In: Auer and Schmidt (eds.), 201–225.
- Shackleton Jr., Robert George 2010. *Quantitative assessment of English-American speech relationships*. PhD thesis, University of Groningen. Available at: www.rug.nl/research/portal/.
- Thun, Harald 2010. Pluridimensional Cartography. In: Lameli, Kehrein and Rabanus (eds.), 506–524.
- TNZN = Grootaers, Ludovic and Kloekie, Gesinus G. 1939–1972: *Taalatlas van Noord- en Zuid-Nederland*. 9 fascs. Leiden: Brill.
- Trudgill, Peter 1975a. *Sociolinguistics: An Introduction*. Reprint. Harmondsworth: Penguin.

- Trudgill, Peter 1975b. Linguistic geography and geographical linguistics. In: *Progress in Geography* 7, 227–252.
- Unwin, David 1981. *Introductory spatial analysis*. London/New York: Methuen.
- Upton, Clive 2010. Designing maps for non-linguists. In: Lameli, Kehrein and Rabanus (eds.), 142–157.
- Veith, Werner H. 2006a. Dialect atlases. In: Brown *et al.* (eds.), 517–528.
- Veith, Werner H. 2006b. Dialects. Early European Studies. In: Brown *et al.* (eds.), 540–560.
- Wattel, Evert and van Reenen, Pieter 2010. Probabilistic maps. In: Lameli, Kehrein and Rabanus (eds.), 495–505.
- Weiß, Helmut 2013. UG und syntaktische (Mikro-)Variation. In: Abraham, Werner and Leiss, Elisabeth (eds.): *Dialektologie in neuem Gewand. Zu Mikro-/Varietätenlinguistik, Sprachenvergleich und Universalgrammatik*. Sonderheft Linguistische Berichte, 171–205.
- Wieling, Martijn 2012. *A quantitative Approach to Social and Geographical Dialect Variation*. PhD thesis, University of Groningen. Available at: www.rug.nl/research/portal/.
- Wurm, Stephen A., Mühlhäusler, Peter, and Tryon, Darrell T. 1996. *Atlas of Languages of Intercultural Communication in the Pacific, Asia, and the Americas*. Berlin/New York: Mouton de Gruyter.

21 Identifying Regional Dialects in On-Line Social Media

JACOB EISENSTEIN

Electronic social media offer new opportunities for informal communication in written language, while at the same time providing new datasets that allow researchers to document dialect variation from records of natural communication among millions of individuals. The unprecedented scale of this data enables the application of quantitative methods to automatically discover the lexical variables that distinguish the language of geographical areas, such as cities. This can be paired with the segmentation of geographical space into dialect regions, within the context of a single joint statistical model—thus simultaneously identifying coherent dialect regions and the words that distinguish them. Finally, a diachronic analysis reveals rapid changes in the geographical distribution of these lexical features, suggesting that statistical analysis of social media may offer new insights on the diffusion of lexical change.

21.1 Dialect in Social Media

Social media comprises a wide range of different internet platforms, including collaborative writing projects such as Wikipedia, on-line communities such as Facebook and MySpace, forums such as Reddit and Stack Exchange, virtual game worlds, business and product reviews, and blogs and microblogs (Boyd and Ellison 2007). These platforms offer a diverse array of ways to interact with friends and strangers—but in the overwhelming majority of cases, interaction is conducted in written language. As social media plays an increasingly ubiquitous part in daily life,¹ writing, therefore, acquires a new social role and a new level of importance. So it is unsurprising that a flowering of diversity in textual styles has been noticed by some observers (Walther and D'Addario 2001; Crystal 2006; Tagliamonte and Denis 2008; Dresner and Herring 2010; Collister 2011; Schnoebelen 2012)—as well as by some critics (Thurlow 2006). Whether this newfound diversity is attributed to the lack of social regulation on social media writing (as compared with other written media), or to the demands of social, near-synchronous communication, it raises natural questions for dialectology: to what extent can geographical variation be observed in social media writing, how does it relate to geographical variation in spoken language, and what are the long-term prospects for this form of variation?

21.1.1 Twitter

Social media services vary in the degree to which large-scale data can be collected, and offer different sorts of geographic metadata. *Twitter* is a social media service that is particularly advantageous on both characteristics. Most messages are publicly readable—this is the default setting—and can therefore be acquired through the streaming Application Programming Interface (API). In addition, many authors choose to include precise geographic coordinates as metadata with each message, although this is not enabled by default. In our data, roughly 1–2% of messages include geographical coordinates. Other methods for geolocating Twitter messages are explored by Dredze *et al.* (2013), but are not considered here.

Twitter is a *microblog* service, and messages (“tweets”) are limited to 140 characters. The basic user interface is shown in Figure 21.1, and some examples of Twitter messages are shown here:

bigdogcoffee Back to normal hours beginning tomorrow.....Monday–Friday 6am–10pm

Sat/Sun 7:30 am–10 pm

crampell Casey B. Mulligan: Assessing the Housing Section <http://nyti.ms/hcUKK9>

THE REAL SHAQ fill in da blank, my new years shaqlution is _____

These messages are chosen from public personae—a small business, a journalist, and a celebrity-athlete—but the majority of the content on Twitter is created by ordinary people. Users construct custom timelines by choosing to *follow* other users, whose messages then appear in the timeline. Unlike Facebook, these social network connections are directed, so that celebrities may have millions of followers (Kwak *et al.* 2010). For this reason, Twitter can be used as a broadcast medium. However, Twitter also enables dialogues in which messages can be “addressed” to another user by beginning the message with a username (Figure 21.1). This motivates another usage scenario, more akin to undirected social networks like Facebook, in which Twitter hosts public conversations

The screenshot shows a Twitter interface. At the top, there's a profile picture of Chuck Grassley, his name, and a blue verified checkmark. To the right, there are options to follow or unfollow him, and a gear icon. Below this, the main tweet is displayed:

Congress shld take action to stop this administration from turning over control of Internet to foreign body so dictators can censor there

Below the tweet, engagement metrics are shown: 73 retweets and 37 likes. The timestamp is 10:24 AM - 29 Mar 2014. Underneath the tweet, there are icons for retweet, reply, favorite, and more. A reply box is open, showing the user's profile picture and the text "Reply to @ChuckGrassley". The user has responded with:

- 29 Mar 2014
@chuckgrassley censor, Senator Grassley. A censor is something that goes off when for example, you post a tweet. I get those signals.

Below this reply, there are icons for retweet, reply, favorite, and more.

Figure 21.1 An example broadcast message followed by a conversational reply, which is addressed to the original author by beginning the message with his username.

(Huberman *et al.* 2008). More than 40% of messages in the dataset described below are addressed to another user.

Representativeness is a major concern with social media data. While social media increasingly reaches across barriers such as age, class, gender, and race, it cannot be said to offer a demographically balanced portrayal of the language community. Fortunately, the demographics of Twitter users in the United States have been surveyed repeatedly by the Pew Internet Research center (Duggan and Smith 2013). Results from late 2013 found that 18% of internet users visit Twitter, with nearly identical rates among men and women. The most recent data is from 2015. 23% of American internet users visit Twitter, with a slightly higher rate for men (25%) than women (21%). Blacks and Hispanics were more likely to visit Twitter than whites (28% to 20%), and young people use Twitter at a much higher rate (32% for ages 18–29, 29% for ages 30–49, 13% for ages 50–64, and 6% for ages 65 and above). Differences across education level and income were not significant, but Twitter is used significantly less often in rural areas. An important caveat is that the *per-message* demographics may be substantially different from these figures, if the usage rate also varies with demographics. Consequently, it is important to remember that quantitative analysis of Twitter text (as with all social media) can describe only a particular demographic segment within any geographical area. This issue could be ameliorated through more fine-grained geographical analysis: for example, US census blocks offer detailed demographic information, and their boundaries are drawn to emphasize demographic homogeneity (Eisenstein *et al.* 2011b). Another approach would be to try to infer the demographic characteristics of individual authors (Argamon *et al.* 2007; Chang *et al.* 2010; Rosenthal and McKeown 2011), and then correct for these characteristics using post-stratification. Such refinements are beyond the scope of this chapter; their integration into social media dialectology must remain a topic for future work.

21.1.2 Related Work

The computer science community has shown great interest in the problem of text-based geolocation: predicting where individuals are from, based on their writings (Cheng *et al.* 2010; Eisenstein *et al.* 2010; Wing and Baldridge 2011; Hong *et al.* 2012). This task can be seen as the converse of the dialectologist's goal of summarizing the linguistic patterns that characterize residents of each geographical area. While predictive methods may yield insights for dialectology, the goals of accurate prediction and comprehensible modeling are not perfectly aligned, and therefore the most useful computational techniques may not necessarily be those which yield the most accurate predictions.

More broadly, several research efforts exploit social media datasets for purposes that touch on issues related to dialectology. Many researchers have attempted to model and predict the spread of on-line “memes” (Leskovec *et al.* 2009; Romero *et al.* 2011), and a related line of work investigates the survival of new words and expressions (Garley and Hockenmaier 2012; Altmann *et al.* 2011). Other researchers have focused on the ways in which such temporal trends are shaped by groups, and on the emergence and evolution of linguistic conventions in on-line communities (Garley and Hockenmaier 2012; Kooti *et al.* 2012; Nguyen and Rosé 2011; Postmes *et al.* 2000; Danescu-Niculescu-Mizil *et al.* 2013b). A further consideration is the role of dyadic social relationships, which may shape linguistic behavior through phenomena such as accommodation (Danescu-Niculescu-Mizil *et al.* 2011), politeness (Danescu-Niculescu-Mizil *et al.* 2013a), power dynamics (Danescu-Niculescu-Mizil *et al.* 2012; Gilbert 2012; Prabhakaran *et al.* 2012), and code-switching (Paolillo 2011). Such research has mainly focused on the linguistic norms of on-line communities, such as forums and chatrooms, rather than on geographical regions in the physical world; it is geographically anchored on-line language variation that constitutes the main focus on this chapter.

21.2 Dataset

The empirical findings in this chapter are based on a dataset that was gathered by Brendan O'Connor from the public “Gardenhose” version of Twitter’s streaming API (Application-Programming Interface), and is first described in a technical report (Eisenstein *et al.* 2012). Within the initial set of messages, only those containing GPS metadata are considered here, so that analysis can be restricted to the United States. The streaming API ostensibly offers a 10% sample of public posts, although Morstatter *et al.* (2013) show that messages containing GPS metadata are sampled at a much higher rate. The dataset was acquired by continuously receiving data from June 2009 to May 2012, and contains a total of 114 million geotagged messages from 2.77 million different user accounts.

Retweets are repetitions of previously posted messages; they were eliminated using both Twitter metadata as well as the “RT” token (a common practice among Twitter users to indicate a retweet). Tweets containing URLs were eliminated in order to remove marketing-oriented messages, which are often automated. Accounts with more than 1,000 followers or followees were removed for similar reasons. All text was downcased and tokenized using the publicly-available Twokenize program (Owoputi *et al.* 2013), and repetitions of more than two characters were normalized to just two characters (e.g., *heyyyyy* → *heyy*). No other textual preprocessing was performed.

21.3 Known Lexical Variables

Before developing quantitative methods for discovering dialect variation in social media, I begin with a simpler question: do regional dialect words from spoken language persist in social media? This investigation will focus on four well-known spoken-language dialect terms.

- *Yinz* is a form of the second-person pronoun, which is associated with the dialect of Southwestern Pennsylvania around the city of Pittsburgh (Johnstone *et al.* 2002). It is exceedingly rare in the Twitter dataset, appearing in only a few hundred messages, out of a total of 100 million. The geographical distribution of these messages is indeed centered on the city of Pittsburgh, as shown in Figure 21.2a.
- *Yall* (also spelled *y'all*) is an alternative form of the second-person pronoun, often associated with the Southeastern United States, as well as African-American English (Green 2002). It is relatively frequent in the dataset, used at a rate of approximately one per 250 messages, making it more than 1,000 times more common than *yinz*. Its geographical distribution, shown in Figure 21.2b, indicates that it is popular in the Southeast, but also in many other parts of the United States.
- *Hella* is an intensifier that is popularly associated with Northern California (Bucholtz *et al.* 2007); it is used in examples such as *i got hella nervous*. Unlike *yinz*, *hella* is fairly common in this dataset, appearing in nearly one out of every 1,000 messages. While the word does appear in Northern California at a higher-than-average rate, Figure 21.2c shows that it is used throughout the country.
- *Jawn* is a noun with Philadelphia origins (Alim 2009) and diffuse semantics:
 1. @name ok u have heard this jawn right
 2. how long u been up in that jawn @name
 3. i did wear that jawn but it was kinda warm this week

Jawn appears at a rate of approximately one per 10,000 messages, with a geographical distribution reflecting its Philadelphia origins (Figure 21.2d). However, there is significant diffusion to nearby areas in New Jersey and New York City.

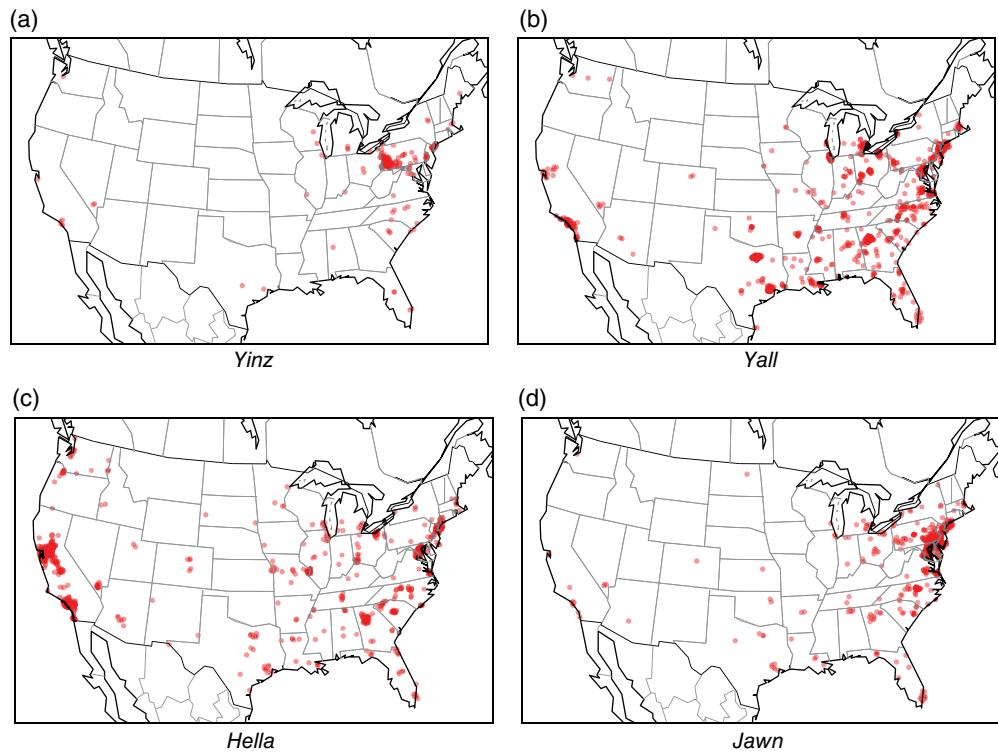


Figure 21.2 Geolocations for messages containing four lexical variables known from previous work on spoken American English. In every case but *yinz*, 1000 randomly selected examples are plotted; for *yinz*, only 535 examples are observed, so all are plotted.

These maps show that existing lexical variables from spoken language do appear in social media text, and they are supported by analysis of other well-known variable pairs, such as *sub/hoagie* and *soda/pop*. However, variables that appear frequently have broad geographical distributions, which only loosely reflect their spoken-language geographical association (in the case of *hella*) or fail to reflect it altogether (*yall*). In the case of these two terms, this may be attributable to their popularity in African-American English, an ethnic dialect that is widely shared across the United States (Eisenstein *et al.* 2011b; Green 2002). Conversely, terms with sharply defined geographical signatures, such as *yinz* and *hoagie*, are exceedingly rare. As noted above, Twitter users are on average younger and less likely to be White; terms such as *yinz* may be rarely used in this demographic. Only *jawn* maintains both a strong geographical signature and widespread popularity, and Alim (2009) traces this term's origins to African-American English.

21.4 Discovering Lexical Variables

The previous section shows that some lexical and phonological variables from spoken language have made the transition to on-line media, and that analysis of such media provides a large-scale but noisy picture of this variation. But the analysis thus far has been unsystematic, focusing on a few hand-chosen examples. A key advantage of big data is that it can speak for itself: we can apply data-driven statistical analysis to find words

with strong geographical associations. This could reveal lexical variables from spoken language that were neglected by our intuition, or might identify variables that are new to electronic media.

Many researchers have examined the problem of automatically identifying words associated with types of documents or groups of authors. Monroe *et al.* (2008) provide a useful overview of relevant statistical techniques in this general area; Eisenstein *et al.* (2010) show how these methods can be applied to dialect analysis; in Chapter 24 of this volume, Grieve describes methods for measuring spatial autocorrelation in linguistic variables. Our approach here will be probabilistic, with the goal of characterizing how the word frequencies change with respect to geography. A simple first-pass approach would be to consider a set of metropolitan areas, and then compute the relative frequency of each term among all messages from within each area. Such frequencies can be computed directly from raw counts, but they are difficult to compare. If we consider the raw *difference* in frequencies, this will overemphasize very common words at the expense of rare ones; for example, an increase of 1% in the frequency of *the* or *and* would be far greater than the total frequency of *yinz* in Pittsburgh. If, on the other hand, we consider the *ratio* of frequencies, this will overemphasize rare words; in the limit, a word that appears just once in the entire dataset will have an infinite ratio of frequencies between two metropolitan areas.

One way to avoid these problems is to reparametrize the probability distribution over words, by applying the *logistic transformation*. This transformation takes the form of a ratio between two non-negative quantities, and the denominator ensures that the function sums to one over all words—thus satisfying the basic requirements of a probability distribution. Specifically, for each region r , we have:

$$P_r(w) = \frac{\exp(m_w + \beta_w^{(r)})}{\sum_i \exp(m_i + \beta_i^{(r)})} \quad (21.1)$$

where $m_w = \hat{\log} P(w)$, the log of the empirical frequency of word w across the entire population (in all regions), and $\beta_w^{(r)}$ is the *deviation* from this empirical log frequency in region r . The numerator exponentiates this sum, ensuring non-negativity, and the denominator sums over all words, ensuring that $\sum_w P_r(w) = 1$. Assuming $\bar{\beta}^{(r)}$ is centered at zero, then a large positive value $\beta_w^{(r)}$ means that word w is substantially more frequent in region r than it is elsewhere; a large negative value means it is less frequent. Sorting by $\beta_w^{(r)}$ proves a very effective way of identifying meaningful words for region r , striking a good balance between uniqueness and frequency. *Comprehensibility* is therefore the first of two advantages for this formulation. Similar observations were made by Monroe *et al.* (2008), in the context of analyzing speech differences between political parties in the United States.

Since \bar{m} is computed directly from the empirical frequencies, the estimation problem is to compute each parameter vector $\bar{\beta}^{(r)}$. We will use a regularized maximum-likelihood criterion, meaning that we choose $\bar{\beta}^{(r)}$ to maximize the log-likelihood of the observed text, $\mathcal{L} = \sum_n \log P(w_n; \bar{\beta}^{(r)}, \bar{m})$, subject to a *regularizer* that penalizes $\bar{\beta}^{(r)}$ for its distance from zero. Regularization (also called shrinkage) reduces the sensitivity of the w parameter estimates to rare terms in the data, by introducing a bias toward uniformity (Murphy 2012). Because of this bias, strong evidence is required before we have $\beta_w^{(r)} \neq 0$; therefore, *robustness* is the second advantage of this formulation. A typical choice for the regularizer is to penalize a norm of the vector $\bar{\beta}^{(r)}$. More details, including an algorithm to estimate $\bar{\beta}^{(r)}$, can be found in a prior publication (Eisenstein *et al.* 2011a). Source code is also freely available.²

This approach can be applied to the dataset described in Section 21.2, with each region corresponding to a *metropolitan statistical area* (MSA). MSAs are defined by the US

government, and include the regional area around a single urban core. Using this approach, the top words for some of the largest MSAs in the United States are:

- **New York:** flatbush, baii, brib, bx, staten, mta, odee, soho, deadass, werd
- **Los Angeles:** pasadena, venice, anaheim, dodger, disneyland, angeles, compton, ucla, dodgers, melrose
- **Chicago:** #chicago, lbvs, chicago, blackhawks, #bears, #bulls, mfs, cubs, burbs, bogus
- **Philadelphia:** jawn, ard, #phillies, sixers, phils, wawa, philadelphia, delaware, philly, phillies

The plurality of these terms are place names (underlined) and geographically specific entities (italicized), such as sports teams (*dodgers*, *sixers*), businesses (*wawa*, a grocery store), and local government agencies (*mta*, which is responsible for mass transportation in New York City). However, there are several other types of words, which are of greater interest for dialect.

- **Dialect words from speech.** The term *jawn* was already discussed as a feature of spoken Philadelphia English, and the terms *burbs* (suburbs) and *bogus* (fake) may also be recognized in spoken language. The term *deadass*—typically meaning “very,” as in *deadass serious*—may be less familiar, and might have passed unnoticed without the application of automated techniques.
- **Alternative spellings.** The spelling *werd* substitutes for the term *word*—but only in the senses identified by Cutler (1999), as in *oh, werd?* (*oh really?*), or as affirmation, as in *werd, me too*. Note that the spelling *word* is also used in these same contexts, but the spelling *werd* is almost never used in the standard sense.

More remotely, *ard* is an alternative spelling for *alright*, as in:

- (4) @name ard let me kno
- (5) lol (*laugh out loud*) u'll be ard

Similarly, *brib* is an alternative spelling for *crib*, which in turn signifies *home*.

- (6) bbq (*barbecue*) at my fams (*family's*) brib
- (7) in da brib, just took a shower

Nationally, *brib* appears at a rate of once per 22,000 messages, which is roughly 5% as often as *crib*. But in the New York City area, *brib* appears at a rate of once per 3,000 messages.

A final example is *baii*, meaning a friend or partner. As shown in the second example below, it may also function as a pragmatic marker: similar to *man* in Cheshire's (2013) study of urban English in the U.K., it can be used without referring to any specific individual.

- (8) look at my baii @name congrats again wish i was there 2 see u walk baii
- (9) i'm outta here baii

- **Abbreviations.** The abbreviation *lol* (*laugh out loud*) is well-known in the discourse about social media text, but several lesser-known abbreviations have strong regional affiliations. These include *lbvs* (*laughing but very serious*) and *mfs* (*motherfuckers*, as in *these mfs are crazy*). My prior work with collaborators at Carnegie Mellon University (Eisenstein *et al.* 2010) identified several other phrasal abbreviations with non-uniform geographical distributions, including *af* (an intensifier signifying *as fuck*), *ctfu* (*cracking the fuck up*), and *lls* (*laughing like shit*).

- **Combinations** A few words appear to combine aspects of multiple types. The word *odee* is a phonetic spelling of the abbreviation *od*, which stands for *overdose*, but it is now used as an intensifier with considerable syntactic flexibility.

- (10) she said she odee miss me
 (11) its rainin odee :(
 (12) for once i'm odee sleepy

The geographical distributions of four of these terms are shown in Figure 21.3. Another measure of the regional specificity of these words can be seen in the cumulative fraction of word counts accounted for by their home cities (Figure 21.4). For example, for the word *yinz*, 77% of the counts come from Pittsburgh; for *ard*, 81% of the counts come from its top two cities of Philadelphia and Baltimore. Each subfigure also includes the overall cumulative proportion of word counts, indicating that 16% of all counts come from the largest metropolitan area, New York, and that this fraction increases slowly to nearly 50% when the ten largest metropolitan areas are considered. This line is what one would expect to see for words with no significant geographical association, and indeed, Figure 21.4a shows *yall* and *hella* track it closely. Figures 21.4b and 21.4c show that many of the strongest geographical orientations belong to automatically discovered social media terms—particularly to terms associated with New York, such as *baii*, *brib*, and *odee*. Figure 21.4d shows that the geographical associations for these social media terms are stronger than the corresponding associations for many place and entity names.

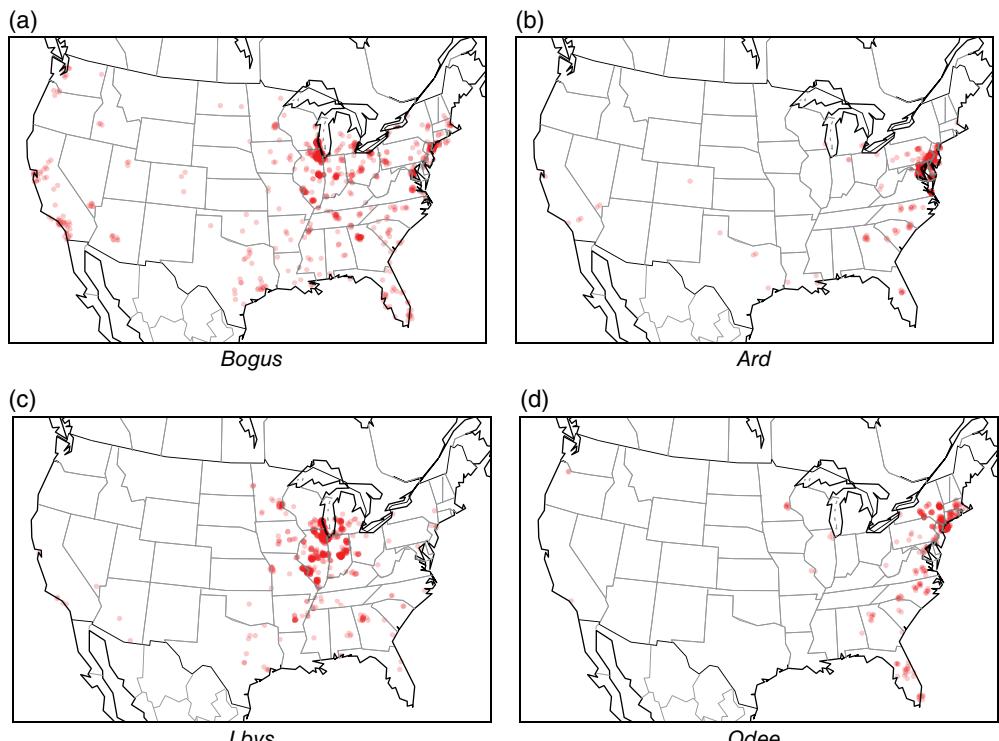


Figure 21.3 Four examples of lexical variables discovered from social media analysis.

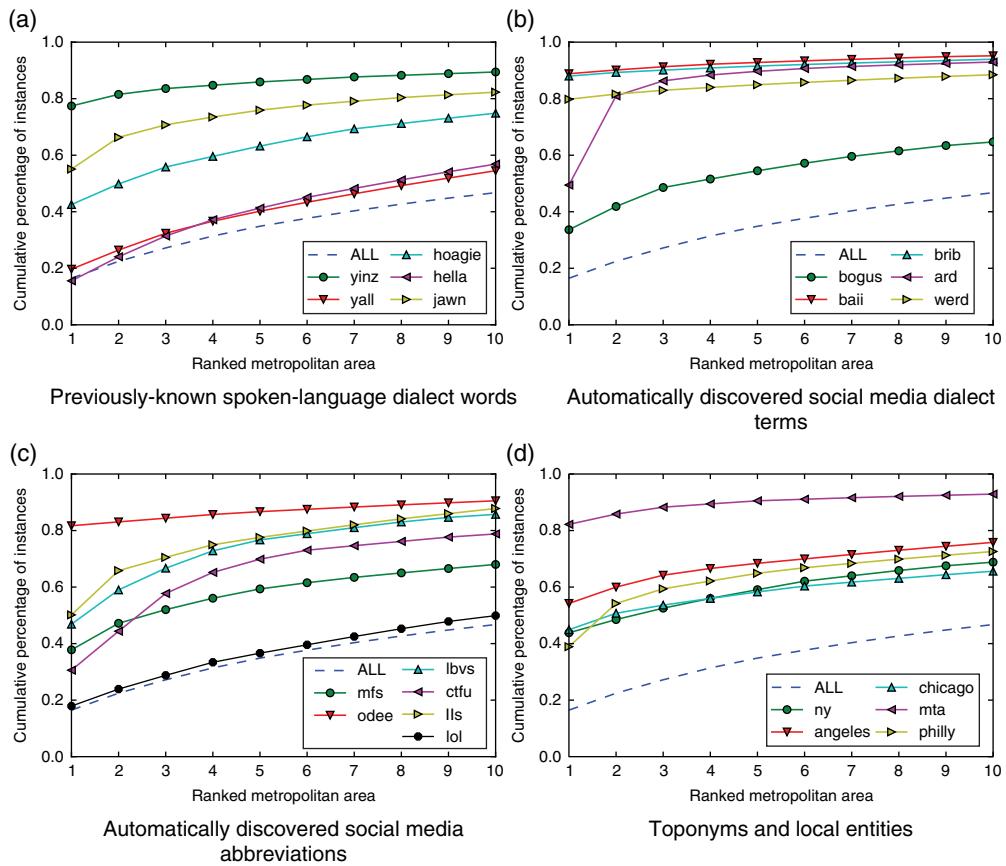


Figure 21.4 Cumulative proportion of counts across top metropolitan areas. The plot for a word will be flat to the extent that its usage is dominated by a single city, and will track the “all” line to the extent that its usage is independent of geography.

21.5 Discovering Dialect Regions

Figure 21.4 shows that many words are strongly associated with individual metropolitan areas—indicated by a high initial value and then a plateau in the graph. But some words, such as *ard*, *lls*, and *ctfu* have a steeper slope for the first few points—indicating that they are strong in two or more metropolitan areas. This suggests that interesting dialect regions may span multiple metropolitan areas. Moreover, such a view could reveal different linguistic patterns, since the words we are currently analyzing were chosen on the basis of their association with individual metropolitan areas. By grouping metropolitan areas, we are more likely to notice variables that belong to smaller cities, which individually might not have large enough word counts to survive the statistical shrinkage applied to individual metropolitan areas.

Other corpus-based approaches to dialectometry have applied clustering to identify dialect regions from matrices of *distances* between more fine-grained geographical areas (Heeringa and Nerbonne 2001; Szemrećsanı 2011). Such distance matrices are typically computed from catalogs of predefined dialect features; for example, Szemrećsanı (2011) uses a set of 57 morphosyntactic features, including the frequency of alternative tense markings

and systems of negation. More details on these approaches can be found in Chapters 18 and 23 of this volume. In social media, the relevant dialect features may not be known in advance; part of the goal is to discover them from raw text. While it is still possible to compute distance functions—for example, by applying a divergence metric between the word frequency distributions in two metropolitan areas (Wing and Baldridge 2011)—such distance metrics can be difficult to apply when data is sparse and high-dimensional, because they are sensitive to the amount and nature of smoothing applied to the word counts.

Recall that the overall goal is to partition the map into a set of regions, which are both spatially compact and linguistically consistent. We can frame this objective probabilistically: we want to choose a set of regions and regional word distributions so that the observed data—the text and its geolocations—have a high likelihood under some reasonable probability model. To do this, we treat the regions, and the assignment of authors to regions, as *latent variables*. We then define a probabilistic model that unifies the latent variables with the observed text and geolocations (Eisenstein *et al.* 2010). Bayesian inference in this model combines the likelihood of the observed data with our prior beliefs about the latent variables (e.g., the typical size of a geographical region), using the machinery of Bayes' law (Murphy 2012; Nerbonne 2007).

How can we define a probabilistic model that unifies the observations and latent variables? Our approach will be “generative,” in that we propose a fictional stochastic process by which the latent variables and observations are generated. This generative process is not intended as a psycholinguistic model of writing; it is simply a way to arrange the variables so that the quantities of interest can be recovered through statistical inference. There are three key quantities: the assignment of authors to dialect regions, the spatial extent of the dialect regions, and the word distributions associated with each region.³ The relationships between these quantities can be formalized by assigning them variable names, and proposing statistical dependencies between the variables, as shown in Algorithm 21.1.

```

for author  $a \in \{1 \dots A\}$  do
    Randomly sample a region  $z_a \sim P(z; \theta)$ , where  $\theta$  defines a prior distribution over regions. For example,  $\theta$  will assign high likelihood to the region containing New York, and low likelihood to more sparsely-populated regions.
    for each tweet  $t \in \{1 \dots T_a\}$  do
        Randomly sample a geolocation  $\bar{x}_{a,t} \sim P(\bar{x}_{a,t}; \varphi_{z_a})$ , where  $\bar{x}_{a,t}$  is a latitude-longitude pair and  $\varphi_{z_a}$  specifies a distribution over such pairs. The form of the distribution over latitude and longitude pairs is Gaussian in prior work (Eisenstein et al. 2010). In general, the Gaussian distribution is inappropriate for latitude and longitude on a sphere, but it is a suitable approximation for a small section of the globe, such as the continental United States.
        for each word token  $n \in \{1 \dots N_{t,a}\}$  do
            Randomly sample a word  $w_{a,t,n} \sim P(w; \vec{m}, \vec{\beta}_{z_a})$ , using the probability distribution defined in Equation 21.1.
        end
    end
end

```

Algorithm 21.1 Stochastic generative model for geolocated text.

The goal of statistical inference is to obtain estimates of the quantities of interest, under which the observed words and geolocations attain high probability. Specifically, we must infer the assignment of authors to regions (z), the prior distribution over regions (θ), the geographical extent of each region (φ), and the word distributions for each region (β). By

summarizing the probabilistic dependencies between the latent and observed variables, Algorithm 21.1 provides the basic specification for an inference procedure. Toolkits for *probabilistic programming* are capable of automatically transforming such a specification directly into executable code for statistical inference, without requiring any manual derivations or programming (Lunn *et al.* 2000; Goodman *et al.* 2008; Stan Development Team 2014).⁴ However, such toolkits were not used in this work, because this model admits relatively straightforward inference through the application of variational expectation maximization (Wainwright and Jordan 2008). The details of this procedure are beyond the scope of this chapter, and are described in prior publications (Eisenstein *et al.* 2010, 2011a). Variational expectation maximization is similar to soft K-means clustering, alternating between two steps: (1) making soft assignments of authors to regions (clusters), and (2) updating the linguistic and geographic centroids of the clusters. Eventually, this procedure converges at a local optimum.

21.6 Change Over Time and Other Next Steps

The analysis thus far has been entirely synchronic, but social media data can also shed light on how on-line language changes over time. Many of the terms mentioned in previous sections, such as *ctfu*, *ard*, *baii*, are nearly unknown in English writing prior to the recent era of computer-mediated communication. Yet dialect lexicons have expanded so rapidly that these terms are now in use among several thousands of individuals, and in some cases, across wide areas of the United States.

Figure 21.5 shows some examples. The term *af* (*as fuck*), is used mainly in Southern California and Atlanta in 2010, but attains widespread popularity by 2012. The term *ion* (meaning *i don't*; it is very rarely used in the chemical sense in this dataset) appears in a few scattered Southern cities in 2010, but spreads widely throughout the South by 2012. The emoticon *--* (indicating ambivalence or annoyance) was popular in several of the largest coastal urban areas in 2010—with remarkably limited popularity in the interior metropolises of Chicago, Dallas, and Houston—but reached widespread urban popularity by 2011, and nearly universal usage by 2012. These examples all offer support for various versions of the gravity model of linguistic change (Trudgill 1974), in which new features spread first between the most populated cities, with a limited role for geographical diffusion. However, other examples are more difficult to explain in terms of population alone: for example, *ctfu* (*cracking the fuck up*) spreads from its origin in Cleveland to adjacent parts of Pennsylvania, and from there to coastal cities along the mid-Atlantic; it does not spread westward to the large, nearby metropolises of Detroit and Chicago until much later. The abbreviation *lbvs* (*laughing but very serious*) is used almost exclusively in Chicago in 2010, and becomes popular in several other Midwestern cities by 2012—although not yet the nearby cities of Detroit and Cleveland. The phonetic spelling *ard* is highly popular in Baltimore and Philadelphia, but does not spread to the neighboring city of Washington DC—a distance of 90 kilometers. Several of the terms most strongly associated with New York City (*odee*, *werd*, *deadass*) also fail to attain much popularity outside their city of origin.

The search for explanations beyond population size and geographical proximity leads inevitably to considerations of race, class, and cultural differences, and these topics have long been seen as central to sociolinguistic research on language variation and change (Gordon 2000; Labov 2011). This strain of sociolinguistics has focused primarily on sound changes such as the Northern Cities Shift (Labov 1994), but the slow pace of sound change and the limited number of measurable linguistic variables pose challenges for the distillation of a quantitative “grand unified theory” that accounts for geography, population size, and demographics. In this sense, social media data offers unique advantages: change is rapid

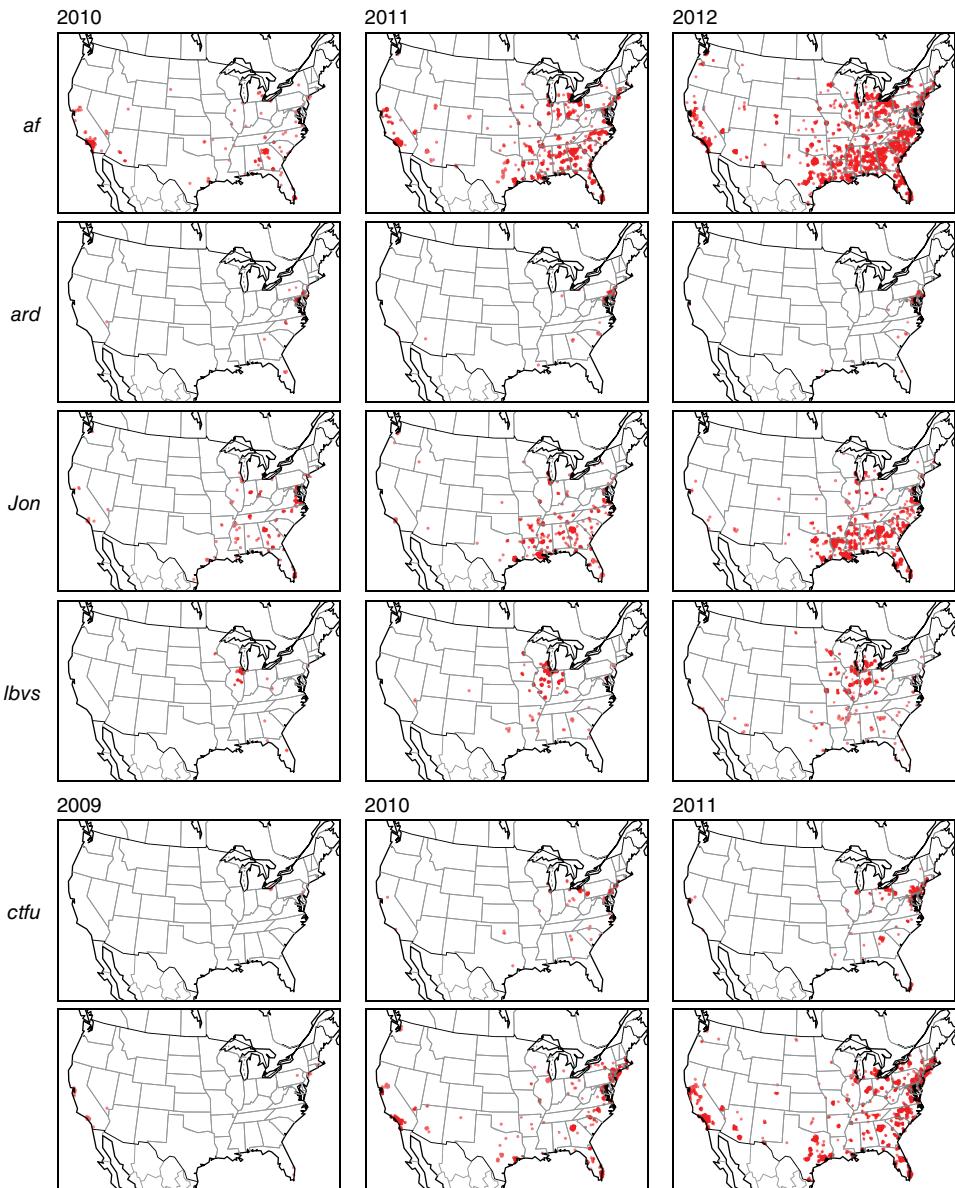


Figure 21.5 Geolocations for messages containing the words *af* (as fuck), *ard* (alright), *ion* (i don't), *lbvs* (laughing but very serious), *ctfu* (cracking the fuck up), and the emoticon -_- (ambivalence or annoyance).

enough to be measurable in real time, and the methods described in this chapter can be used to identify hundreds of linguistic features whose frequency correlates with both time and place. While such correlational analysis has already shown that many lexical features have strong demographic profiles (Eisenstein *et al.* 2011b), the role of demographics in the diffusion of these terms has yet to be definitively captured, though this is a topic of ongoing investigation. As Labov (1994) shows, the tightly interconnected nature of the phonological system means that a single sound change can cause a series of other changes, which may take several generations to play out. This property would seem to be lacking from lexical

change—although the lexicon features its own patterns of interdependence (Pierrehumbert 2010)—and if so, this would limit the extent to which generalizations can be drawn between these two systems. But from a social perspective, it may be precisely this relative simplicity that makes the lexicon the ideal laboratory in which to disentangle the complex social phenomena that regulate language variation and change.

Social media data may have still more to tell us about dialect. A clear priority is to export these methods to dialects outside the United States, particularly to places where variation is better described by continua rather than discrete regions (Heeringa and Nerbonne 2001), which may require new computational methods. Another topic of interest is the relationship between written and spoken dialect: to what extent is phonological regional variation transcribed into written language in social media? Preliminary work suggests that several phonological variables are often transcribed in Twitter writing (Eisenstein 2015), but these phenomena (such as “g-dropping”) mainly correlate with register variation, rather than with geographical dialects. Still another key question is how the use of geographical variables is affected by other properties of the author: for example, younger people may write more informally, but older authors may be more likely to use traditional variables such as *yinz*. Finally, sociolinguistics has been increasingly concerned with the role of language variation in fine-grained conversational situations (Jaffe 2012)—a phenomenon that is very difficult to measure at scale without social media data. Recent work shows that in conversational dialogues, authors modulate the frequency with which they use local variables depending on the size and identity of the likely audience (Pavalanathan and Eisenstein 2015), lending support to theories such as accommodation (Giles *et al.* 1991) and audience design (Bell 1984). It is hoped that further studies in this direction will shed light on how dialect is perceived, and how it is deployed to create and reflect social relationships.

Acknowledgments

Thanks to Brendan O’Connor for providing the data on which this chapter is based, and for many insightful conversations over a fun and productive long-term collaboration on this research. Thanks are also due to John Nerbonne, Shawn Ling Ramirez, and Tyler Schnoebelen for providing helpful editorial suggestions on this chapter. This work benefitted from collaborations and discussions with David Bamman, Scott F. Kiesling, Brendan O’Connor, Umashanthi Pavalanathan, Noah A. Smith, Tyler Schnoebelen, and Eric P. Xing, and was supported by a grant from the National Science Foundation.

Biographical Note

Jacob Eisenstein is Assistant Professor in the School of Interactive Computing at the Georgia Institute of Technology, where he leads the Computational Linguistics Laboratory. He received a doctorate in Computer Science from the Massachusetts Institute of Technology in 2008.

NOTES

1 Facebook reported 757 million daily active users in December 2013 (<http://investor.fb.com/releasedetail.cfm?ReleaseID=821954>, retrieved on April 5, 2014). Twitter reported 271 million monthly active users in August 2014 (<https://about.twitter.com/company>, retrieved on August 3, 2014).

- 2 <https://github.com/jacobeisenstein/SAGE>.
- 3 The models described by Eisenstein *et al.* (2010) and Eisenstein *et al.* (2011a) are more ambitious, attempting to distinguish latent *topics*, which capture variation that is independent of geography. This is beyond the scope of this chapter.
- 4 A repository of software packages is found at <http://probabilistic-programming.org>.

REFERENCES

- Alim, H. Samy. 2009. Hip hop nation language. In Duranti, A., editor, *Linguistic Anthropology: A Reader*, pages 272–289. Wiley-Blackwell, Malden, MA.
- Altmann, Eduardo G., Janet B. Pierrehumbert, and Adilson E. Motter. 2011. Niche as a determinant of word fate in online groups. *Plos one*, 6(5):e19009.
- Argamon, Shlomo, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. 2007. Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*, 12(9).
- Bell, Allan. 1984. Language style as audience design. *Language in Society*, 13(2): 145–204.
- Boyd, Danah, and Nicole B. Ellison. 2007. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1): 210–230.
- Bucholtz, Mary, Nancy Bermudez, Victor Fung, Lisa Edwards, and Rosalva Vargas. 2007. Hella nor cal or totally so cal? the perceptual dialectology of california. *Journal of English Linguistics*, 35(4): 325–352.
- Chang, Jonathan, Itamar Rosenn, Lars Backstrom, and Cameron Marlow. 2010. ePluribus: Ethnicity on social networks. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, pages 18–25, Menlo Park, California. AAAI Publications.
- Cheng, Zhiyuan, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*, pages 759–768.
- Cheshire, Jenny. 2013. Grammaticalisation in social context: The emergence of a new English pronoun. *Journal of Sociolinguistics*, 17(5): 608–633.
- Collister, Lauren B. 2011. *-repair in online discourse. *Journal of Pragmatics*, 43(3): 918–921.
- Crystal, David. 2006. *Language and the Internet*. Cambridge University Press, second edition.
- Cutler, Cecilia A. 1999. Yorkville crossing: White teens, hip hop and African American English. *Journal of Sociolinguistics*, 3(4): 428–442.
- Danescu-Niculescu-Mizil, Cristian, Michael Gamon, and Susan Dumais. 2011. Mark my words! linguistic style accommodation in social media. In *Proceedings of the Conference on World-Wide Web (WWW)*, pages 745–754.
- Danescu-Niculescu-Mizil, Cristian, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the Conference on World-Wide Web (WWW)*, pages 699–708, Lyon, France.
- Danescu-Niculescu-Mizil, Cristian, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013a. A computational approach to politeness with application to social factors. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 250–259, Sophia, Bulgaria.
- Danescu-Niculescu-Mizil, Cristian, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013b. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the Conference on World-Wide Web (WWW)*, pages 307–318.
- Dredze, Mark, Michael J. Paul, Shane Bergsma, and Hieu Tran. 2013. Carmen: A Twitter geolocation system with applications to public health. In *AAAI Workshop on Expanding the Boundaries of Health Informatics Using Artificial Intelligence*, pages 20–24.
- Dresner, Eli, and Susan C. Herring. 2010. Functions of the nonverbal in CMC: Emoticons and illocutionary force. *Communication Theory*, 20(3): 249–268.
- Duggan, Maeve, and Aaron Smith. 2013. Social media update 2013. Technical report, Pew Research Center.
- Eisenstein, Jacob. 2013. Phonological factors in social media writing. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 11–19, Atlanta.
- Eisenstein, Jacob, Amr Ahmed, and Eric P Xing. 2011a. Sparse additive generative models of text. In *Proceedings of the International Conference*

- on Machine Learning (ICML)*, pages 1041–1048, Seattle, WA.
- Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pages 1277–1287, Stroudsburg, Pennsylvania. Association for Computational Linguistics.
- Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2012. Mapping the geographical diffusion of new words. Technical Report 1210.5268, ArXiV.
- Eisenstein, Jacob, Noah A. Smith, and Eric P. Xing. 2011b. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 1365–1374, Portland, OR.
- Garley, Matt, and Julia Hockenmaier. 2012. Beefmoves: dissemination, diversity, and dynamics of English borrowings in a German hip hop forum. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 135–139, Jeju, Korea.
- Gilbert, Eric. 2012. Phrases that signal workplace hierarchy. In *Proceedings of Computer-Supported Cooperative Work (CSCW)*, pages 1037–1046.
- Giles, Howard, Justine Coupland, and Nikolas Coupland. 1991. *Contexts of accommodation: Developments in applied sociolinguistics*. Cambridge University Press.
- Goodman, Noah D., Vikash K. Mansinghka, Daniel M. Roy, Keith Bonawitz, and Joshua B. Tenenbaum. 2008. Church: A language for generative models. In *Proceedings of Uncertainty in Artificial Intelligence (UAI)*, pages 220–229.
- Gordon, Matthew J. 2000. Phonological correlates of ethnic identity: Evidence of divergence? *American Speech*, 75(2): 115–136.
- Green, Lisa J. 2002. *African American English: A Linguistic Introduction*. Cambridge University Press, Cambridge, U.K.
- Heeringa, Wilbert, and John Nerbonne. 2001. Dialect areas and dialect continua. *Language Variation and Change*, 13(3): 375–400.
- Hong, Liangjie, Amr Ahmed, Siva Gurumurthy, Alexander J. Smola, and Kostas Tsoutsouliklis. 2012. Discovering geographical topics in the twitter stream. In *Proceedings of the Conference on World-Wide Web (WWW)*, pages 769–778, Lyon, France.
- Huberman, Bernardo, Daniel M. Romero, and Fang Wu. 2008. Social networks that matter: Twitter under the microscope. *First Monday*, 14(1).
- Jaffe, Alexandra, ed. 2012. *Stance: Sociolinguistic Perspectives*. Oxford Studies in Sociolinguistics. Oxford University Press.
- Johnstone, Barbara, Neeta Bhasin, and Denise Wittkofski. 2002. “Dahntahn” Pittsburgh: Monophthongal /aw/ and Representations of Localness in Southwestern Pennsylvania. *American Speech*, 77(2): 148–176.
- Kooti, Farshad, Haeryun Yang, Meeyoung Cha, P. Krishna Gummadi, and Winter A Mason. 2012. The emergence of conventions in online social networks. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, pages 194–201, Menlo Park. AAAI Publications.
- Kwak, Haewoon, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media? In *Proceedings of the Conference on World-Wide Web (WWW)*, pages 591–600, New York. ACM.
- Labov, William. 1994. *Principles of Linguistic Change*, volume 1: Internal Factors. Blackwell Publishers.
- Labov, William. 2011. *Principles of Linguistic Change*, volume 3: Cognitive and Cultural Factors. WileyBlackwell.
- Leskovec, Jure, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of Knowledge Discovery and Data Mining (KDD)*, pages 497–506.
- Lunn, David J., Andrew Thomas, Nicky Best, and David Spiegelhalter. 2000. WinBUGS – a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, 10(4): 325–337.
- Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin’ words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4): 372–403.
- Morstatter, Fred, Jurgen Pfeffer, Huan Liu, and Kathleen M. Carley. 2013. Is the sample good enough? Comparing data from Twitter’s Streaming API with Twitter’s Firehose. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, pages 400–408, Menlo Park, California. AAAI Publications.
- Murphy, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective*. The MIT Press.
- Nerbonne, John. 2007. The exact analysis of text. In Mosteller, F. and Wallace, D., editors, *Inference and Disputed Authorship: The Federalist Papers*. CSLI: Stanford, third edition.

- Nguyen, Dong, and Carolyn P. Rosé. 2011. Language use as a reflection of socialization in online communities. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 76–85.
- Owoputi, Olutobi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part of-speech tagging for online conversational text with word clusters. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 380–390, Stroudsburg, Pennsylvania. Association for Computational Linguistics.
- Paolillo, John C. 2011. Conversational codeswitching on usenet and internet relay chat. *Language@Internet*, 8(3).
- Pavalanathan, Umashanthi, and Jacob Eisenstein. 2015. Audience-modulated variation in online social media. *American Speech*, 90(2).
- Pierrehumbert, Janet B. 2010. The dynamic lexicon. In Cohn, A., Huffman, M., and Fougeron, C., editors, *Handbook of Laboratory Phonology*, pages 173–183. Oxford University Press.
- Postmes, Tom, Russell Spears, and Martin Lea. 2000. The formation of group norms in computer-mediated communication. *Human communication research*, 26(3): 341–371.
- Prabhakaran, Vinodkumar, Owen Rambow, and Mona Diab. 2012f. Predicting overt display of power in written dialogs. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 518–522.
- Romero, Daniel M., Brendan Meeder, and Jon Kleinberg. 2011. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the Conference on World-Wide Web (WWW)*, pages 695–704.
- Rosenthal, Sara, and Kathleen McKeown. 2011. Age prediction in blogs: A study of style, content, and online behavior in pre- and Post-Social media generations. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 763–772, Portland, OR.
- Schnoebelen, Tyler. 2012. Do you smile with your nose? Stylistic variation in Twitter emoticons. *University of Pennsylvania Working Papers in Linguistics*, 18(2):14.
- Stan Development Team. 2014. Stan: A c++ library for probability and sampling, version 2.2.
- Szmrecsanyi, Benedikt. 2011. Corpus-based dialectometry: a methodological sketch. *Corpora*, 6(1):45–76.
- Tagliamonte, Sali A., and Derek Denis. 2008. Linguistic ruin? LOL! Instant messaging and teen language. *American Speech*, 83(1): 3–34.
- Thurlow, Crispin. 2006. From statistical panic to moral panic: The metadiscursive construction and popular exaggeration of new media language in the print media. *Journal of Computer-Mediated Communication*, pages 667–701.
- Trudgill, Peter. 1974. Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. *Language in Society*, 3(2): 215–246.
- Wainwright, Martin J., and Michael I. Jordan. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2): 1–305.
- Walther, Joseph B., and Kyle P. D'Addario. 2001. The impacts of emoticons on message interpretation in computer mediated communication. *Social Science Computer Review*, 19(3): 324–347.
- Wing, Benjamin, and Jason Baldridge. 2011. Simple supervised document geolocation with geodesic grids. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 955–964, Portland, OR.

22 Logistic Regression Analysis of Linguistic Data

JOHN C. PAOLILLO

22.1 Introduction

Linguistic analyses are often faced with the situation of analyzing empirical data in the form of categories. This chapter addresses this need by presenting the logistic regression framework. Other models are used for categories, including ones more closely related to the dimension-reducing techniques of Nerbonne and Wieling (this volume). Here, we ignore this complexity to focus on the presentation of the logistic regression framework, and how it can be used to pose and answer questions regarding language variation. This chapter is organized as follows. We begin with a general discussion of how category data are relevant to dialect studies. Next we examine some sample data, and consider how models of this data are constructed and evaluated using logistic regression. This is followed by an example analysis of the sample data, and general conclusions about the application of the framework.

22.2 Category Data

Category data have values that are names for mutually exclusive conditions of the same phenomenon. For example, we may distinguish front and back vowels, recognizing a variable, perhaps “frontness” with two possible category values. Systems of category variables often cross-classify phenomena, so we have, for example high and low vowels in addition to front and back ones, and height and frontness cross-classify the observable vowels. Category variables may have two values (or *levels*), in which case they are *binary*, or more than two, in which case they are *polytomous*; these values may be ordered or otherwise structured. For example, for place of articulation, a complete set of values might include labial, dental, alveolar, palatal, and velar as mutually exclusive categories. There is a natural order of these categories, from front to back or vice versa, that is linguistically relevant, and which may play a role in the correct analysis.

Category variables are analyzed by replacing them with continuous-valued *contrast* variables that are orthogonal and linearly independent. In general, a category variable with m categories is replaced with $m - 1$ contrasts (Fox 1997, 196–202). Table 22.1 illustrates three types of contrasts for a hypothetical variable *City*, with four levels. In each, the three new variables are used to replace the values of the *City* variable for every observation. Special functions in software are typically used to generate contrasts.

Table 22.1 Three types of contrasts for a four-level variable, *City*.

City	Treatment			Sum			Helmert		
	N	P	A	B-N	B-P	B-A	B-N	P-N/B	A
Boston	0	0	0	-1	-1	-1	-1	-1	-1
New York	1	0	0	1	0	0	1	-1	-1
Philadelphia	0	1	0	0	1	0	0	2	-1
Arlington	0	0	1	0	0	1	0	0	3

The first set of contrasts are *treatment contrasts* (also called *effects coding*), in which each new variable is coded 1 for one and only one category and 0 for all others. These contrasts treat Boston as the reference category and express only the difference between Boston and the other cities. They do not compare the remaining cities with each other, and they leave Boston to be characterized implicitly. Treatment contrasts are typically implemented so that the default reference category is the alphabetically first category. This is seldom helpful, and the analyst has to be prepared to understand and take control of the software's method for generating contrasts.

The second set of contrasts, called *sum contrasts* also uses Boston as a reference category, but by opposing it, pairwise, with each of the other categories. Sum contrasts allow each city in Table 22.1 to be expressed in terms of the full set of cities as a whole. The third type of contrasts, *Helmert contrasts*, are most useful when one wishes to incorporate ordering information in the analysis, as can be seen in the placement of the -1 and 0 values; this coding corresponds to a North-South ordering of the cities along the Eastern Coast of the US Other types of contrasts include nested dichotomies (Fox 1997, 474). Hence, contrasts can be used to address questions about a broad range of category structures in linguistic analysis.

When a variable represents the specific phenomenon of interest around which an analysis will be framed, it is known as the *dependent* or *response* variable.¹ Other observed variables are known as *independent* or *predictor* variables. Response and predictor variables may both be continuous or category variables. When the response is a continuous variable, linear models are used, when it is a category variable, logistic regression may be used. Continuous and category predictors do not necessarily affect the choice of model. When predictors are continuous, they are often called *covariates*.

22.3 Example Data

Table 22.2 presents sample data for *was/were* variation in York English, arranged as a *contingency table*, which we use for discussion and the presentation of example analyses below. The original data are from Sali Tagliamonte, and are a subset of the data analyzed in Tagliamonte (1998). This table shows the cross-classification of observations by six variables: the realized form of *was/were* (the response variable), affirmative and negative sentences, person and number (first- and third-person singular versus second-person or plural), gender (male/female), age (older, middle, younger), and individual speaker.² The combinations of observed variables that make up the contingency table are called *cells*; Table 22.2 has 92 cells, with a count of *was* and *were* values for each.³

Table 22.2 Contingency table for *was/were* variation by individual speaker (single character labels), affirmative/negative, standard *was/were* contexts, gender, age, and speaker.

		<i>Singular first and third person</i>		<i>Plural or second person</i>	
		<i>Affirmative</i>	<i>Negative</i>	<i>Affirmative</i>	<i>Negative</i>
	<i>Speaker</i>	<i>was : were</i>	<i>was : were</i>	<i>was : were</i>	<i>was : were</i>
Females					
<i>Younger</i>	W	379 : 4	24 : 10	9 : 57	0 : 2
	h	97 : 1	6 : 2	13 : 25	0 : 4
	n	123 : 5	9 : 0	1 : 24	0 : 1
	d	172 : 2	15 : 4	8 : 17	0 : 7
<i>Middle</i>	f	134 : 1	7 : 0	1 : 55	0 : 4
	a	330 : 4	9 : 2	44 : 64	1 : 4
	R	147 : 2	12 : 0	3 : 29	0 : 6
	t	139 : 6	16 : 0	12 : 35	0 : 5
<i>Older</i>	c	321 : 10	8 : 1	8 : 51	0 : 3
	g	226 : 3	12 : 0	7 : 48	0 : 8
	å	104 : 2	3 : 2	1 : 42	0 : 4
	o	136 : 0	5 : 0	8 : 26	0 : 3
	M	60 : 4	7 : 0	12 : 53	0 : 10
Males					
<i>Younger</i>	y	167 : 13	3 : 6	8 : 27	0 : 5
	H	200 : 6	4 : 1	2 : 41	0 : 2
<i>Middle</i>	A	171 : 2	7 : 1	2 : 39	0 : 1
	≠	134 : 36	0 : 1	4 : 30	0 : 1
	s	90 : 3	3 : 0	12 : 46	0 : 0
	m	201 : 4	8 : 0	21 : 38	0 : 3
<i>Older</i>	q	152 : 1	2 : 0	11 : 29	0 : 0
	e	69 : 0	1 : 1	9 : 29	0 : 1
	r	239 : 11	15 : 1	1 : 70	0 : 3
	j	154 : 9	5 : 0	8 : 58	0 : 0

The realization of *was/were* is the main phenomenon of interest, so the proportion of *was* realized in a cell is the response variable, symbolized *y*; the remaining variables are predictors. The observed counts are greatest where *was/were* appears in a standard context: *was* is preferred in first- and third-person singular, and *were* in second-person or plural. The counts of non-standard *was/were* are many times smaller, though not usually zero. Affirmative polarity might have more non-standard use than negative polarity, although possibly not in the second person or plural, for which we have few observations of *was* (only one). Negative polarity is observed more rarely than affirmative polarity, leaving these patterns somewhat in doubt. Individuals also appear to vary considerably in their *was/were* use; theories of language change predict that if *was/were* is changing in York, male and female speakers should vary systematically in this respect, as should older and younger speakers.⁴ The original data files used only single character codes for all variables, so individual speakers are indicated by single characters.

22.4 Model Fundamentals

To compare the counts in a contingency table, we need to have a *model*, which is an expression of the values that we expect to find under different combinations of observed variables. Normally, we *estimate* a model from the available data; this requires a *model statement* expressing the expected relation among the variables. *Parameter* values are computed from the data that are most likely under those assumptions; when entered back into the model statement, these values *predict* values close to what is observed. The differences between the predicted and observed values are aggregated into a statistic of *model fit*. Each parameter represents a hypothesis or claim whose effect does not depend on the others; they are *independent*. The total number of these parameters measures the complexity of the model, or its *degrees of freedom*. Parameter values, the model fit statistic, and the degrees of freedom are used together with a *distributional theory* for the model to conduct significance tests.

This description fits the well-known chi-squared test, familiar to many from introductory statistics courses. This test is more properly called the *chi-squared test for the model of independence*; we could apply it to Table 22.2 by collapsing along different dimensions of the table. For example, to investigate gender differences within grammatical contexts, we might aggregate all individual speakers, as well as positive and negative polarity, to get the values in Table 22.3.⁵

The model for this test is a table of expected values based on the row and column proportions and the grand total. Model fit is measured by the chi-squared statistic, a sum of individual deviances between observed and expected values for each table cell, and the distributional theory of the chi-squared allows us to assess our test values against a criterion value, for example, 3.84 for $p \leq 0.05$ at 1 degree of freedom for both sub-tables in Table 22.3. Under this we might conclude that female and male speakers of York English appear to be significantly different in their use of *was/were* in first- and third-person singular grammatical contexts (because $p < 0.05$), but not in second-person or plural contexts (because $p > 0.05$).⁶

Application of the chi-squared test in this fashion has a number of problems. First, only two dimensions can be tested at a time, and to run tests on multi-way data we need to either condition (as we did for grammatical context) or aggregate (as we did for speaker and polarity), limiting our attention to one potential relationship at a time. Worse yet, the different questions we might ask *confound* each other. For example, we cannot know what the contributions of age and polarity are to the patterns in Table 22.3, because they are no longer available in the analysis. If we instead aggregate by gender so that we can compare age and *was/were* in the two grammatical contexts, we find that in first- and third-person singular,

Table 22.3 Chi-squared test for the model of independence for gender and *was/were* use, carried out separately for first- and third-person singular versus second-person or plural grammatical contexts.

	First and third singular			Second-person or plural		
	Was	Were	Total	Was	Were	Total
Female	2501	65	2566	128	587	715
Male	1625	96	1721	78	423	501
Total	4126	161	4287	206	1010	1216
Chi-sq			26.43		Chi-sq	1.14
p			0.000000274		p	0.286

was/were use is not significantly different by age (5.27 at 2 degrees of freedom, $p = 0.07$), but in second-person or plural it is (13.39 at 2 degrees of freedom, $p = 0.001$), results that appear to at least partly contradict those of Table 22.3. In obtaining these two results, we also re-use the same data, but the test assumes that the data are independently sampled. It is relatively easy to demonstrate using hypothetical data that both the significance and direction of apparent patterns in tables like these can change depending on how the data are aggregated (Simpson's paradox). Furthermore, the method for computing the model of independence does not generalize to multi-way tables like Table 22.2, and the chi-squared test gives us no guidance in reasoning about the lack of fit even in two-way tables. For these reasons, we need a more general approach to analyzing multi-way tables.

Such an approach is developed for a model statement with the form in (22.1), in which separate terms are added together to predict the response y : this is the Generalized Linear Model (GLM) form (Agresti 1990, 1996; McCullagh and Nelder 1989).

$$f(y) = a + b_1x_1 + b_2x_2 \dots + e \quad (22.1)$$

In (22.1), the variable y is the response, for example, the proportion of *was* in the cells of Table 22.2. The *link function* $f(\bullet)$ transforms the y values into a scale useful for analysis. Different functions offer models for different kinds of data. One such function is the natural logarithm $f(y) = \ln y$; it is typically used with count data, giving a *log-linear* or *Poisson* model. Another is the *logit* or log-odds function: $f(y) = \text{logit}(y) = \ln(y / (1 - y))$; it is used with proportions, giving a *logistic regression* model. The log-linear model is mathematically more general: a logistic regression model can always be expressed as a log-linear model, not vice versa (Paolillo 2002, 185). The data in Table 22.2 can be modeled either way, but sampling techniques used in variation studies seldom support interpreting log-linear main effects, whereas logistic regression effects are relatively independent of sampling. Here, only logistic regression analysis is illustrated.⁷

The term a in the model statement is the intercept, a parameter representing a reference point common to all observations; the other model parameters are expressed relative to a . The x_1 , x_2 , and so on, are the predictor variables; there are as many of these as are needed in the analysis. Crucially, we assume that these x are independent; the model parameters may be distorted and unreliable if this assumption is violated. Such violations are referred to as *multicollinearity*, which should be routinely checked in regression models. The b_1 , b_2 , and so on, are the model parameters, one for each x variable. The last term e represents the deviance or *residual* variation of each observation from the value predicted by the other terms in the model. Often this term is omitted from the model statement, in which case the y is usually replaced by \hat{y} , where the hat notation $\hat{\cdot}$ explicitly indicates we are talking about an estimate of y .

The model's degrees of freedom is the number of parameters estimated from the data. Normally, this will be the number of a and b values used in the model. Sometimes it is required that we know the *residual degrees of freedom*; this is the total number of observations (total degrees of freedom) minus the model degrees of freedom. Both are often called simply "degrees of freedom" without distinguishing model or residual, a practice that easily leads to confusion.⁸ Categorical variables require one degree of freedom for each contrast variable, hence for a variable with m categories, there will be $m - 1$ degrees of freedom.

The model fit statistic for logistic regression models is G^2 , also called the likelihood-ratio chi-squared of the model, is given in (22.2), where y is the observed value and \hat{y} is the expected (estimated) value under the model. Higher values of (22.2) indicate a poorer fit of the model to the observed data.

$$G^2 = 2 \sum y \ln(y/\hat{y}) \quad (22.2)$$

The distributional theory for the model allows us to compare G^2 with criterion values in order to conduct significance tests. For example, the *likelihood ratio test* is conducted by computing the G^2 for two models, a *constrained* model where some b parameters are set to zero, and an *unconstrained* model where all the b values are allowed to take their best estimates. The constrained model is *nested* within the unconstrained model (Bishop *et al.* 1975), and the difference in the G^2 of the two models is chi-squared distributed at the difference in degrees of freedom of the two models (the number of $b_i x_i$ terms in the excluded set). A criterion probability value is chosen, and values of G^2 that exceed that are interpreted as "significant": $p \leq 0.05$ is customary, although, depending on the number of tests to be done, a stricter criterion like the Bonferroni correction may be used.⁹ Significance suggests it is unsafe to ignore the contribution of the parameters tested, whereas non-significance warns against their interpretation. One may test a set of parameters (corresponding to a polytomous category variable) or a single parameter (corresponding to a binary variable or a covariate). In this way, a regression model like (22.1) allows the analyst to focus the interpretation of observations on relationships that are likely to be meaningful.¹⁰

The model in (22.1) is intended to allow us to obtain estimates of the b values given some data, but its form does not fully determine what the b values should be: different scales of the x values result in different b values. Two models with the same x variables but different scales are said to have different *parameterizations*. For binary categories, the direction of the contrast changes the parameterization and affects the interpretation of the model, for example, whether one considers the variation among females relative to males or that of males relative to females. The choice is arbitrary, but of utmost importance in the interpreting the b values. For polytomous variables, the set of contrasts determines what significance tests are evaluated: each test asks if some $b_i x_i$ can be set to zero, meaning that different parameterizations of a category variable, whereas they make equivalent predictions when taken as a set, provide different information about the relationships in the data.

When the response variable y is a polytomous variable, the structure of the category system is especially important: for a response with m levels, the $m - 1$ contrasts represent different response variables, and a separate model can be estimated for each (Bishop *et al.* 1975). For this approach to be worth the trouble of keeping the data comparable across all of the models, one must have very clear hypotheses at the outset. Example applications are rule ordering in /s/ lenition in Caribbean Spanish (Rousseau 1989; Sankoff and Rousseau 1989) and invariant *be* realization in AAVE (Labov 1969) as demonstrated in Paolillo (2002, 143–146).¹¹

The equation in (22.1) describes a model with *main effects*, that is, independent effects for each of the predictor variables. Sometimes it is important to consider effects involving more than one predictor variable simultaneously. Such effects are called *interaction effects*,¹² and they take the form of the $b_{ij} x_i x_j$ term in (22.3). Depending on software, interactions can be created automatically, or manually by creating a new variable whose values are the product $x_i x_j$. If different parameterizations of a category variable are considered, the number of possible products $x_i x_j$ can be large, and multi-collinearity can result, so care must be taken if interactions are coded manually.

$$\text{logit}(y_{i,j}) = a + b_i x_i + b_j x_j + b_{i,j} x_i x_j \dots + e \quad (22.3)$$

Interactions represent complex conditional relations among variables and models that contain them are generally more complex than those containing only main effects. Many hypotheses can be framed entirely in terms of main effects: preference for *was* in first- and third-person singular, negative polarity, by female speakers and younger ages are all represented by main effects.

Some questions require interaction effects, such as whether the preference for *was* is greater when first- and third-person singular co-occur with negative polarity, or if two groups disagree on the direction of preference for *was/were* variation in a specific environment, as might be the case if the two groups were diverging. Both cases require a *two-way interaction* to investigate (person/number by polarity or group by person/number). Three-way and *higher-order* interactions are possible, though harder to interpret, and researchers tend to avoid them. It is generally not safe to exclude investigation of interaction effects *a priori*, and linguistic applications of regression should systematically consider potential interactions (Paolillo 2011; Sigley 2003). In general, the best guidance for considering interaction effects comes from thoroughly understanding the substantive social, dialectological, and linguistic issues relevant to an analysis.

Logistic regression analyses may be conducted using many software packages; two that deserve special mention are Varbrul and R. The Varbrul family of programs was designed expressly for logistic regression; its latest representative, GoldVarb Lion (Sankoff, Tagliamonte and Smith 2012) has many legacy features, for example, variables can only be categories (using single-character codes) and many important model diagnostics are not available; guidance for its use is in the GoldVarb documentation, Paolillo (2002) and references therein.

R (R Development Core Team 2012) is a powerful programming language and environment, with a full range of built-in statistical functions. The `glm()` function offers a powerful interface to the entire GLM family and provides many useful model diagnostics. As a command-line environment, R requires a greater commitment to learn than many packages, but it is more consistent and flexible. Many books are available, with Baayen (2008) and Gries (2009) addressing a linguistic audience, though neither treats `glm()` in any depth, for which Crawley (2002, 2007) are more complete.¹³ For examples of logistic regression applied to linguistic analysis see references in Gorman and Johnson (2013) and Paolillo (2002:24–7).

22.5 Example Regression Analysis

We begin by stating a model of the variation we want to estimate. We use the R function `glm()` to fit the model statement in (22.4), which maps transparently onto (22.1), and results in the model summarized in Table 22.3. By default, a reference cell parameterization is used (plural or second person, affirmative, female, middle age group), so we have labeled the parameters in Table 22.3 with the levels of the category variables they represent.

$$\text{was} \sim 1 + \text{age} + \text{gender} + \text{persnum} + \text{polarity} \quad (22.4)$$

Table 22.4 Logistic regression model of York English *was/were* variation, with reference-cell parameterization.

	Estimate	Std. Error	z	p
<i>Intercept</i>	-1.1742	0.1121	-10.474	< 2e-16
<i>Sg1&3</i>	5.0374	0.1210	41.643	< 2e-16
<i>Negative</i>	-1.8504	0.1990	-9.301	< 2e-16
<i>Male</i>	-0.5922	0.1169	-5.065	4.08e-07
<i>Older</i>	-0.1871	0.1301	-1.438	0.1506
<i>Younger</i>	-0.2577	0.1476	-1.746	0.0808
				.

Table 22.4 presents the summary of parameter estimates for the model in (22.4), along with information about the standard error, z (sometimes given as t) and p values associated with them, which together comprise the *Wald test*. The null hypothesis of this test is that a specific parameter $b = 0$, and parameters are normally distributed around zero when this is true. Parameters are considered non-zero, and therefore, safe to interpret if the p -value of the Wald test is significant.

In Table 22.4, the p values associated with the intercept, person/number, polarity and gender are significant (being well below any criterion we would use) and the two age parameters are non-significant and should not be interpreted. We therefore interpret the significant parameter values. The reference cell is plural or second person contexts, affirmative polarity sentences used by females in the middle age group. This cell is fully characterized by the intercept, which is negative and moderately sized, hence a low rate of *was* use is predicted for this cell. First- and third-person singular contexts have a large positive parameter value, hence greater (almost categorical) *was* use is predicted for those contexts. Negative polarity decreases the predicted *was* use, and males also use *was* less than females.

Slightly different interpretations would be made with a sum contrast parameterization, as given in Table 22.5. Fortunately, the significance tests work quite similarly in this example. The intercept is close to zero and not significant, meaning that *was* and *were* are similarly common in the data. The person/number factor has a large effect, in the direction of decreased *was* in the plural or second person context; this is the standard pattern, and therefore, expected. Negative polarity shows increased use of *was*, as does female gender; this is evidence of a non-standard pattern in the distribution of *was/were*. Age has no significant effect, so we cannot link non-standard *was/were* to inter-generational change.

In contrast to the Wald tests in Table 22.4 or 22.5, likelihood ratio tests test an entire set of contrasts at once and so are independent of that variable's parameterization. Table 22.6 presents the likelihood ratio tests for the models in both Table 22.4 and 22.5. The degrees of

Table 22.5 Logistic regression model of York *was/were* variation, with sum contrast parameterization.

	Estimate	Std. Error	z	p
Intercept	-0.02508	0.09857	-0.254	0.7991
Person/number	-2.51868	0.06048	-41.643	< 2e-16 ***
Polarity	0.92519	0.09948	9.301	< 2e-16 ***
Gender	0.29610	0.05846	5.065	4.08e-07 ***
Age 1	0.14825	0.07920	1.872	0.0612 .
Age 2	-0.03884	0.07749	-0.501	0.6162

Table 22.6 Likelihood ratio tests for parameter groups in Logistic regression model of York *was/were* variation.

	Degrees of Freedom	Deviance	P
Person/number	1	3218.0	< 0.0001
Polarity	1	68.4	< 0.0001
Gender	1	25.2	< 0.0001
Age	2	3.6	0.1653
Residual	83	304.4	< 0.0001

freedom column represents the total number of parameters needed for each categorical variable. The deviance is the difference in the G^2 statistic of two models (with and without the variable in question), which is chi-squared distributed at the degrees of freedom for that variable, yielding the associated p value. These tests compare closely to what we have seen in the Wald tests, though in general they tend to be more conservative.

The test of the model's residual deviance, like a chi-squared test for independence, tells us if there is variation that is not adequately explained by the parameters in (22.4). In Table 22.6, the test is significant and there are 83 degrees of freedom left, meaning that there should be many opportunities to discover meaningful predictor interactions in the data: our model is not yet a good model of the data. To see where the model might be improved, we generate the model's predictions and examine the fit of the predictions to the data. These are computed from (22.1) using the appropriate coding for the x variables, as in (22.5), for younger females in the first-and third-person singular positive polarity context. Here, we use the treatment contrasts, although the predictions are the same under both parameterizations.

$$\begin{aligned} \text{logit}(p_{\text{was}}) &= a + b_{1\&3\text{sg}} x_{1\&3\text{sg}} + b_{\text{neg}} x_{\text{neg}} + b_{\text{male}} x_{\text{male}} + b_{\text{older}} x_{\text{older}} + b_{\text{younger}} x_{\text{younger}} \quad (22.5) \\ &= -1.1742 + (5.0374) * 1 + (-1.8504) * 0 + (-0.5922)(-0.1871) * 0 + (-0.2577) * 1 \\ &= -1.1742 + 5.0374 - 0.2577 \\ &= 3.6055 \end{aligned}$$

Residual deviances (the error terms e in (22.1)) are computed from the difference of the predicted logit and the logit of the observed proportion in each cell. The model assumes that these deviances are normally distributed; if they are not, then the model is probably not appropriate to the data, and should be reconsidered. The usual diagnostic for this is a quantile-quantile normal or *QQ-normal plot*, illustrated in Figure 22.1, where the standardized deviance of the residuals is plotted against the theoretical quantile values obtained by predicting them from a standard normal distribution. If the normality assumption is good, then the vertical and horizontal scales will be similar, and the plotted points will be symmetrically distributed around the origin and fall close to the $y = x$ line (the dashed line in the plot). In Figure 22.1, these assumptions are clearly violated: the scale of the standardized residuals is much larger than predicted by the theoretical quantiles, and many points in both tails lie well off of the $y = x$ line. In other words, the residuals have an expanded variance, with much heavier tails than would be expected under a normal distribution. We should, therefore, be critical of this model.

Further inspecting the residuals may tell us what direction our model may need to be improved. One common diagnostic is a plot of residuals against fitted values, as illustrated in Figure 22.2. If residuals are normally distributed, there should be a relatively even spread across the entire range of fitted values. Any form of pattern in the distribution otherwise represents a potential violation of model assumptions. Categories, of course, limit the number of predicted values one can have, and these typically show up as bands of points at specific fitted values. In the fitted model, there are only 24 possible combinations of categories, and so no more than 24 fitted values. We examine the distribution of points to see if they appear to be normally distributed around these values.

While it is customary to plot all of the residuals in one plot, it can be hard to ascertain what predictors may be responsible for patterns in the residuals. For this reason, in Figure 22.2

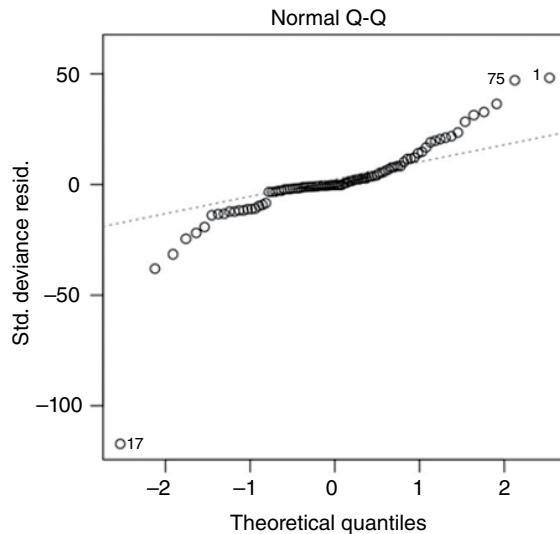


Figure 22.1 QQ-normal plot of model residuals.

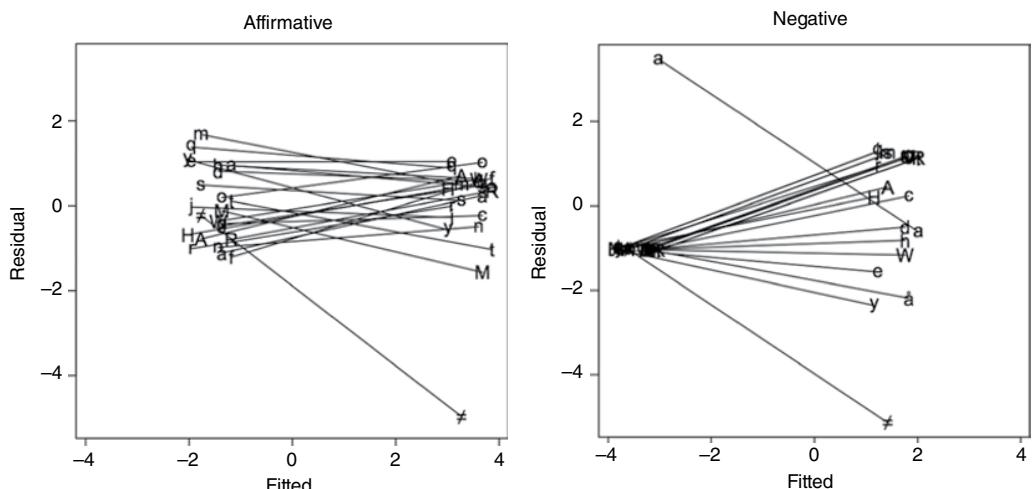


Figure 22.2 Plot of residuals versus fitted values for the model, separated by affirmative and negative polarity contexts; lines connect points belonging to the same individual.

we plot residuals separately for affirmative and negative polarity; both plots share the same x and y scales, and so can be easily compared. Also in both plots we use lines to connect first- and third-person singular with second-person or plural contexts belonging to the same speaker. The speaker labels are plotted at each end to further clarify the relationships.

In Figure 22.2, affirmative and negative polarity behave quite differently with respect to the model residuals: in negative polarity second-person or plural contexts, where *was* is predicted to be rarest, there are exactly two values of residual, whereas in affirmative contexts there is considerable variation in both person/number contexts. The residual for negative polarity, second-person or plural contexts for speaker *a* has the single datum of *was* used in this context.

A further pattern observable in Figure 22.2 is that speakers vary considerably more than the groups to which they belong. For example, within the affirmative second person or plural contexts, male speakers are spread over a broader range of residuals than their fitted values are separated by age. In comparison to the first- and third-person singular contexts, there is quite a bit of variation in the size and direction of the residuals (as seen in the slopes of the lines) this suggests that there is a large amount of inter individual variation, both in overall propensity to use *was* and in the effect of person/number, which is not accounted for in the model in (22.4).

The criticisms we have just made of the model in (22.4) suggest various directions one might attempt to modify and improve the model. The G^2 test of the residual suggests that there is unused information relevant to predicting the distribution of *was*. The principal information we have not used is the individual speaker identities, and we should examine if those can be incorporated into the model somehow. The residuals also draw our attention to the general lack of information about negative polarity contexts, suggesting that these contexts, although judged “significant” by the Wald and G^2 tests, offer insufficient data to support the estimated parameter value. Proceeding prudently would require excluding these instances from further analysis.

Using individual speakers alongside gender and age results in a problem of multicollinearity, that is, if speaker, gender, and age are coded separately, three of the contrast variables will be linearly dependent on the remaining set. This is because the grouping variables (speaker and age) code speakers as well, and so are redundant with them.¹⁴ When such variables are used in an analysis, the corresponding parameters are *aliased*, meaning they are excluded from the model (they are set to zero). Mathematically this is not a problem, but interpretively it may be, because one will not be able to test for significance or interpret either the full set of differences among speakers, or possibly worse, the intended group parameters.

For demonstration purposes, we adopt an approach to this problem that remains within the general framework of (22.1) by manually coding contrasts for speaker to include variables corresponding to gender and age contrasts as in Table 22.7 (see also Paolillo 2013).¹⁵ Basically, the speakers are simply coded with the sum contrasts of the groups, as can be seen by comparison with Tables 22.1 and 22.2; we rely on a feature of the `contrasts()` function in R to complete the full set of 22 contrast variables. The model in (22.6) may then be fit to the data for the affirmative polarity contexts.

$$was \sim 1 + speaker + persnum \quad (22.6)$$

This model shows considerably better conformance to the model assumptions, as we can see from the diagnostic plots in Figure 22.3. The QQ-normal plot of residuals is a much better fit to normal, although the extreme high- and low-valued residuals still deviate from normal. The residual by fitted value plot confirms this improved conformance to model assumptions, although the speaker ≠ from the middle age group of males is still an extreme outlier. One might wish to drop this speaker from the analysis, seeking some other reason for this individual’s exceptional behavior (a low rate of *was* in second person or plural contexts); one needs far greater familiarity with the data collection than presently available to us to justify this move. We can, however, feel safer in the interpretation of the estimates from this analysis, presented in Table 22.8, than in those of Tables 22.4 and 22.5.

We are uninterested in the speaker parameters Spkr 1-19 (for the contrasts provided automatically), because they only serve to collectively distinguish speakers from one another, and are not interpretable as representing individual speakers. The person/number parameter is negative and significant, indicating lower *was* use in second person or plural contexts.

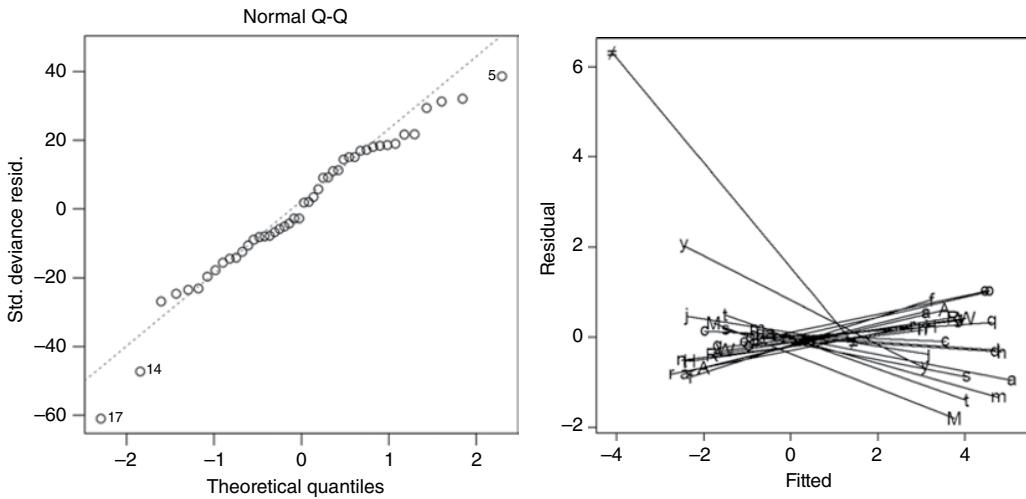


Figure 22.3 Model diagnostics for the model in (22.5) fit to data from the affirmative polarity contexts of Table 22.2.

Table 22.7 Speaker contrasts incorporating gender and age contrasts.

Speaker	Gender	Age 1	Age 2	Speaker	Gender	Age 1	Age 2
W	1	-1	-1	y	-1	-1	-1
h	1	-1	-1	H	-1	-1	-1
n	1	-1	-1	A	-1	1	0
d	1	-1	-1	≠	-1	1	0
f	1	1	0	s	-1	1	0
A	1	1	0	m	-1	1	0
R	1	1	0	q	-1	0	1
t	1	1	0	e	-1	0	1
c	1	0	1	r	-1	0	1
g	1	0	1	j	-1	0	1
å	1	0	1				
o	1	0	1				
M	1	0	1				

Gender is significant and positive, indicating a greater use of *was* by females. The two age parameters are both not significant. We have largely confirmed the structure of the earlier model, with the exception that inter-speaker variation is now an explicit part of our model and negative polarity is excluded from consideration. Our conclusions are conditioned on affirmative polarity, and do not generalize to negative polarity.

The analysis of the deviance of this model is presented in Table 22.9. Here, gender and age are tested within the speaker parameter group, which is significant. The residual is much smaller than previously, both because of the exclusion of the negative polarity contexts from the analysis and because of the additional variation absorbed by individual speaker variation. It is nonetheless significant at the residual 22 degrees of freedom, indicating that there possibly still is variation remaining to be explained.

Table 22.8 Parameter values for the model York *was/were* variation in (22.5).

	<i>Estimate</i>	<i>Std. Error</i>	<i>z</i>	<i>p</i>	
(Intercept)	0.973659	0.067261	14.476	<2.00E-16	***
Person/number	-2.769163	0.075695	-36.583	<2.00E-16	***
Gender	0.223886	0.065953	3.395	0.000687	***
Age 1	-0.008173	0.090509	-0.09	0.928053	
Age 2	0.004753	0.089814	0.053	0.957799	
Spkr 1	-0.117103	0.275902	-0.424	0.671248	
Spkr 2	0.894605	0.360003	2.485	0.012955	*
Spkr 3	1.720537	0.345519	4.98	6.37E-07	***
Spkr 4	-0.772995	0.311779	-2.479	0.013164	*
Spkr 5	0.189998	0.299169	0.635	5.25E-01	
Spkr 6	1.047346	0.327861	3.194	0.001401	**
Spkr 7	0.335706	0.294112	1.141	0.253694	
Spkr 8	0.405616	0.29218	1.388	0.165065	
Spkr 9	1.680185	0.276109	6.085	1.16E-09	***
Spkr 10	0.090571	0.328321	0.276	0.782655	
Spkr 11	-0.77612	0.351746	-2.206	0.02735	*
Spkr 12	0.898435	0.34762	2.585	0.009751	**
Spkr 13	1.870994	0.320582	5.836	5.34E-09	***
Spkr 14	0.04363	0.252231	0.173	0.862669	
Spkr 15	-0.282766	0.349287	-0.81	0.418197	
Spkr 16	0.949376	0.312195	3.041	0.002358	**
Spkr 17	0.021135	0.31171	0.068	9.46E-01	
Spkr 18	0.205031	0.272605	0.752	0.451981	
Spkr 19	0.205117	0.302923	0.677	0.498326	

Table 22.9 Analysis of the deviances in the model of (22.5) and Table 22.8.

	<i>Degrees of Freedom</i>	<i>Deviance</i>	<i>P</i>
Person/number	1	3095.2	<0.0001
Speaker	22	155.3	<0.0001
Residual	22	73.0	<0.0001

Whether this model is satisfactory depends in part on our interpretive goals. If we had hoped to identify a pattern of variation in *was* use that distinguished different groups of people, then we have done that with gender: females show a greater use of the non-standard *was* than do males. Such a goal is often compatible with the goal of identifying a linguistic marker that distinguishes dialects. As long as we can show a distinction in the rate of variation among speakers belonging to two or more dialect groups, then we can claim to have identified such a marker.

If our hope was to confirm a specific hypothesis for *was/were* variation in negative polarity sentences (Tagliamonte 1998), we have not been able to do that because of data insufficiency. If this question is genuinely important, there is no way to address it other than to obtain more data. Likewise, we are unable to make strong claims about the distribution of *was/were* over age, or to reveal more of the grammatical patterning of *was/were* in York English

without more data. Although other grammatical variables and speakers were included in the original data set used in Tagliamonte (1998), the grammatical variables were pared down to those thought most clearly justified, and only speakers with a sufficient data were included. Estimation of logistic regression models is generally not considered reliable with fewer than 100 observations (Long 1997, 54), and since individual speaker was regarded as a necessary level of analysis (as I hope is demonstrated above), only speakers with at least 100 observations each were used, resulting in the 23 speakers of Table 22.2. Even with the research questions whittled down to two dimensions of linguistic context and two regarding social context, we find that this is not a sufficient amount of data to justify a full analysis.

These observations illustrate the delicate balancing act required in logistic regression analysis of linguistic data. One naturally seeks a complete analysis of the data, which requires using variables that distinguish specific contexts of observation in ways that are theoretically relevant, such as the origin of a change in a specific negative polarity context, or in a specific social or demographic group. The same variables raise challenges of data imbalance or insufficiency, which may violate the assumptions of the regression model. Since inferences from the model are not valid when its assumptions are not met, analysis must proceed carefully, checking that use of the model is justified, and that the remaining data suffice to answer the questions asked of it. Argumentation of this sort, which is crucial to the proper application of the model, is often omitted when regression models are presented.

22.6 Summary

We have discussed here a framework for the logistic regression analysis of linguistic variables. This framework allows fitting and testing hypotheses about variation that are conditioned by linguistic and social factors, and which permit an account of individual variation. The hypotheses that characterize our understanding of language, its variation and change are far from simple. For example, the hypothesis “Group g shows increased use of a particular linguistic variant in a particular environment across generations,” whereas typical in variation studies, is actually a complex statement representing a three-way group, generation, and context interaction. If this hypothesis must hold in the context of individual variation, as well as possible lexical item-specific variation (cf. Baayen *et al.* 2008), then this three-way interaction may need to be tested in a model with crossed effects for speaker and lexical item. The research design and data gathering considerations for such a study are considerable. Properly understanding the nature of the question and its relation to hypothesis tests in a statistical model framework is the first step toward assessing the validity of the hypothesis.

NOTES

1 Both terms suggest experimental research, but are used with other designs.

2 This subset is also treated in Paolillo (2013); a different subset of the same data was analyzed in Tagliamonte and Baayen (2012).

3 An alternative arrangement treats *was/were* as an additional variable and has 184 cells with a count of observations in each.

4 This particular example does not include a geographic or regional component. Questions of this nature take the same form as the treatment of gender and age in this example.

5 One could also collapse the grammatical context variable, but it is less clear how to interpret the observed relationships if this is done.

6 This example is computed without Yates’ correction for two-by-two tables (Woods *et al.* 1986, 146–147).

- 7 Reliance on opportunistic rather than representative sampling in variationist studies favors logistic regression. Log-linear modeling and logistic regression are often available together in statistical software.
- 8 The degrees of freedom used in the chi-squared test of the model of independence is the residual degrees of freedom.
- 9 For m significance tests, the Bonferroni correction is to use p/m as the criterion.
- 10 The converse does not hold: failing to find significance does not license one to dismiss a possible relationship from future consideration; doing so is an independence claim and requires substantive justification (Paolillo 2011).
- 11 Log-linear modeling may be used to specify and estimate a model equivalent to polytomous logistic regression, when other options are not available.
- 12 This sense of interaction needs to be distinguished from the sense involved in “constraint interaction,” especially when linguistic constraints are intended: linguistic constraint interaction need not be modeled by interaction effects (Paolillo 2002, 2011).
- 13 Crawley (2002) treats S-plus, a close relative of R, and has more material on introductory statistics and research design, whereas Crawley (2007) focuses on R and advanced statistical techniques.
- 14 Multi-collinearity may also arise from properties of the sample, possibly distorting the parameter estimates; care must be taken to avoid it (see Fox 1997, 354–363 for recommendations and discussion). The usual diagnostic is a correlation matrix of the x variables in the model, with high correlations among variables indicating problematic multi-collinearity.
- 15 This is equivalent to using speaker as a blocking variable (Bates 2010), a technique closely related to mixed-effects modeling (Baayen *et al.* 2008; Bates 2010; Jaeger 2009; Johnson 2009; Tagliamonte and Baayen 2012), which addresses multi-collinearity by introducing additional information among the predictors (Fox 1997, 362; Gelman and Hill 2007, 393).

REFERENCES

- Agresti, Alan. 1990. *Categorical Data Analysis*. New York: Wiley.
- Agresti, Alan. 1996. *An Introduction to Categorical Data Analysis*. New York: Wiley.
- Baayen, R. Harald. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. New York: Cambridge University Press.
- Baayen, R. Harald, Davidson, Douglas J., and Bates, Douglas M. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59.4: 390–412.
- Bates, Douglas J. 2010. lme4: Mixed-effects modeling with R (draft). To appear, from Springer. Downloaded from <http://lme4.r-forge.r-project.org/lMMwR/lrgprt.pdf>, accessed May 21, 2015.
- Bishop, Yvonne M.M., Fienberg, Stephen E., and Holland, Paul W. 1975. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.
- Crawley, Michael J. 2002. *Statistical Computing: An Introduction to Data Analysis Using S-Plus*. New York: Wiley.
- Crawley, Michael J. 2007. *The R Book*. New York: Wiley.
- Fox, John. 1997. *Applied Regression Analysis, Linear Models, and Related Methods*. Thousand Oaks, CA: Sage.
- Gelman, Andrew, and Hill, Jennifer. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, UK: Cambridge University Press.
- Gorman, Kyle, and Johnson, Daniel Ezra. 2013. Quantitative analysis. In R. Bayley, R. Cameron and C. Lucas, eds., *The Oxford Handbook of Sociolinguistics*, 214–240. Oxford University Press.
- Gries, Stefan Th. 2009. *Quantitative Corpus Linguistics with R: A Practical Introduction*. New York: Routledge.
- Jaeger, Florian. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59: 434–446.
- Johnson, Daniel Ezra. 2009. Getting off the GoldVarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass*, 3: 359–383.
- Labov, William. 1969. Contraction, deletion and the inherent variability of the English copula. *Language*, 45: 715–762.

- Long, J. Scott. 1997. *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.
- McCullagh, Peter, and Nelder, James A. 1989. *Generalized Linear Models*, 2nd Edition. Boca Raton, FL: Chapman & Hall/CRC.
- Paolillo, John C. 2002. *Analyzing Linguistic Variation: Statistical Models and Methods*. Stanford, CA: CSLI Publications.
- Paolillo, John C. 2011. Independence claims in linguistics. *Language Variation and Change*, 23(02): 257–274.
- Paolillo, John C. 2013. Individual effects in variation analysis: Model, software and research design. *Language Variation and Change*, 25: 89–118.
- R Development Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>
- Rousseau, Pascale. 1989. A versatile program for the analysis of sociolinguistic data. In Fasold, R.; and Schiffrin, D., Eds. *Language Change and Variation*. Amsterdam: Benjamins. 359–409.
- Sankoff, David, and Rousseau, Pascale. 1989. Statistical evidence for rule ordering. *Language Variation and Change*, 1: 1–18.
- Sankoff, David, Tagliamonte, Sali, and Smith, Eric. 2012. *Goldvarb Lion: A multivariate analysis application*. Department of Linguistics, University of Toronto. <http://individual.utoronto.ca/tagliamonte/goldvarb.html>
- Sigley, Robert. 2003. The importance of interaction effects. *Language Variation and Change*, 15: 227–253.
- Tagliamonte, Sali. 1998. Was/were variation across the generations: View from the city of York. *Language Variation and Change*, 10: 153–191.
- Tagliamonte, Sali, and Baayen, R. Harald. 2012. Models, forests and trees of York English: Was/were variation as a case study of statistical practice. *Language Variation and Change*, 24: 135–178.
- Woods, Anthony, Fletcher, Paul, and Hughes, Arthur. 1986. *Statistics in Language Studies*. New York: Cambridge University Press.

23 Statistics for Aggregate Variationist Analyses

JOHN NERBONNE AND MARTIJN WIELING

23.1 Aggregation and Perspectives from the Aggregate

Dialect geographers have viewed the distribution of most individual linguistic features (pronunciations, allomorphy, or lexical choice) as complex, and moreover, as non-overlapping. Features often overlap very poorly, resulting in isoglosses that do not “bundle.” Nerbonne (2009) maps nine features often discussed in German dialectology, showing how poorly they overlap in detail, but noting that they mostly distinguish the north from the south. Bloomfield’s (1933) discussion of dialects came to a similar conclusion, leading him to quote Grimm’s (1819: IV) famous dictum that “every word has its own history.” The view has been disputed, as Schirmunksi’s (1962: 78ff) account of Wenker’s reception among the neo-grammarians documents, but it is now accepted. The conclusion that individual features have very noisy geographic distributions has been seen to threaten the dialectological enterprise. Gaston Paris concluded on this basis that “there was no geography of *dialects*, only of individual linguistic features” (quoted by Goebl, Ch. 7).

Séguy (1971, 1973) introduced the term “dialectometry” as he took the liberating step of AGGREGATING over large sets of features (over 400) to then examine how well geographic neighbors agreed; poor aggregate agreements indicated dialect boundaries. The simple step of examining aggregate distributions proved to be enlightening. Goebl (1982) introduced local perspectives on the larger dialect space, examining the distribution of linguistic differences (or equivalently, similarities) from each site in his (Italian) data collection, pointing out *inter alia* that sites with skewed distributions might be transition zones. We also refer to differences as distances, noting that the measures used in dialectometry generally satisfy the conditions imposed on mathematical distances (symmetry, zero between identical elements, and the “triangle inequality”: $d(a,b) \leq d(a,c) + d(c,b)$, for all c). Seguy (1971) examined the distribution of aggregate differences (in Gascogne) as a function of geography, displaying a sub-linear curve, which Nerbonne (2010) argues contradicts Trudgill’s (1974) gravity theory of dialect divergence.¹ These early works motivate the aggregate perspective.

There are various ways of probing linguistic data to obtain a measure of difference between sites (or samples, in case we are interested in non-geographic differences as well²), for example, measuring the percentage of concepts that is realized in different ways lexically, or by using the percentage of syntactic features that are realized differently (see Ch. 20, this volume, for more sophisticated methods). Equivalently, we may always measure similarities and convert these to differences. In the interest of brevity we shall abstract away from the

concrete details of how the distances we analyze below are obtained. The presentation below proceeds from the assumption that we have a table of aggregate differences between pairs of sites in our data. We will use a small example from Heeringa (2004: 146) in the following section.

23.2 Clustering

A traditional question in dialect geography is whether there are DIALECT AREAS, that is, regions within which variation is low, but which are relatively distinct with respect to varieties outside the region. There are many CLUSTERING techniques that search for groups in data (Kaufmann and Rosseeuw 2005; Manning *et al.* 2008: Ch.16–17); we focus here on the ones popular in dialectology. All the techniques we examine search for groups in the linguistic data without reference to location.³ If, on the basis of aggregate differences, a group is identified and confirmed, and if it indeed constitutes a geographic region, then this is already a modest success.

Some clustering algorithms, such as k-means clustering (Manning *et al.* 2008: 515ff), aim to partition the data, that is, determine groups such that each variety (or sampling point) belongs to exactly one. These “flat” algorithms thus assume that there is no hierarchical structure in the grouping, where a given variety might be assigned not only to Vologda, but also to North Russian, which is part of Russian, which in turn finally belongs to East Slavic (see Zhobov and Alexander, Ch. 31, this volume). It is uncontroversial, however, that dialect groups may be hierarchically structured, which leads to a focus on HIERARCHICAL CLUSTERING ALGORITHMS, which we will concentrate on. A cross-cutting distinction separates algorithms which work bottom-up, or agglomeratively, from those which work top-down, or divisively. We focus on agglomerative algorithms, which have enjoyed the most dialectological attention, but we will return to divisive algorithms at the end of this section.

23.2.1 Hierarchical Agglomerative Clustering

Johnson (1967) realized that different hierarchical agglomerative algorithms can be characterized in similar ways. They all begin with a sample \times sample table of aggregate differences such as that in Table 23.1 (below). They all then seek the two (groups of) sites for which the differences are minimal and fuse these two to create a cluster. The table is then revised so that the two minimally different groups are eliminated while the result of fusing them is included. This leaves the table with the difference values of several cells unspecified (see Table 23.2).

Table 23.1 An example table of aggregate differences between Dutch sites. The cells in the diagonal are empty, with the implicit value zero since there are no differences between a site and itself. The cells below and to the left of the diagonal are blank, because these would be the same as those above and to the right, since the differences between sites *a* and *b* are the same as those between *b* and *a*. This also means that the row for the last site in the table can be empty. We omit this row as the example is developed further.

Grouw	Haarlem	Delft	Hattem	Lochem
Grouw	42	44	46	47
Haarlem		16	36	38
Delft			38	40
Hattem				21
Lochem				

Table 23.2 The (partial) table after fusing Haarlem and Delft but before determining the differences between the recently fused element and the other elements. The missing values are signaled by question marks, and the following value is the mean of the distances from two components of the fusion to the other site. From Heeringa (2004: 147).

	Grouw	Haarlem and Delft	Hattem	Lochem
Grouw		? (43)	46	47
Haarlem and Delft			? (37)	? (39)
Hattem				21

Once we have obtained a filled-in table, we are in a position to iterate, again choosing the closest (groups of) sites, fusing them, and assign the necessary distance to the new elements (and eliminating the components of the fusion). Johnson (1967) showed that several sorts of hierarchical agglomerative clustering approaches could be characterized by the function used to determine the new distances in the step immediately following the fusion. One may use the arithmetic mean between the two components (and shown in parentheses in Table 23.2), which is referred to as the “Unweighted Pair-Group Method using Arithmetic averages” (UPGMA). Heeringa (2004) also discusses a version which uses a mean weighted by group sizes (WPGMA), and a pair of methods ((un-)weighted by group size) which determine a centroid in the abstract space of differences. In addition, one may simply assign the new element either the smallest difference value available (among all the pairs with one element in the new group and one in the old one), which is referred to as “nearest neighbor clustering” or “single-link,” or assign the new element the largest distance, which is referred to as “furthest neighbor” or “complete-link.” Finally, Ward’s method proceeds from the insight that assigning a single distance to the elements newly fused introduces a kind of “error” in treating the fused elements as the same. It then is designed to minimize the error. Prokić and Nerbonne (2008) review other popular techniques.

Figure 23.1 shows the output of clustering, a dendrogram, that is, a tree with the varieties as leaves, which are joined to form more substantial branches reflecting the fusion process. Dendograms are popular not only for illustrating groups, but also for showing COPHENETIC DISTANCES. The cophenetic distance between any two sites is their common distance to the first encompassing node. In Figure 23.1 the cophenetic distance between Hattem and Lochem is 21.

23.2.2 Stochastic Clustering

All the popular hierarchical agglomerative algorithms suffer from instability, meaning that small differences in the input table can sometimes lead to large differences in the output dendrogram, occasionally very large (Prokić and Nerbonne 2008).

In order to overcome the inherent instability in clustering, two approaches have been used, bootstrap clustering and noisy clustering (Nerbonne *et al.* 2008), both of which add a STOCHASTIC element.⁴ In bootstrap clustering, clustering is repeated a fixed number of times, usually 100 or 1,000 times, whereas the choice of elements in the aggregate sample is allowed to vary. A common choice is to fix the number of elements in the aggregate at the total number available, and then to choose the elements randomly with replacement. If one begins with 200 words, then sample size is fixed at 200. When selecting elements with replacement, one may select the same element more than once, which will then force other elements to be omitted. Another option is to repeat clustering adding different small amounts of noise (e.g., 0.3 standard deviations) to the distance matrix at each iteration. In either case the result is a

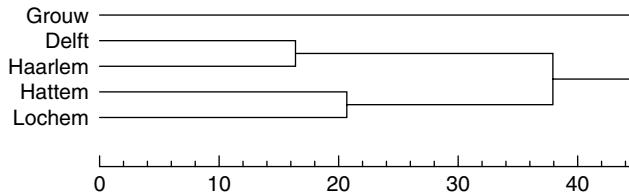


Figure 23.1 A dendrogram displaying the results of UPGMA clustering on the distance matrix in Table 23.1. From Heeringa (2004: 147). The horizontal distance from the leaves on the left to a branching point shows the cophenetic distance.

dendrogram where each internal node is associated with a percentage indicating how often the sites below the node (its leaves) emerged during the stochastic process. One may be fairly confident of clusters that emerge 90% or more of the time.

23.2.3 Other Clustering Techniques

Since dialectologists often present their results in maps of dialect areas, and since clustering produces groups that normally project onto areas, clustering thus facilitates comparison to earlier work.

Before closing this section, we note that there are perhaps hundreds of alternative clustering algorithms that have not yet been used on dialect data (see the NIPS conferences, <http://nips.cc>), meaning that there is clearly room for further experimentation. Divisive algorithms begin with the entire set of samples, seeking a split that leaves the subgroups as internally similar as possible. Manning *et al.* (2008: Ch. 17.6) claim that divisive algorithms are computationally more effective and that they may also be more accurate, since they base their decisions on the entire data set, not just on local evidence (pairs of varieties or groups of varieties). Given their potential advantages and the scant attention they have received in dialectology, more work on this topic would be desirable.

Two principles are important in further experiments. First the potentially hierarchical structure of dialect relations should not be ruled out (as in k-means clustering, see Manning *et al.* 2008: 515ff), and second, the algorithm should be required to assign a reliability to the clusters it proposes.

23.3 Dimension Reduction

23.3.1 Multidimensional Scaling

Multi-dimensional scaling (MDS, Kruskal and Wish 1978) was introduced to dialectology by Embleton (1993). It inputs an input distance matrix such as Table 23.1 and assigns coordinates to each element in a small number of dimensions. Given the MDS coordinates we may derive assigned distances using the Euclidean formula, and the “dimension reduction” is successful to the degree that the derived distances agree with the input distances. This success is indicated by the STRESS in the solution, where less stress is better, or by the correlation of the input distances with the MDS-derived distances, where greater correlation coefficients are naturally preferred. MDS analyses must be accompanied by one of these two numbers if they are to be published. One should also attempt to interpret the dimensions, and in the case of dialectological applications we prefer to see both geographic interpretations (how are the dimensions projected onto a map) and linguistic ones (what linguistic features do these sites share). See Wieling *et al.* (2007: Fig. 6), Prokić (2010: §3.5.1) and Nerbonne (2010a: Map 2405) for examples.

We also need to ask how many dimensions to use in a solution. Each additional dimension will reduce stress (but less than all the previous dimensions) and improve the correlation with the input matrix, so a common way to determine the optimal number of dimensions to be used is to plot, for example, stress as a function of the number of dimensions in what is called a scree plot.

Since each additional dimension reduces stress less than previous ones, there will always be a point where the curve in the scree plot begins to flatten out, so that additional dimensions in the flat portion of the curve no longer account for much variance (Johnson 2009: 209), indicating that further dimensions may be ignored. A further question concerns whether one may use metric versions of MDS or whether one should stick to a non-metric version, given the categorical nature of linguistic data. But since it is fine to treat large aggregates as numerical data, that is, metrically, we have no compunction about using the (simpler) metric version.

Applications of MDS to dialect data typically represent the data well in two or maximally three dimensions, and this has led to a popular innovation in dialect mapping where each dimension is assigned a color—usually red, green, or blue—and each site is assigned a color intensity corresponding to its coordinate in the MDS solution (Nerbonne *et al.* 1999), arguably the first representation of dialect continua based on exact techniques.

MDS does not partition dialect sites into non-overlapping dialect areas, and it also does not suffer from the instability we noted in non-stochastic clustering routines. Moreover, it analyzes the same sort of distance matrix, which is input to clustering. This means that it can be used to examine clustering results in more detail.

23.3.2 Principal Component Analysis and Factor Analysis

PRINCIPAL COMPONENT ANALYSIS (PCA) and FACTOR ANALYSIS (FA) are two related dimension-reducing techniques, which input not distance matrices, but rather matrices of sites and numeric features. They differ in that PCA attempts to account for all the variance in the matrix, while FA is more modest, taking aim only at the variance that is shared among the input variables. This means that noise and variables that share no variance with others are ignored in FA.

Following Tabachnik and Fidell (2001: Ch.13), we shall (mostly) discuss PCA and FA together, although we will naturally give some examples of each. In particular, we use the term FACTOR to refer both to PCA components and FA factors. This eliminates some awkward formulations.

The need to provide numeric features naturally entails representing linguistic data numerically. This may be straightforward, as when Labov (2001: 286ff, 345ff) applies PCA to formant frequencies, but Shackleton (2010, Ch. 3 & App. B) translates the vowels in *The Pronunciation of English in the Atlantic States* (Kurath and McDavid 1961) and *The Dialect Structure of Southern England* (Kurath and Lowman 1970) into numerical values on a set of features in order to apply PCA. This involves coding each token of a vowel numerically, so that the vowel in *wool* was <5,3,2,1>, representing maximal height, maximal backness, maximal roundness, and non-rhoticity. See Shackleton (2010: 188, Table B.1). Using FA, on the other hand, Clopper and Paolillo (2006) use formant frequencies and vowel duration for fourteen American vowels, whereas Nerbonne (2006) translates transcribed vowels into (numeric) feature vectors somewhat as Shackleton does. Grieve *et al.* (2011) identify lexical alternatives and then use their relative frequencies in an analysis using FA, but see Grieve (this volume) as well for several other crucial steps in that analysis. Pickl (2013) also uses relative lexical frequencies as input to FA.

The result of applying the analysis is a set of factors smaller than the original set of variables. The goal is to interpret one's data not on the basis of, say, 20 vowels, or 200 or more phoneme tokens, but rather in terms of a much smaller number of factors in the solution. We note two

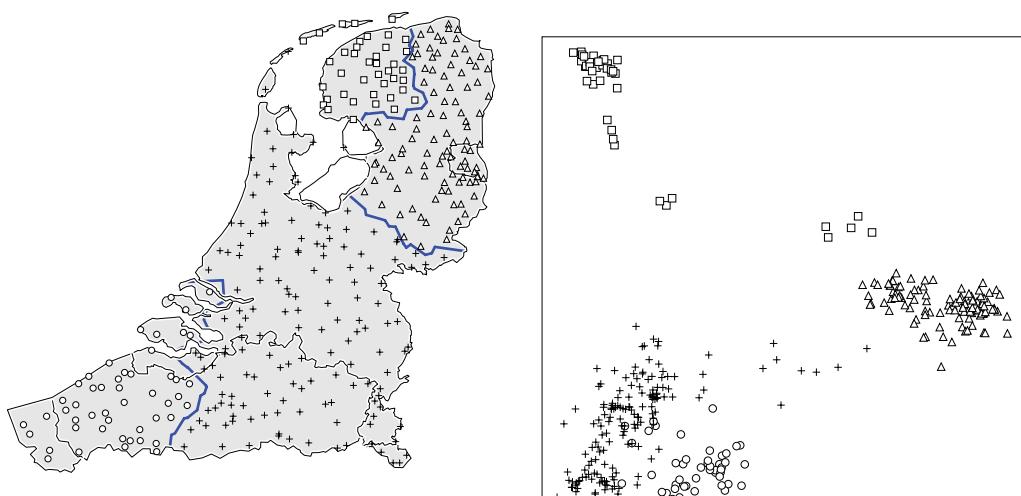


Figure 23.2 Left: a map of the Netherlands showing the largest four clusters obtained using Ward's method. We shall discuss the Frisian area in the Northwest, the Lower Saxon area in the Northeast, the West Flemish (including Zealand) area in the Southwest, and the large Franconian area in the middle. On the right, a projection of the same data into two dimensions using MDS ($r=0.76$). The MDS graph shows that there is an “archipelago” of Frisian sites (boxes) that are quite similar to Lower Saxon, and that Lower Saxon is otherwise fairly distinct, whereas the West Flemish and Franconian sites are less well distinguished. The MDS perspective adds to the clustering. From Gabmap, www.gabmap.nl.

further interesting outputs of PCA and FA. On the one hand we obtain for each factor and variable, the **LOADING**, that is, the degree to which the two correlate, which facilitates in interpreting the variable. If a factor correlates highly with [i,i,e,ɛ] and [æ] (which therefore have high loadings for this factor), but not with other vowels, the factor may be interpreted straightforwardly, giving dimension reduction the potential to contribute to the deeper linguistic analysis of aggregate analyses. Tabachnik and Fidell (2001) suggest the loadings should have values greater than 0.32 if one is to interpret them. On the other hand, PCA and FA also provide **Scores** for each factor and site in the input matrix, contributing to the geographical interpretation.

Turning to the question of how many factors to retain in analyses, several considerations may play a role. First, just as in MDS, we want to interpret the factors we decide to retain in the solution. The geographic or social interpretation is important, but given that we analyze individual variables jointly in PCA and FA (via factor loadings), it is most intriguing to seek linguistic interpretation, especially when it suggests that a more abstract level of linguistic structure might be influential. Second, each factor in PCA and FA is associated with an **EIGENVALUE** derived from the input matrix, which represents variance, in particular where an eigenvalue of 1.0 is roughly the variance associated with a single case (site or speaker). This is the source of a common rule of thumb to disregard eigenvalues less than 1.0.⁵ Scree plots are used for PCA and FA, just as in MDS, and in a similar way. Since factors are ordered in importance, one uses the scree plot to determine which initial sequence of factors is to be used by examining the decreasing sequence of eigenvalues in a given solution (Tabachnik and Fidell 2001: 621). A third desideratum in choosing which factors to retain is the wish to explain a good deal of the variance, where the rule of thumb is to strive for 70% explained variance. But this can clash with the wish to identify and interpret the factors and the total variance explained.

Clopper and Paolillo (2006) show that two factors account for 73% of the variance in their vowel pronunciation data, whereas Nerbonne (2006), using hundreds of vowel tokens as

variables, was able to explain only 35% of variance using three factors. Leinonen (2010: 109) can account for 60.1% of the variance using 10 factors, and Pickl (2013) retains 20 factors in order to reach 59.3% explained variance.

Before closing the discussion of PCA and FA, we note that the factors may be ROTATED in order to improve interpretability (Tabachnik and Fidell 2001: Ch. 13). It is most common to use a rotation in which the first factor explains a maximum of variance, and each successive factor accounts for a maximum of remaining variance. Geometrically, each successive factor is orthogonal to the others in this so-called VARIMAX rotation. Alternatively, one may choose to use OBLIQUE rotations, especially when this might facilitate interpretation, in which case there will be overlap in the variance claimed by different factors. All of the dialectology studies examined have opted for the varimax rotation.

FA may also be used in a hypothesis-testing fashion, but since this “confirmatory” FA has not been used much in dialectology, it will not be presented here. Due to the opportunity to hypothesize about factors, FA is also preferred when there is some theory relating the factors under study (Brown 2009). See, however, Pickl (2013), who uses confirmatory FA to test whether Bach’s (1950) conjecture that the areal extent of the use of a form correlates with its frequency. PCA is always used in an exploratory fashion.

Leinonen (2008) shows the advantage of deciding on PCA and FA on a case-by-case basis: she first uses PCA on the band-filtered vowel spectra of 19 vowels of over 1,000 Swedish speakers to obtain a representation of vowel quality, and found that PC1 and PC2 were readily interpretable as the first and second formants.⁶ She ignores the first filter and it turns out that she can compensate for sex differences by ignoring the women’s second filter while performing separate PCAs for men and women. In a second step, Leinonen uses FA to uncover more abstract dimensions in Swedish geographic and social variation, for example, a factor that could be interpreted as lower vowel height for the two long vowels [œ:] and [æ:] in positions before /r/. Leinonen’s work suggests that one should not regard PCA and FA as simple alternatives but rather as techniques for specific purposes.

23.3.3 Related Techniques

Cichocki (2006) experiments with correspondence analysis (CA), first developed as a counterpart to PCA for categorical data by Benzécri (1992). Just as in PCA the input is a matrix of sites × data, but the data may be categorical in the case of CA. Uiboaed *et al.* (2013) use CA to analyze the geographic distribution of lexically specific constructions (so-called “collostructions”) in Estonian.

Leino and Hyvönen (2008) experiment with a range of more advanced “components” and conclude notably “if in doubt, start with factor analysis” (p. 186). Prokić and Van de Cruys (2010) experiment with a three dimensional matrix (site × site × 20 phonetic-correspondences), which they reduce to a set of most important correspondences using the tensor-reduction techniques PARAFAC. The interesting technique requires a substantial amount of data.

Wieling and Nerbonne (2011) report on bipartite spectral graph clustering (BSGC), which, despite its name, is less indebted to clustering than to the dimension-reducing techniques reported on in this section. It combines dimension reduction with techniques from graph theory to provide a similar sort of result to PCA and FA, namely a sketch of the affinities of cases (sites) with one another and also of affinities between cases and linguistic features. Space limitations prevent a longer discussion here, but we should note that Wieling and Nerbonne (2011) proposed a numeric evaluation of the features identified by BSGC, which involves measuring how representative they are within an area and also how distinctive they are with respect to other areas. Prokić *et al.* (2012) generalize this work to include numeric features such as edit distance.

23.4 Regression Models

REGRESSION analyses seek to explain or predict a single dependent variable on the basis of one or more independent variables. In textbook cases weight is predicted on the basis of height, or university success on the basis of high school success, aptitude and discipline. In such cases both the independent (predictor) variables and the dependent (or response) variable are numeric, but we are free to construe potential categorical predictors numerically, for example, as taking the values zero and one. LOGISTIC REGRESSION, in which categorical variables are predicted, is the subject of the previous chapter in this handbook.

As the bulk of dialectological work aims at characterizations of the relations among varieties and strives to characterize those both at an aggregate linguistic level and also with respect to the linguistic details involved, that is, the components of the aggregate, it is not surprising that regression, with its focus on a single dependent variable, has played a lesser role (than clustering or dimension-reducing techniques). But regression analyses have played a role in characterizing the relation between geographic distance and aggregate linguistic differences, and new regression techniques have been used in conjunction with aggregate analyses which will be presented in Sections 23.4.1 and 23.4.2.

Trudgill (1974) urged more attention for the theoretical question of how geography and demography influence linguistic variation, and his paper has sparked a stream of studies in the intervening years. Nerbonne and Heeringa (2007) studied the effect of distance and population size on aggregate pronunciation differences in Lower Saxon dialects using a regression analysis. Not to anyone's surprise, they found a robust relation, where more distant settlements had more different varieties than closer ones, but they also noted that the response variable (aggregate pronunciation differences) failed to have a linear relationship with geographic distance. Instead it was sub-linear, so that they were able to show a linear relation between the logarithm of geographic distance and linguistic distance. Nerbonne (2010) showed that the same sub-linear relation holds for pronunciation in six other language areas, namely American English on the Eastern seaboard, the entire Dutch area, Germany, Gabon Bantu, Norway and Bulgaria. Spruit *et al.* (2009) show that the same sub-linear relation holds for vocabulary, but perhaps not for syntax, where a linear model was marginally better. Szemrecanyi (2012) finds completely different results using morpho-syntactic frequencies in corpus data. Because Seguy's (1971) paper also graphed the influence of geography sub-linearly, Nerbonne (2010) proposed to call this SEGUY'S CURVE.

23.4.1 Mixed-Effects Models

In standard regression models, we assume that the independent variables are non-random, that is, fixed. The sex or gender of participants in a dialect survey is a clear example of a FIXED EFFECT, which is normally controlled for in survey design. Only two sexes are polled and any future work would use exactly these two, so that they are repeatable. In contrast, the words in a data set are assumed to be a random sample from a much larger population of potential words. If we repeated the survey, we might use new words. The choice of words is, therefore, a RANDOM EFFECT. Taking into account the structural variability associated with these random effects allows for generalizable results, with *p*-values which are not over-confident (Baayen *et al.* 2008). Regression models using both fixed and random effects are known as MIXED-EFFECTS MODELS (Pinheiro and Bates 2000).⁷

Keune *et al.* (2005) presented a very early use of mixed-effects models for language variation, in which the authors contrast the use of Dutch adjectives and adverbs using the suffix *-lijk* and its reduction in speech, and they treat the choice of words as a random variable.

In one of the studies reported, the dependent variable is the lexical frequency with which these words appear in newspapers in the Netherlands and Flanders with different registers (“quality,” national or regional). As the authors note, the mixed-effects analysis effectively builds a model for each individual word, so that one can note not only that there is a general tendency for *-lijk* to appear more frequently in the Netherlands and in more formal newspapers, but also that some words go against the grain, and moreover that that there is an interaction between register and country. In the study on phonological reduction using the *Corpus of Spoken Dutch* (<http://lands.let.ru.nl/cgn/>), it is shown that reduction is more common in the Netherlands than in Flanders, more common among men than women, more common when the word was predictable, and less common at the end of utterances. But some words go against the grain with respect to reduction as well. Keune and colleagues further note that the mixed-effects analysis has the welcome consequence that sociolinguists no longer need to identify alternatives to serve in variable rule analysis of the Varbrul sort (see Paolillo’s chapter, this volume). It becomes instead possible to examine a large number of linguistic phenomena simultaneously.

Tagliamonte and Baayen (2012) examine the *was/were* alternation in York, where *was* is often used in plural existential sentences. They analyze 300 tokens from only 40 individuals, and turn to mixed-effects modeling treating speakers as a random effect, eliminating problems connected with imbalanced numbers of tokens per individual. Their re-analysis shows that one influence on the choice between *was* and *were*, the polarity of the utterance (whether it is used in construction with negation), needed to be re-thought due to its interactions with other predictors.

The following section (23.4.2) discusses further studies in which mixed-effects analysis has been used in combination with generalized additive modeling. Baayen (2008: Ch. 7) is a step-by-step presentation of how to conduct mixed effects analyses in the R `lme4` package with several examples. Jäger (2008) presents mixed-effects logistic regression in a general comparison of regression designs vs. categorical designs. Johnson (2008) presents `Rbrul`, a package by the author, and a detailed comparison of how mixed-effects analyses compare to the popular logistic regression package, Varbrul (Sankoff and Labov 1979). Winter (2013) provides a tutorial in mixed-effects regression modeling for linguists.

Although mixed-effects models have been clearly gaining in popularity, Paolillo (2013) has criticized their use, arguing for including speakers as a fixed effect (rather than a random effect), since “the sample of individuals [may be] nonrandom” and modeling it as a random factor would be in error. His fixed-effects approach, however, is “limit[ing] the scope of inference” (Bolker *et al.* 2009), that is, undercutting the ability of the analysis to generalize from sample to population. We, therefore, recommend mixed-effects regression as the appropriate method to analyze linguistic data involving participants responding to multiple items.

23.4.2 Generalized Additive Modeling

In contrast to methods from geo-statistics, which focus mainly on identifying the aggregate geographical pattern in the data (see Grieve, this volume), another approach, generalized additive modeling (GAM, Hastie & Tibshirani 1990, Wood 2006), is able to *simultaneously* detect the aggregate geographical pattern, whereas also identifying the importance of other relevant social and lexical predictors.

A GAM is an extension of a generalized linear regression model. As noted above, in linear regression, the response variable is assumed to have a linear relationship with one or more predictor variables. The response variable must be numerical (such as pronunciation distance from a certain reference variety), whereas the predictor variables can be either numerical (such as the speaker’s age) or categorical (such as the speaker’s gender).

The generalized linear regression model is a generalization of the linear regression model in such a way that the response variable (transformed via a link function) has a linear relationship to the predictor variables (and may have an arbitrary distribution). For example, logistic regression is a form of a generalized linear regression model which uses the log-odds (logit) link function. Logistic regression (see Paolillo, this volume) is the appropriate form to analyze binary data and is frequently used in sociolinguistics (e.g., Tagliamonte and Baayen 2012).

A generalized additive model extends the generalized linear regression model by allowing the (possibly transformed) response variable to have a non-linear relationship with one or more predictors. The non-linear relationships are modeled via smooths, which can have different basis functions and basis dimensions. The basis function indicates how the smooths are built up. For example, a cubic spline basis function is constructed by connecting sections of cubic polynomials. The basis dimension indicates the upper limit for how complex the smooth can be (i.e., how many degrees of freedom are invested). The higher this number, the more wiggly the smooth can become. To prevent overfitting (i.e., too wiggly curves), GAMs are estimated using penalized likelihood estimation and cross-validation. The best, computationally feasible, smoothing basis for a single predictor or multiple isotropic predictors (i.e., predictors having the same scale) is the thin plate regression spline (Wood 2003). This basis is constructed by a weighted sum of geometrically simpler curves (or surfaces in the multidimensional case). When multiple predictors need to be combined (i.e., interact) which are on a different scale, a tensor product can be used combining different smoothing bases. Importantly, a mixed-effects regression approach (see Chapter 7 of Baayen 2008; Baayen *et al.* 2008), which is necessary for taking into account the structural variability associated with the random-effect factors (such as speakers and words) is also possible within the GAM framework. In that case, random-effect factors are modeled as smooths as well (see Chapter 6 of Wood 2006).

Rather than constructing a regression model in which geography is simplified as geographical distance (e.g., Nerbonne and Heeringa 2007), a GAM can model complex geographical patterns directly via a two-dimensional smooth of longitude and latitude. The first study to use a generalized additive modeling approach for the aggregate analysis of dialect variation was conducted by Wieling, Nerbonne, and Baayen (2011). They only used a generalized additive model to represent geography and used the fitted values of this model as a new predictor in a linear mixed-effects regression model. Their dataset consisted of pronunciation distances (compared to standard Dutch) for 562 words in 424 locations in the Netherlands. Besides finding support for a complex geographical pattern of pronunciation distances from standard Dutch with greater pronunciation distances from standard Dutch in the peripheral areas (see Figure 23.3), they simultaneously identified the importance of various social and lexical features. For example, larger communities were found to use pronunciations closer to standard Dutch, and more frequent words were more different from standard Dutch (indicating resistance to standardization).

Wieling *et al.* (forthcoming; see also Wieling 2012, Ch. 7) created a single generalized additive mixed-effects regression model to determine pronunciation distances from standard Catalan. The Catalan dataset (Valls, Wieling, and Nerbonne 2013) consisted of 357 words in 40 dialectal varieties located in Catalonia, Andorra and Aragon. In each location 8 speakers (born between 1930 and 1996) were interviewed. Similar to the results of Wieling, Nerbonne, and Baayen (2011), geography was found to be a highly important, non-linear predictor. Furthermore, a clear border effect was observed between Aragon and the other two regions: younger speakers in Catalonia and Andorra (where Catalan is an official language), but not in Aragon (where Catalan is not an official language) had pronunciations closer to the standard Catalan language than the older speakers. In addition, Wieling *et al.* (forthcoming) found support for other social and lexical variables, such as word category and the year of birth of the speakers.

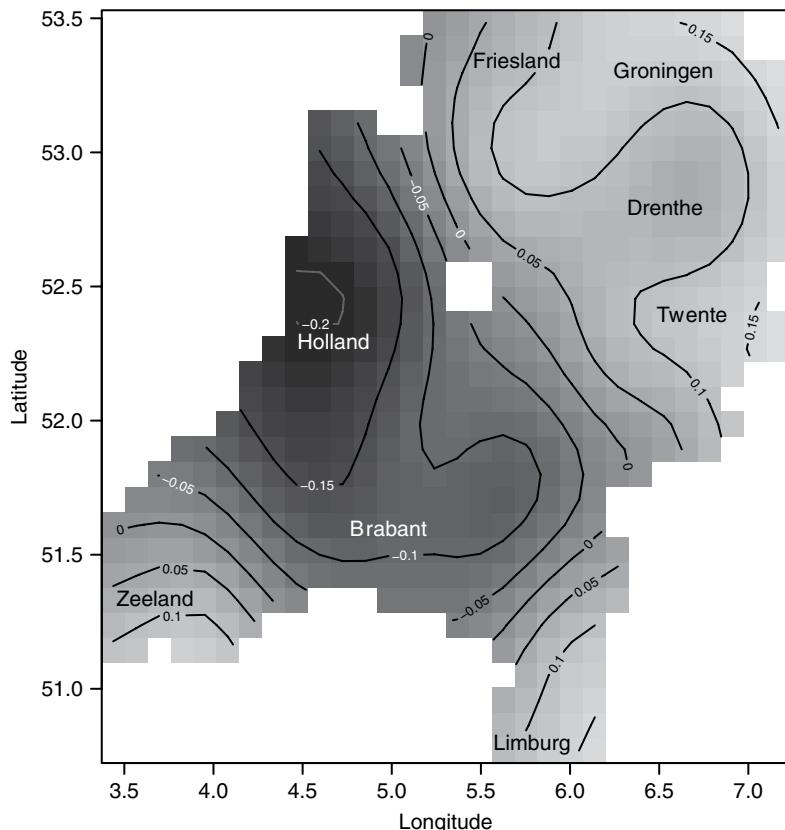


Figure 23.3 Contour plot obtained with a generalized additive model. The contour plot shows a regression surface of pronunciation distance (from standard Dutch) as a function of longitude and latitude obtained with a generalized additive model using a thin plate regression spline. The (black) contour lines represent aggregate distance isoglosses, darker shades of gray indicate smaller distances closer to the standard Dutch language, lighter shades of gray represent greater distances. Note that the empty square indicates the location of the IJsselmeer, a large lake in the Netherlands. Reprinted (including caption) from Wieling (2012) with permission.

The generalized additive modeling approach can also be applied to study other types of aggregate variation. Wieling *et al.* (2014) investigated Tuscan lexical variation in a dataset consisting of 170 concepts for 2,060 speakers (in 213 localities). Their response variable was binary, with a 1 indicating that the speaker used a non-standard Italian form and a 0 indicating the use of the standard Italian form. Consequently, they used a generalized additive mixed-effects *logistic* regression model for this data. They also used a more sophisticated approach to modeling geography: they allowed the geographical pattern to vary depending on concept frequency and speaker age. In effect, they used a four-dimensional tensor product smooth ($\text{longitude} \times \text{latitude} \times \text{concept frequency} \times \text{speaker age}$). Their results highlighted the potential of the generalized additive modeling approach and showed distinctive differences in the geographical patterns associated with speaker age and concept frequency. For example, whereas younger speakers, especially in the area around Florence, were more likely to use standard Italian lexical forms than older speakers, this did not appear to be the case for the low-frequency concepts. In that situation the younger speakers used more

general definitions, which did not coincide with the specific (old-fashioned) standard Italian form (e.g., they used ‘swine’ rather than ‘boar’ to denote a male pig).

As attempting to use a relatively new and complex statistical method might seem daunting, paper packages for the studies of Wieling *et al.* (2011), Wieling *et al.* (forthcoming) and Wieling *et al.* (2014) have recently been made available at the Mind Research Repository (<http://openscience.uni-leipzig.de>). These paper packages include all data and R commands (using the R package mgcv; Wood 2006) needed to fit the generalized additive models and replicate the results of these studies.

In sum, the advantage of the generalized additive modeling approach is that it allows one to directly incorporate the complex influence of geography on the aggregate patterns, whereas simultaneously considering the importance of other social and lexical variables. Furthermore, the availability of paper packages enables other researchers to easily apply these methods to their data as well.

23.5 Conclusions and Prospects

Dialectal material is often analyzed today using a variety of multivariate techniques. This chapter has been slanted to methods for the analysis of large aggregates and to techniques that seek to identify groups together with their common speech habits. In the case of geographical data, these might be areas and their lexical peculiarities, but we have sketched how more advanced techniques are poised to combine the analysis of geographical and social variables, contributing technically to the Chambers-Trudgill program of understanding variationist linguistics—dialectology and sociolinguistics—in a unified way.

We would like to close this chapter by identifying a challenge. As attractive as the GAMs and mixed-effects models are, they remain regression models, focused on the prediction of a single criterion variable. This works quite well for research that can be formulated in such a focused way, for example, on the relative proximity of local varieties (including different cross-cutting social distinctions) to a single standard language. By contrast, the techniques in the first part of the chapter, including clustering, MDS, but also factor analysis and related techniques, did not require the identification of a single criterion variable, but instead could be used fairly directly on a dialectologist’s table of sites and linguistic variables (or on the site \times site summary of that table’s differences). These techniques provide a more global picture of the landscape of language varieties, but they are (now) poorly equipped to include a range of other variables such as class, education, gender and the like. The challenge thus is to find a way to combine the virtues of the two approaches.

NOTES

- 1 See Chap. 7 (Goebel) and Chap. 20 (Heeringa and Prokić) above for more on the theory and computational underpinnings of dialectometry, respectively.
- 2 We emphasize geographical differences in this chapter (but see Sec. 23.4.2), but aggregating techniques may also be applied to social varieties or social dialectology (Boberg 2005).
- 3 See Chap. 25 (Grieve) for geo-statistical techniques that include a bias for geographically coherent areas.
- 4 The rest of the presentation assumes the discussion of statistics in the introduction to this section of the handbook.
- 5 Although Costello and Osborne (2005) find this one of the least reliable criteria.
- 6 But obviating the need for formant tracking, which has a higher rate of failure.

- 7 Clark (1973) had criticized that language differences were often treated as fixed effects even though the words used in studies and experiments were actually a random sample of a larger population of potential words, and therefore, should be treated as a random effect (as is done in the mixed-effects regression framework).

REFERENCES

- Baayen, R. Harald. 2008. *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.
- Baayen, R. Harald, Doug J. Davidson, and Doug M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4): 390–412.
- Bach, Adolf. 1950. *Deutsche Mundartforschung: ihre Wege, Ergebnisse und Aufgaben*. Heidelberg: Winter.
- Benzécri, Jean-Paul. 1992. *Correspondence analysis handbook*. New York: Marcel Dekker.
- Boberg, Charles. 2005. The North American regional vocabulary survey: New variables and methods in the study of North American English. *American Speech*, 80(1): 22–60.
- Bloomfield, Leonard. 1933. *Language*. New York: Holt, Rhinehart and Winston.
- Bolker, Benjamin M., Mollie E. Brooks, Connie J. Clark, Shane W. Geange, John R. Poulsen, M. Henry H. Stevens, and Jada-Simone S. White. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution* 24(3): 127–135.
- Brown, James. 2009. Principal components analysis and exploratory factor analysis—definitions, differences, and choices. *Japan Association for Language Teaching, Testing and Evaluation Special Interest Group Newsletter* 13.1: 26–30.
- Cichocki, Wladyslaw. 2006. Geographic variation in Acadian French /r/: What can correspondence analysis contribute toward explanation? *Literary and Linguistic Computing* 21.4: 529–541.
- Clark, Herbert H. 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of verbal learning and verbal behavior* 12.4: 335–359.
- Clopper, Cynthia G., and John C. Paolillo. 2006. North American English vowels: A factor-analytic perspective. *Literary and Linguistic Computing* 21.4: 445–462.
- Costello, Anna and Jason Osborne. 2005. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment, Research and Evaluation* 10: 1–9.
- Embleton, Sheila. 1993. Multidimensional scaling as a dialectometrical technique: Outline of a research project. In: Reinhard Köhler and Burghardt Rieger (eds.) *Contributions to quantitative linguistics*. 267–276. Dordrecht: Kluwer.
- Goebl, Hans. 1982. *Dialektometrie; Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie. (Philosophisch-Historische Klasse Denkschriften 157)* Vienna: Verlag der Österreichischen Akademie der Wissenschaften.
- Grieve, Jack, Dirk Speelman, and Dirk Geeraerts. 2011. A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change* 23(2): 193–221.
- Grimm, Jacob. 1819. *Deutsche Grammatik*. 1. Theil, Göttingen: Dieterich.
- Hastie, Trevor J., and Robert J. Tibshirani. 1990. *Generalized Additive Models*. London: Chapman & Hall/CRC.
- Heeringa, Wilbert J. 2004. *Measuring dialect pronunciation differences using Levenshtein distance*. Ph.D. thesis, University of Groningen.
- Jaeger, T. Florian. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of memory and language* 59(4): 434–446.
- Johnson, Keith. 2008. *Quantitative methods in linguistics*. Malden, MA: Blackwell.
- Johnson, Stephen C. 1967. Hierarchical clustering schemes. *Psychometrika* 32.3: 241–254.
- Kaufman, Leonard, and Peter J. Rousseeuw. 2005, 1990. *Finding groups in data: an introduction to cluster analysis*. Vol. 344. Hoboken: Wiley.
- Johnson, Daniel Ezra. 2009. Getting off the GoldVarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass* 3: 359–383.
- Joseph B. Kruskal, and Myron Wish. 1978. *Multidimensional scaling*. Vol. 11. London: Sage.
- Keune, Karen, Mirjam Ernestus, Roeland van Hout, and R. Harald Baayen (2005). Variation

- in Dutch: From written MOGELIJK to spoken MOK. *Corpus Linguistics and Linguistic Theory* 1(2): 183–223.
- Kurath, Hans, and Guy S. Lowman. 1970. *The Dialectal Structure of Southern England*. Tuscaloosa: University of Alabama Press.
- Kurath, Hans, and Raven McDavid. 1961. *The Pronunciation of English in the Atlantic States: Based upon the Collections of the Linguistic Atlas of the Eastern United States*. Ann Arbor: University of Michigan Press.
- Labov, William. 2001. *Principles of linguistic change Vol. 2: Social factors*. (Language in Society 29). Oxford: Oxford University Press.
- Leino, Antti, and Saara Hyvönen. 2008. Comparison of component models in analysing the distribution of dialectal features. *International Journal of Humanities and Arts Computing* 2(1–2): 173–187.
- Leinonen, Therese. 2008. Factor analysis of vowel pronunciation in Swedish dialects. *International Journal of Humanities and Arts Computing* 2(1–2): 189–204.
- Leinonen, Therese. 2010. *An acoustic analysis of vowel pronunciation in Swedish dialects*. Ph.D. thesis, University of Groningen.
- Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge: Cambridge University Press.
- Nerbonne, John. 2006. Identifying linguistic structure in aggregate comparison. *Literary and Linguistic Computing* 21(4): 463–475.
- Nerbonne, John. 2009. Data-Driven Dialectology. *Language and Linguistics Compass* 3(1): 175–198. DOI: 10.1111/j.1749-818x.2008.00114.x
- Nerbonne, John. 2010. Measuring the Diffusion of Linguistic Change. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365: 3821–3828. DOI: 10.1098/rstb.2010.0048.
- Nerbonne, John. 2010a. Mapping aggregate variation. In: Alfred Lameli, Ronald Kehrein and Stephan Rabanus (eds.) *Language and Space. International Handbook of Linguistic Variation* 2: 476–495. Berlin: Mouton De Gruyter.
- Nerbonne, John, and Wilbert Heeringa. 2007. Geographic distributions of linguistic variation reflect dynamics of differentiation. In *Roots: Linguistics in Search of its Evidential Base*, edited by Sam Featherston and Wolfgang Sternefeld, 267–297. Berlin: Mouton De Gruyter.
- Nerbonne, John, Wilbert Heeringa, and Peter Kleiweg. 1999. Edit distance and dialect proximity. In David Sankoff and Joseph Kruskal (eds.) *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, 2nd ed., Stanford: CSLI Press.
- Nerbonne, John, Peter Kleiweg, Wilbert Heeringa, and Franz Manni. 2008. Projecting dialect distances to geography: Bootstrap clustering vs. noisy clustering. In: Christine Preisach, Lars Schmidt-Thieme, Hans Burkhardt and Reinhold Decker (eds.) *Data Analysis, Machine Learning and Applications*, 2008. 647–654. Berlin: Springer.
- Paolillo, John C. 2013. Individual effects in variation analysis: Model, software, and research design. *Language Variation and Change* 25(1): 89–118.
- Pickl, Simon. 2013. *Probabilistische Geolinguistik: Geostatistische Analysen lexikalischer Variation in Bayerisch-Schwaben*. Stuttgart: Franz Steiner.
- Pinheiro, José C. and Douglas M. Bates (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer.
- Prokić, Jelena. 2010. *Families and resemblances*. Ph.D. thesis, University of Groningen.
- Prokić, Jelena, Çağrı Çöltekin, and John Nerbonne. 2012. Detecting shibboleths. *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*. 72–80. Shrodsburg, PA: Association for Computational Linguistics.
- Prokić, Jelena and John Nerbonne. 2008. Recognising groups among dialects. *International Journal of Humanities and Arts Computing* 2.1–2: 153–172.
- Prokić, Jelena, and Tim Van de Cruys. 2010. Exploring dialect phonetic variation using PARAFAC. *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*. Shrodsburg, PA: Association for Computational Linguistics.
- Sankoff, David, and William Labov. 1979. On the uses of variable rules. *Language in Society* 8(2–3): 189–222.
- Schirmunski, Viktor. 1962. *Deutsche Mundartkunde. Vergleichende Laut- und Formenlehre der deutschen Mundarten*. Berlin: Akademie Verlag.
- Séguy, Jean. 1971. La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane* 35(138): 335–357.
- Séguy, Jean. 1973. La dialectométrie dans l'Atlas linguistique de Gascogne. *Revue de Linguistique Romane* 37(145): 1–4.
- Shackleton Jr, Robert G. 2010. *Quantitative assessment of English-American speech relationships*. Ph.D. thesis, University of Groningen.
- Spruit, Marco René, Wilbert Heeringa and John Nerbonne. 2009. Associations among linguistic levels. *Lingua* 119(11): 1624–1642.

- Szemerécsanyi, Benedikt 2012. Geography is overrated. In: Sandra Hansen, Christian Schwarz, Philipp Stoeckle and Tobias Streck (eds.) *Dialectological and Folk Dialectological Concepts of Space. 215–231. (Current Methods and Perspectives in Sociolinguistic Research on Dialect Change 17)* Berlin: De Gruyter.
- Tabachnick, Barbara and Linda S. Fidell. 2001. *Using multivariate statistics*. Boston: Allyn and Bacon.
- Tagliamonte, Sali, and R. Harald Baayen. 2012. Models, forests and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change*, 24(2): 135–178.
- Trudgill, Peter. 1974. Linguistic Change and Diffusion: Description and Explanation in Sociolinguistic Dialect Geography. *Language in Society* 2: 215–246.
- Uiboaed, Kristel, Cornelius Hasselblatt, Liina Lindström, Kadri Muischnek and John Nerbonne. 2013. Variation of verbal constructions in Estonian dialects. *LLC: Journal of Digital Scholarship in the Humanities* 28.1: 42–62.
- Valls, Esteve, Martijn Wieling, and John Nerbonne. 2013. Linguistic advergence and divergence in north-western Catalan: A dialectometric investigation of dialect leveling and border effects. *LLC: The Journal of Digital Humanities Scholarship*, 28(1): 119–146.
- Wieling, Martijn. 2012. *A Quantitative Approach to Social and Geographical Dialect Variation*. PhD dissertation. Groningen: University of Groningen.
- Wieling, Martijn, Wilbert Heeringa and John Nerbonne. 2007. An aggregate analysis of pronunciation in the Goeman-Taeldeman-van Reenen-Project data. *Taal en Tongval* 59: 84–116.
- Wieling, Martijn, Simonetta Montemagni, John Nerbonne, and R. Harald Baayen. 2014. Lexical differences between Tuscan dialects and standard Italian: Accounting for geographic and socio-demographic variation using generalized additive mixed modeling. *Language* 90(3): 669–692.
- Wieling, Martijn and John Nerbonne. 2011. Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features. *Computer Speech & Language* 25.3: 700–715.
- Wieling, Martijn, John Nerbonne, and R. Harald Baayen. 2011. Quantitative Social Dialectology: Explaining Linguistic Variation Geographically and Socially. *PLOS ONE*, 6(9): e23613.
- Wieling, Martijn, Esteve Valls, R. Harald Baayen, and John Nerbonne. forthcoming. Border effects among Catalan dialects. In *Mixed Effects Regression Models in Linguistics*, edited by Dirk Speelman, Kris Heylen and Dirk Geeraerts (eds.). New York: Springer.
- Winter, Bodo. 2013. Linear models and linear mixed effects models in R with linguistic applications. *arXiv preprint arXiv:1308.5499*.
- Wood, Simon N. 2003. Thin plate regression splines. *Journal of the Royal Statistical Society: Series B*, 65: 95–114.
- Wood, Simon N. 2006. *Generalized Additive Models: An Introduction with R*. London: Chapman & Hall/CRC.

24 Spatial Statistics for Dialectology

JACK GRIEVE

24.1 Introduction

Spatial statistics are used to analyze the values of variables that are measured over a series of locations (see Fotheringham *et al.* 2000; Haining 2003; Fortin and Dale 2005; Ripley 2005; Bivand *et al.* 2008; Chun and Griffiths 2013). Unlike most statistical methods, spatial statistics take into consideration the location of each observation, which allows these methods to focus on the analysis of spatial patterns by comparing the values of a variable at nearby locations. Although spatial statistics have clear application in dialectology, providing a statistical foundation for the analysis of regional linguistic variation, they have rarely been applied until quite recently (see Lee and Kretzschmar, 1993; Kretzschmar, 1996, 2003; Grieve 2009, 2011, 2012, 2013, 2014; Grieve *et al.* 2011, 2013a, 2013b; Sibler *et al.* 2012; Asnaghi 2013; Szemrecsanyi 2013; Tamminga 2013). The goal of this chapter is, therefore, to present a practical introduction to common spatial statistics that are especially useful for dialectologists.

In particular, this chapter introduces methods for spatial autocorrelation analysis, variogram analysis, and spatial interpolation, which can be used to test for spatial clustering in the values of linguistic variables, model spatial patterns exhibited by linguistic variables, and predict the values of linguistic variables at unobserved locations. Taken together, these methods constitute a powerful, standardized, and statistically grounded toolbox of techniques for the spatial analysis of regional linguistic variation, and provide principled solutions to many of the most persistent methodological problems in dialectology. All of these analyses can be carried out using common statistical and geographical software packages, including the R statistical computing environment (see Bivand *et al.* 2008), which was used to conduct the analyses for this chapter. These methods are described below, but first two regional linguistic variables, which are used to illustrate the methods being discussed, are introduced and mapped.

24.2 Data

Two regional linguistic variables are used throughout this chapter to illustrate the application of spatial statistics: *not* contraction following forms of the verb *to do* (Grieve 2011) and *everyone/everybody* alternation (Grieve *et al.* 2013). As required for spatial analysis, these

alternation variables are *spatially referenced* in the sense that each observation is associated with a particular location, defined by a longitude and latitude. Specifically, each of these variables were measured over 200 cities from across the contiguous United States based on a 25 million word dialect corpus of letters to the editor (Grieve 2009, 2011, 2012). As is common in dialect studies, as well as in sociolinguistics more generally, these linguistic variables are *alternation variables* (Labov 1972), which consist of a set of synonymous variant linguistic forms. The alternation variables were measured by counting the occurrences of the relevant forms (*don't*, *do not*, *doesn't*, *does not*, *didn't*, *did not*, *everyone*, *everybody*) in the 200 city sub-corpora. The percentage of occurrences of the first variant (contracted *not*, *everyone*) was then calculated for each city by dividing its frequency by the sum of the frequencies of both variants and multiplying this value by 100. For example, if a city sub-corpus contains 30 occurrences of *everyone* and 20 occurrences of *everybody*, then the percentage of *everyone* in that corpus is 60%. In addition, two categorical variables were derived from these two continuous variables by assigning all locations with percentages larger than or equal to the median for that variable as locations where the first variant occurs and all locations with percentages smaller than the median as locations where the second variant occurs. It should also be noted that despite focusing on alternation variables here, the methods introduced in this chapter can be applied to any type of linguistic variable, including quantitative acoustic variables (e.g., Grieve *et al.* 2013) and the frequency of individual linguistic forms (e.g., Szmrecsanyi 2013).

The maps plotting the percentages of *don't/do not* and *everyone/everybody* alternation across the 200 city sub-corpora are presented in Figures 24.1 and 24.2. To create these maps, the observed percentages for each variable were divided into quartiles, with locations characterized by relatively high percentages of the first variant (contracted *not*, *everyone*) being plotted in darker shades of gray and locations characterized by a relatively high percentages of the second variant (full *not*, *everybody*) being plotted in lighter shades of gray. Because both alternations consist of two variants, only one map is needed to show the distribution of both forms. Note that this approach to mapping has no effect on the results of the spatial analyses that follow, which are based on the values of the variables and their locations. Also note that although these variables are measured across observation *points*, the statistics introduced below can also be applied to variables measured across observation *areas* (e.g., states).



Figure 24.1 *Not* contraction raw map.



Figure 24.2 *Everyone/Everybody* alternation raw map.

Overall, these maps do not show clear regional patterns, although contracted *not* appears to be more common in the West, full *not* appears to be more common in the East, and *everyone* appears to be slightly more common in the Mid-Atlantic States. These results are not unusual: dialect maps for individual linguistic variables are often noisy, requiring careful analysis to identify the underlying patterns of regional variation (Nerbonne 2009). This is why traditional dialectologists plot isoglosses and why dialectometrists generally forgo the analysis of individual variables altogether, focusing instead on the analysis of multiple variables using multivariate statistics. Spatial statistics, as introduced in this chapter, offer an alternative approach to making sense of dialect maps, allowing for underlying patterns of regional variation to be identified in the values of individual linguistic variables through a replicable and statistically justified procedure.

24.3 Spatial Autocorrelation Analysis

The most commonly applied spatial statistics in dialectology are measures of spatial autocorrelation, which allow for forms of *spatial dependence* to be detected in the values of individual spatially referenced variables. Spatial dependence occurs when there is a relationship between the values of a variable and the locations over which those values are measured (Haining 2003; Fortin & Dale 2005). The opposite of spatial dependence is *spatial independence*, which occurs when the values of a variable are distributed at random across space. Spatial dependence can be *stationary* or *non-stationary*. Spatial stationarity is characterized by a homogeneous regional pattern that is independent of location, with a variable exhibiting a consistent overall pattern of spatial change across the map. Spatial non-stationarity is characterized by a heterogeneous regional pattern that is dependent on location, where the nature of the pattern changes depending on what part of the map is analyzed (Chun and Griffiths 2013). For example, spatial non-stationarity can appear as a complex and irregular pattern, such as small clusters of high and low values dispersed across a region. Following this distinction, statistical methods for analyzing spatial autocorrelation can be grouped into *global statistics*, which produce one value that summarizes the overall degree of spatial

autocorrelation across the entire map and which are therefore especially useful for analyzing stationary patterns, and *local statistics*, which are measured at each location and which are therefore especially useful for analyzing non-stationary patterns. Before discussing these techniques, however, the concept of a spatial weights matrix is introduced.

24.3.1 Spatial Weights Matrix

To conduct a spatial autocorrelation analysis it is necessary to define a *spatial weights matrix* (SWM), which provides the model of spatial association upon which a spatial autocorrelation analysis is based (Cliff and Ord 1969, 1973, 1981; Haining 2003; Fortin and Dale 2005). Specifically, a SWM is a location-by-location data matrix that specifies the spatial relationship between every pair of locations as a weight, with a relatively strong weight generally indicating that those locations are close together and a relatively weak weight generally indicating that those locations are far apart.

A SWM can be categorical or continuous. Categorical SWMs are usually binary, with pairs of nearby locations being assigned a weight of 1 and all other pairs of locations a weight of 0. There are many possible ways to define nearby locations, but when working with point data the simplest approach is to set a maximum distance after which pairs of locations are no longer considered neighboring. Pairs of locations within this distance are assigned a weight of 1 and pairs of locations outside this distance are assigned a weight of 0. This type of SWM is *symmetric*, with the weight for the pairing of location A with location B being the same as the weight for the pairing of location B with location A. Alternatively, a binary SWM can be defined by assigning pairs of locations a weight of 1 if the second location is one of the nearest neighbors of the first location, for some specific number of neighbors. This type of SWM is generally *non-symmetric*, because even if location A is one of the nearest neighbors of location B, location B may not be one of the nearest neighbors of location A. When dealing with area data, a binary SWM can also be defined by assigning a weight of 1 to all pairs of areas that share a common border. In this case the SWM is referred to as a *contiguity matrix*. A contiguity matrix can also be defined for point data by dividing the map into contiguous areas centered around each location, for example through some form of tessellation.

Continuous SWMs consist of weights that can take a range of values, most commonly between 0 and 1. Continuous SWMs are generally based on the distance between locations, with pairs of locations that are close together being given proportionally higher weights than pairs of locations that are far apart. Most commonly, continuous SWMs for point data are calculated based on inverse distance

$$w_{ij} = \frac{1}{d_{ij}^p}$$

where a pair of locations i and j is assigned a weight w_{ij} equal to the reciprocal of the distance d_{ij} between those locations, with p most commonly set at 1 or 2 (Fortin and Dale 2005). Inverse distance weights are relatively high only for the closest locations, quickly falling off as distance between locations increases.

In addition to selecting the type of SWM (e.g., maximum distance, nearest neighbors, contiguity, inverse distance), it is also necessary to make various other decisions depending on the type of SWM selected (e.g., distance threshold, number of nearest neighbors, contiguity definition, inverse distance power). Furthermore, when conducting a global analysis, it is also possible to standardize the SWM. The goal of standardization is to ensure that each location contributes equally to the calculation of global spatial autocorrelation. The most common approach to standardization is *row standardization*, where the weights for each location are standardized to sum to one. If the SWM is not standardized, then the contribution of each

location to a global measure is proportional to the strength of the weights for that location—which may or may not be ideal. For example, given an unstandardized binary maximum distance SWM, an isolated location will contribute less to the global spatial autocorrelation measure than a location with many neighboring locations.

The multitude of possible SWMs is both an advantage and a disadvantage of spatial autocorrelation analysis: this flexibility allows spatial autocorrelation analysis to be sensitive to a wide range of different spatial patterns, but selecting an SWM is a complex decision that requires careful consideration and can have a substantial effect on the results of the analysis. Contiguity matrices were standard in early research on global (Moran 1948) and local (Getis and Ord 1992) spatial autocorrelation, but it has been argued that continuous SWMs are usually more realistic, because the influence of locations on each other usually does not drop off entirely after some distance (Cliff and Ord 1969; Fotheringham *et al.* 2000). Regardless, the nature of the variable, region, and locations under analysis should be taken into consideration when selecting a SWM, rather than defaulting to any one type. It is also generally advisable to select a simple and standard SWM, as this facilitates the interpretation of the results and limits the possibility of problems arising due to the use of an untested SWM. Finally, the results of varying the SWM on the spatial autocorrelation analysis should be examined directly, so as to guard against unstable results and maximize the chances of detecting important patterns of spatial autocorrelation.

Five SWMs are used in this chapter to demonstrate spatial autocorrelation analysis. The first SWM is an unstandardized, symmetric, and categorical maximum distance SWM, with pairs of locations separated by 600 km or less assigned a weight of 1 and all other pairs of locations assigned a weight of 0. The second SWM is the row-standardized version of the first SWM. The third SWM is a non-symmetric and categorical nearest neighbor SWM, which is naturally standardized, with the three nearest neighbors for each location assigned a weight of 1 and all other pairings assigned a weight of 0. The fourth SWM is an unstandardized, symmetric, continuous inverse distance SVM, with pairs of locations assigned a weight based on inverse distance (with $p=1$). Finally, the fifth SWM is the row-standardized version of the fourth SWM.

24.3.2 Global Spatial Autocorrelation Analysis

Global spatial autocorrelation statistics test the overall degree of spatial dependence in the values of a spatially referenced variable by comparing the values of the variable to each other across space. Because global spatial autocorrelation statistics produce a single value that summarizes spatial dependence across an entire map, these methods are more useful for analyzing variables characterized by stationarity spatial patterns, which can reasonably be described by a single measure. Spatial autocorrelation can be both positive and negative. Positive spatial autocorrelation occurs when the values of a variable at nearby locations tend to be similar to a greater degree than would be expected by chance, whereas negative spatial autocorrelation occurs when the values of a variable at nearby locations tend to be different to a greater degree than would be expected by chance (e.g., a checkerboard pattern). Positive spatial autocorrelation is therefore associated with *spatial clustering*, whereas negative spatial autocorrelation is associated with *spatial dispersion*. Two of the most common global spatial autocorrelation statistics are introduced below: the join count statistic for categorical data and Moran's *I* for continuous data.

24.3.2.1 Join Count

The join count statistics are the most common measures of spatial autocorrelation for categorical data (Moran 1984; Cliff and Ord 1981; Fortin and Dale 2005), as well as the first spatial statistics applied in dialectology (Lee and Kretzschmar 1993; Kretzschmar 1996, 2003). Given a binary variable, where each observation is associated with one of two values, the join count statistics are calculated by first counting the total number of pairs of nearby

locations over which the variable is measured. The total number of pairs of nearby locations that both exhibit the first value, the second value, and different values are then counted and the proportion of neighboring locations with identical value is compared to the number of neighboring pairs of locations that would be expected to agree at random.

The join count statistics consist of three separate measures: J_{BB} , which is the number of neighboring locations that share the first value; J_{WW} which is the number of neighboring locations that share the second value; and J_{BW} which is the number of neighboring locations that exhibit different values. These three values can be calculated for a spatially referenced binary variable x as

$$J_{BB} = \frac{1}{2} \left(\sum_{i=1}^n \sum_{j=1}^n w_{ij} x_i x_j \right), i \neq j$$

$$J_{BW} = \frac{1}{2} \left(\sum_{i=1}^n \sum_{j=1}^n w_{ij} (x_i - x_j)^2 \right), i \neq j$$

$$J_{WW} = \frac{1}{2} \left(\sum_{i=1}^n \sum_{j=1}^n w_{ij} \right) - (J_{BB} + J_{BW}), i \neq j$$

where x_i and x_j are the values of the variable at locations i and j , n is the total number of locations, and w_{ij} is the value of a binary, symmetric, and non-standardized SWM, which assigns a weight of 1 to all neighboring pairs of locations and a weight of 0 to all other pairs of locations. Note that although the join count equations involve summing across all pairs of locations, the results depend only on the values of the variable at neighboring locations, because the spatial weights associated with non-neighboring locations are always equal to 0.

The results of the join count analysis of the categorical versions of the two alternation variables are presented in Table 24.1, calculated using the non-standardized, binary, 600 km maximum distance SWM. These results show that *not* contraction, *everyone*, and *everybody* show significant levels of clustering across the 200 locations, although varying the maximum distance used to define the SWM detects significant levels of clustering for full *not* as well.

24.3.2.2 Moran's I

The original and most common global autocorrelation statistic for continuous data is Moran's I , which is used to test for positive and negative spatial autocorrelation (Moran 1948; Cliff and Ord 1969, 1973, 1981; Haining 2003; Fortin and Dale 2005). Moran's I is calculated by comparing the value of a spatially referenced variable at each location to the values of the variable at other locations, where this comparison is weighted based on the proximity of the locations, as specified by the SWM. For example, given a binary SWM, the value of the variable at each location is only compared to its values at neighboring locations and all of these comparisons are given equal weight. Alternatively, given a continuous SWM, the value of the variable at each location is compared to its value at every other location, but these comparisons are inversely weighted based on the distance between locations. The value of Moran's I usually ranges between -1 and +1, with a significant positive value indicating spatial clustering and a significant negative value indicating spatial dispersion. Moran's I , therefore, provides a useful method for testing if the values of a linguistic variable are distributed at random across a set of locations or if they exhibit a pattern of spatial dependence. For this reason, Moran's I has been applied in numerous recent dialect studies (Grieve 2009, 2011, 2012, 2014; Grieve *et al.* 2011, 2013a; Asnaghi 2013; Szemrecsanyi 2013; Tamminga 2013).

Table 24.1 Join count results.

Variable	J_{BB}			J_{WW}		
	Observed	Expected	p	Observed	Expected	p
Not Contraction	882	696	<.001	571	642	.153
Everyone/Everybody	678	825	.009	701	529	< .001

In a similar manner as the Pearson correlation coefficient, global Moran's I is calculated for a spatially referenced variable x as

$$I = \frac{n}{\sum_{i}^n \sum_{j}^n w_{ij}} \frac{\sum_{i}^n \sum_{j}^n w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_{i}^n (x_i - \bar{x})^2}$$

where n is the number of locations over which the variable is measured, x_i and x_j are the values of the variable at locations i and j , \bar{x} is the mean value of variable across all locations, and w_{ij} is the value of the SWM for the pair of locations i and j . To calculate the significance of Moran's I , a z-score is calculated, either under the assumption of randomization or normality (Fortin and Dale 2005). Much like how the Pearson product-moment correlation coefficient measures the relationship between two variables based on their covariance divided by the product of their standard deviations, global Moran's I measures the spatial relationship between the values of a single variable based on the covariance between a variable and itself across a set of locations, weighted based on the distance between those locations, divided by the squared standard deviation of that variable.

The results of the Moran's I analysis for the continuous versions of the two alternation variables are presented in Table 24.2, calculated using all five SWMs, with p -values generated under the assumption of randomization. The global spatial autocorrelation analysis finds that both *not* contraction and *everyone/everybody* alternation show significant positive spatial autocorrelation (i.e., spatial clustering), although spatial clustering *everyone/everybody* alternation is only identified by the two 600 km maximum distance SWMs. Row standardization did not have a substantial effect on the results of these analyses.

24.3.3 Local Spatial Autocorrelation Analysis

When a map is characterized by non-stationary spatial dependence, a single value cannot provide a complete summary of that pattern. In such cases, the utility of global spatial autocorrelation statistics is limited, potentially failing to identify significant patterns of clustering. For this reason, local spatial autocorrelation statistics, which produce a value for each location summarizing the degree of spatial autocorrelation around that location, were developed for analyzing and mapping more complex patterns of regional variation at the local level. The two most common types of local spatial autocorrelation statistics are introduced below: the local Getis-Ord G statistics and local Moran's I .

24.3.3.1 Local Getis-Ord G_i and G_i^*

Local Getis-Ord G_i and G_i^* are common spatial autocorrelation statistics that are used to identify local patterns of spatial clustering (Getis and Ord 1992; Ord and Getis 1995; Fotheringham *et al.*

Table 24.2 Global Moran's *I* results.

Spatial Weights Matrix		Not Contraction		Everyone/Everybody	
Type	Standardized	<i>I</i>	<i>p</i>	<i>I</i>	<i>p</i>
600 km Maximum Distance	no	0.136	<0.001	0.025	0.037
	yes	0.151	<0.001	0.044	0.010
Three Nearest Neighbors	yes	0.125	0.007	-0.061	0.856
	Inverse Distance	0.061	<0.001	-0.007	0.552
		0.057	<0.001	-0.007	0.584

2000; Haining 2003; Fortin and Dale 2005). The local Getis-Ord statistics produce one value per location, which measures the degree to which that location is part of a high value cluster (i.e., *hot spot*) or a low value clusters (i.e., *cold spot*). This is accomplished by comparing the values of the variable at surrounding locations to the mean, where this comparison is weighted based on the proximity of those locations to the central reference location, as specified by the SWM. The local Getis-Ord statistics return a *z*-score for that reference location (see Ord and Getis 1995), where a relatively high positive *z*-score indicates that the location is in the midst of high value locations, a relatively low negative *z*-score indicates that the location is in the midst of low value locations, and a *z*-score around zero indicates that the location is in the midst of a region of variability or transition. These *z*-scores are calculated for each location and mapped to identify clusters of high- and low-value locations. In effect, the local Getis-Ord statistics produce smoothed maps that identify underlying patterns of regional variation in the values of a variable, while filtering out the non-regional variation in the values of the variable. The local Getis-Ord statistics, therefore, approximate the isogloss method, identifying regions where the different values of a linguistic variable predominate, which is why these statistics have been applied in numerous recent dialect studies (Grieve 2009, 2011, 2012, 2013, 2014; Grieve *et al.* 2011, 2013a, 2013b; Sibler *et al.* 2012; Asnaghi 2013; Tamminga 2013).

The G_i *z*-scores are calculated for a location i for a spatially referenced variable x as

$$G_i = \frac{\sum_{j=1}^n w_{ij}x_j - \bar{x}\sum_{j=1}^n w_{ij}}{s\sqrt{\frac{(n-1)\sum_{j=1}^n w_{ij}^2 - \left(\sum_{j=1}^n w_{ij}\right)^2}{n-2}}}, j \neq i$$

and the G_i^* *z*-scores are calculated as

$$G_i^* = \frac{\sum_{j=1}^n w_{ij}x_j - \bar{x}\sum_{j=1}^n w_{ij}}{s\sqrt{\frac{n\sum_{j=1}^n w_{ij}^2 - \left(\sum_{j=1}^n w_{ij}\right)^2}{n-1}}}$$

where \bar{x} is the mean and s is the standard deviation of the variable, excluding location i for G_i . To identify local patterns of spatial clustering, these equations compare the values of the

variable at each location to the mean value of that variable, where these comparisons are weighted based the distance between each location and a central reference location, as defined by a SWM. The two versions of the local Getis-Ord statistic differ in terms of how the central reference location for which the statistic is calculated is treated, with the location included in the calculation of G_i^* and excluded from the calculation of G_i . It is usually preferable to include the location in the calculation, as it is part of any cluster that may be identified around it, but only certain SWMs allow G_i^* be calculated, as the weight for the comparison between a location and itself must be defined and must be larger than 0. The specific SWM that is selected may therefore require that the G_i statistic be applied. Note that standardization of the SWM is of no consequence, as each location is examined individually.

The local G_i spatial autocorrelation maps are presented for the two variables in Figures 24.3 and 24.4, based on the 600 km maximum distance binary SWM, and the local G_i^* spatial autocorrelation maps are presented in Figures 24.5 and 24.6 based on the inverse distance SWM. These maps distinguish between locations with positive (i.e., high value cluster) and negative (i.e., low value cluster) z-scores, highlighting locations with z-scores ≥ 1.96 and ≤ -1.96 , which correspond to a two-tailed test at the .05 significance level. These maps clearly identify local patterns of spatial clustering in both variables, showing that *not* contraction is more common across the West and full *not* is more common in the East, especially the Northeast, and that *everyone* is more common in the Northeast and *everybody* is more common in the Southwest. The results for *everyone/everybody* alternation are somewhat surprising given the mixed results of the global spatial autocorrelation analysis, exemplifying how a local analysis can identify patterns of spatial clustering that a global analysis can miss.

Comparing these maps, it is clear that both approaches identify the same general patterns, although there are small differences. Overall, the maps based on the maximum distance SWM are smoother, primarily because the z-scores at each location tend to be based on a larger number of heavily weighted locations including the location itself. The regions identified in the two sets of maps are also slightly different. For example, although both approaches identify a similar Northeastern *everyone* region, the core of this region shifts between the Mid-Atlantic States based on the inverse



Figure 24.3 *Not* contraction Getis-Ord G_i^* map (600 km maximum distance).

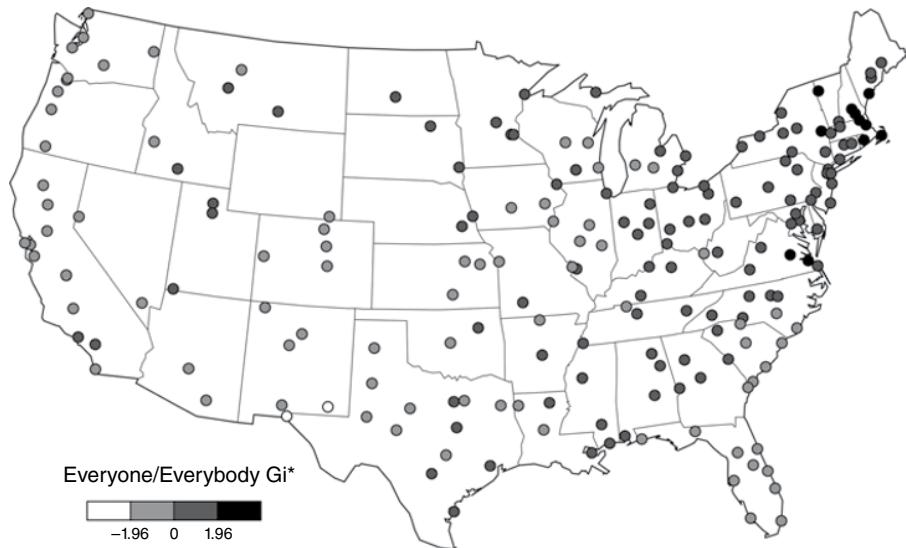


Figure 24.4 *Everyone/Everybody* Getis-Ord G_i^* map (600 km maximum distance).



Figure 24.5 *Not contraction* Getis-Ord G_i map (inverse distance).

distance SWM and New England based on the maximum distance SWM. Both maps are correct—they identify spatial clustering on slightly different criteria—but the inverse distance map arguably aligns more closely to a subjective analysis. More generally, these same basic results were obtained using both local Getis-Ord statistics and a range of different SWMs, up to a 1000 km maximum distance cutoff and 40 nearest neighbors, after which relatively clear spatial patterns were smoothed over in the local spatial autocorrelation maps.



Figure 24.6 Everyone/Everybody Getis-Ord G_i map (inverse distance).

24.3.3.2 Local Moran's I

Another common measure of local spatial autocorrelation is local Moran's I (Anselin 1995), which is calculated for a location i for a spatially referenced variable x as

$$I_i = \frac{(x_i - \bar{x})}{\frac{1}{n} \sum_k^n (x_k - \bar{x})^2} \sum_{j=1}^n w_{ij} (x_j - \bar{x})$$

Local Moran's I yields a value for each location representing the degree of spatial clustering or spatial diffusion around that location. Associated z -scores can then be calculated and mapped to identify regions of spatial clustering (i.e., high positive z -scores) and dispersion (i.e., low negative z -scores). The statistic does not distinguish between high- and low-value clusters, however, which generally makes local Moran's I less applicable in dialectology than the local Getis-Ord statistics. Nevertheless, local Moran's I is useful for identifying areas of spatial dispersion within a map that contains spatial clusters, as well as for detecting spatial outliers. The two local spatial autocorrelation statistics can therefore be used together, with the local Getis-Ord G statistics being used to identify high and low value clusters and with the local Moran's I statistic being used to identify areas of transition.

The local Moran's I map for the continuous version of *not* contraction, calculated using the maximum distance, binary SWM, is presented in Figure 24.7. Although not as definitive as the local G maps, the areas of spatial clustering by the two statistics roughly align. The local Moran's I map also allows for relatively large regions of variability between clusters to be identified, such as in the Lower Midwest and the Central States. In addition, analyzing the most extreme negative z -scores allows for outliers to be detected. For example, *do not* contraction outliers are identified in Barnstable, Charlotte, Madison, Salisbury, and Savannah, all of which can be verified by inspecting the raw maps for this variable.



Figure 24.7 *Not contraction* local Moran's I map (600 km maximum distance).

24.4 Variogram Analysis

A variogram is a model of the spatial variance in the values of a spatially referenced variable (Isaaks and Srivastava 1989; Cressie 1993; Haining 2003; Fortin and Dale 2005; Wackernagel 2010). When plotted, a variogram (also known as a *semivariogram*) shows how the variance in the values of a variable across locations changes depending on the distance between locations. Generally, a variogram shows that variance increases as distance between locations increases, with locations that are closer together tending to exhibit similar values to a greater degree than locations that are farther apart. Variogram analysis along with approaches for spatial interpolation that are based on the variogram are referred to collectively as *geostatistics*. These methods have only recently been applied in dialectology (Grieve 2013).

In geostatistics, a distinction is made between empirical and theoretical variograms. Given a variable measured over a series of locations, an *empirical variogram* can be constructed by plotting the variance in the values of the variable at all pairs of locations that are separated by a distance that falls within a certain distance interval for a series of distance intervals. For example, variances could be measured for all locations separated by 1 and 500 km, 501 and 1000 km, and so on. Specifically, the value of the empirical variogram $\hat{\gamma}$ for a spatially referenced variable x for all pairs of locations i and j that are separated by a distance that falls within the distance interval h as

$$\hat{\gamma}(h) = \frac{1}{2n} \sum_{(i,j) \in n} (x_i - x_j)^2$$

where n is total number of pairs of locations that are separated by a distance that falls within that distance interval. This calculation is then repeated for a series of generally non-overlapping distance intervals and these variances are plotted against the distance intervals, to create an empirical variogram, which shows how the variance in the values of a variable changes depending on the distance between locations.

The empirical variograms based on the G_i^* z-scores for *not* contraction (Figure 24.5) is plotted in Figure 24.8, as a series of 8 white circles in the foreground, based on eight equally sized distance intervals, calculated using Euclidean distance. In addition, in the background of this figure the *variogram cloud* is plotted, which shows the squared difference in values for every pair of locations in the distribution of the variable. The variogram cloud is basis for the empirical variogram, which is created by taking an average of the squared differences between all pairs of locations in the variogram cloud that fall within each of the specified distance intervals. Note that the raw data for *not* contraction was not used to generate the example variogram because it does not show clear regional patterns and therefore is not as well suited for exemplifying variograms as locally autocorrelated z-scores.

Given an empirical variogram, which is defined for certain distance intervals, it is possible to estimate the theoretical variogram for that variable, which is defined for all possible distances, by fitting a function to the empirical variogram. Various functions are commonly used to fit theoretical variograms to empirical variograms including exponential, Gaussian, and spherical models (see Fortin and Dale, 2005). The theoretical variogram for *not* contraction is plotted in Figure 24.8 as a line based on a Gaussian model. Other models were tested and found to produce similar theoretical variograms. The shape of the variogram also remains relatively stable if the number of intervals is varied.

The basic pattern exhibited by a variogram can be described by three values. The *nugget* is the variance of the variogram at distance zero. In the variogram for *not* contraction, the nugget is very close to zero, indicating that the values of the variable at nearby locations tend to be similar. Variograms then generally increase quickly with distance until they plateau at a certain variance, which is known as the *sill*. For example, the variogram for the locally autocorrelated *not* contraction map has a sill of approximately 4.5. An increasing slope, especially at relatively small distances, indicates that a variable exhibits a clear pattern of spatial clustering. The steepness of this slope also reveals how gradually the values of a variable change across space, with a steep slope being associated with more sudden changes and a moderate slope being associated with more gradual changes. Finally, the distance at which the sill is reached is known as the *range*. After this distance the amount of spatial influence is

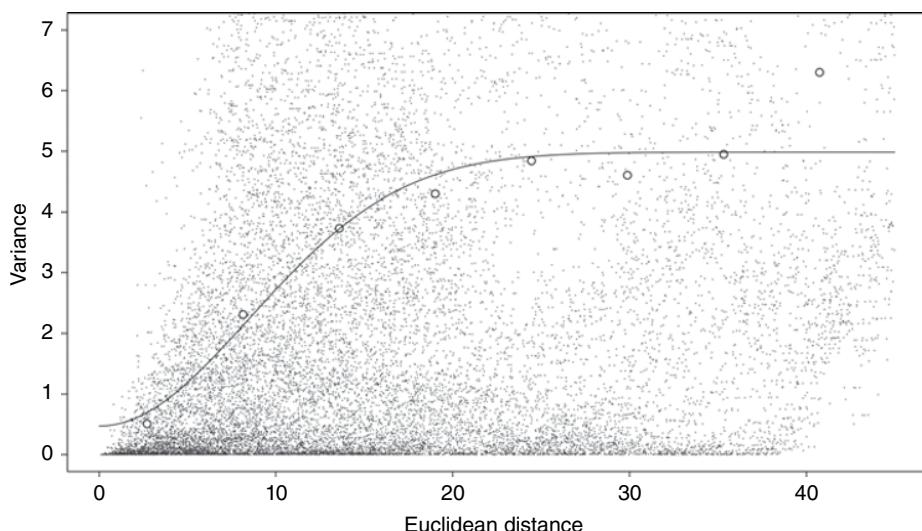


Figure 24.8 *Not* contraction Getis-Ord G_i^* variogram.

negligible. For example, the variogram for *not* contraction has a range of approximately 15, when measured as Euclidean distance.

24.5 Spatial Interpolation

The final topic covered by this review is spatial interpolation, which involves estimating the value of a variable at an unobserved location based on the values of the variable at observed locations. Although there are many different approaches to spatial interpolation, this section describes two of the most common options: inverse distance weighted interpolation and ordinary kriging. These methods can be applied to a variety of problems in dialectology. Their most basic application is to predict the value of a linguistic variable at a specific location of interest, but these methods can also be used to prepare sets of spatially referenced linguistic variables that are not defined across the same set of locations for comparison or aggregation. For example, spatial interpolation can be used to replace missing values or to estimate the values of a variable across a consistent set of reference locations, which can be particularly useful when working with data from different dialect surveys (see Grieve 2013). Interpolation is also useful for visualizing patterns of regional variation, for example by interpolating a variable across a very dense grid of locations, which can also be used to plot isoglosses.

24.5.1 Inverse Distance Weighted Interpolation

Inverse distance weighted interpolation (Ripley 2005; Cressie 1993; Bivand *et al.* 2008) is a relatively simple way to estimate the value of a variable at an unobserved location based on its observed values, taking into consideration the distance between the unobserved and observed locations. Specifically, the estimated value of a variable at an unobserved location is calculated as the average value of the variable at surrounding locations, weighted based on the inverse of the distance between those locations and the unobserved location, so that nearby locations are given greater weight:

$$\hat{x}_0 = \frac{\sum_i^n d_{0i}^{-p} x_i}{\sum_i^n d_{0i}^{-p}}$$

where \hat{x}_0 is the estimated value of the variable at the unobserved location, n is the total number of observed locations, d_{0i} is the distance between the unobserved location and an observed location i , and p is the inverse distance weighting power, which is generally set at 1 or 2. The interpolation of the variable at an unobserved location can also be restricted to only the nearest observed neighbors.

Inverse distance weighted interpolation, with $p=2$, was used to estimate the G_i^* z-scores for *not* contraction (see Figure 24.5) at seven unobserved locations. The interpolated values are mapped in Figure 24.9, marked with a double circle. By comparing these estimated values to the surrounding values, it can be seen that inverse distance weighted interpolation has produced reasonable estimates that are largely in line with the observed values of the variable at surrounding locations.

Although inverse distance weighted interpolation generally produces satisfactory results, there are inherent limitations with this approach to spatial interpolation. Most important, the

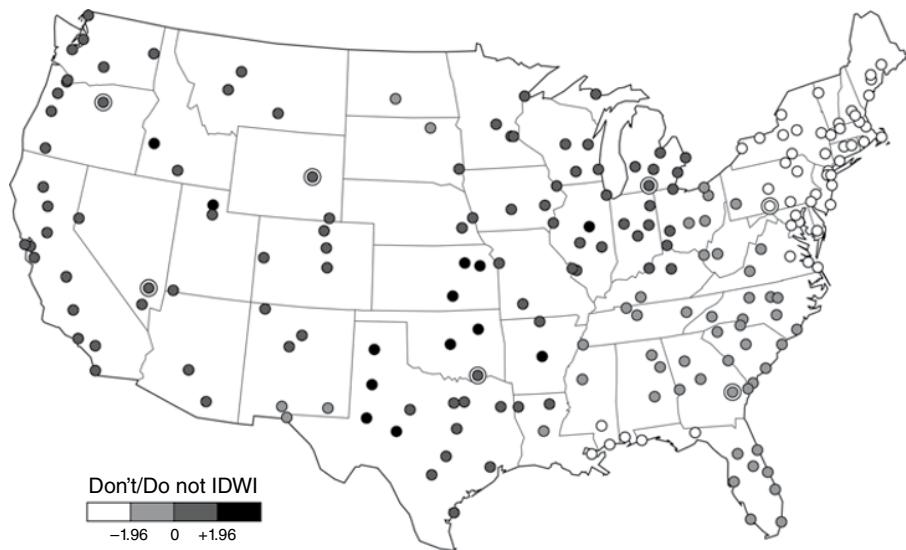


Figure 24.9 Not contraction inverse distance weighted interpolation.

interpolated value of the variable is always bounded by the minimum and maximum observed values of the variable, which can lead to inaccuracies. For example, consider a variable that steadily increases across a region from east to west over a regular grid of observation points. If the unobserved location whose value is being interpolated falls within this grid, then the estimation will generally be accurate. If, however, the unobserved location falls outside the eastern or western edges of the region, then the estimation will generally be inaccurate, because the value should either be smaller or larger than the minimum or maximum observed values of that variable, given the overall trend. Because the spatial relationship between the observed values of variable being interpolated are not taken into consideration, inverse distance weighted interpolation is not as accurate as more complex approaches to interpolation, such as ordinary kriging. Nevertheless, inverse distance weighted interpolation is a simple and usually effective method for spatial interpolation.

24.5.2 Ordinary Kriging

Ordinary kriging is the most common geostatistical method for interpolating the value of a spatially referenced variable at an unobserved location (Isaaks and Srivastava 1989; Bivand *et al.* 2008; Wackernagel 2010). Ordinary kriging estimates the value of a variable at a particular unobserved location by taking a weighted average of the values of the variable at observed locations, where these weights are based on both the distance separating that observed location from the unobserved location and the theoretical variogram for that variable, which provides a model of how the value of the variable changes across space. This contrasts with inverse distance-weighted interpolation, which only takes into consideration the values of the variable at observed locations and their distance from the unobserved location.

Specifically, ordinary kriging estimates the value of a variable at an unobserved location \hat{x}_0 as an unbiased linear combination of the weighted values of the variable across n observed locations as

$$\hat{x}_0 = \sum_{i=1}^n w_i x_i$$

where n is the number of observed locations, x_i is the value of the variable at the observed location i , and w_j is the kriging weight associated with location i , which is based on the value of the theoretical variogram for the distance separating the unobserved location and location i .

The first stage of ordinary kriging, therefore involves computing a theoretical variogram for the variable being interpolated, as described in Section 24.4. A theoretical variogram is necessary, rather than an empirical variogram because ordinary kriging requires that the value of the variogram be defined for the distances between the unobserved location and each of the observed locations, which are not generally instantiated in the empirical variogram. Given a theoretical variogram, the kriging weights can then be calculated for each of the observed locations, based on the distance separating that location from the location where the variable is being estimated, by solving the ordinary kriging equation system

$$\begin{cases} \sum_{j=1}^n w_j \gamma(x_i - x_j) + \mu = \gamma(x_i - x_0) & \text{for } i = 1, \dots, n \\ \sum_{j=1}^n w_j \end{cases}$$

where $\gamma(x_i - x_j)$ is the value of the theoretical variogram for the distance separating the observed locations x_i and x_j , $\gamma(x_i - x_0)$ is the value of the theoretical variogram for the distance separating the observed location x_i and unobserved location x_0 , and μ is a Lagrange multiplier.

The first n equations in the ordinary kriging system equate the value of the theoretical variogram for the distance between the unobserved location and one observed location to a linear combination of the weighted values of the theoretical variogram for the distances between that observed location and every other observed location. The final equation in the ordinary kriging system requires that the kriging weights sum to 1 to minimize the mean estimation error, thereby fulfilling what is known as the *unbiasedness condition*. In turn, because this equation system contains one more equation than unknowns, a Lagrange multiplier is introduced to allow the system to be solved to obtain the kriging weights.

Ordinary kriging was used to estimate the local G_i^* z-scores for *not* contraction across the same seven unobserved locations as in the previous section. The estimates are very similar to the values estimated through inverse distance weighted interpolation, and results in map that is identical to Figure 24.9, except that the location in Oregon is estimated as having a value larger than 1.96. Note that the estimates do not change substantially when other variogram models are used to fit the theoretical variogram.

Ordinary kriging was also used to interpolate the G_i^* z-scores for *not* contraction over a 250,000 regularly spaced grid of locations, as mapped in Figure 24.10. This application of ordinary kriging allows for high-density visualization of regional linguistic variation. In addition, because the values of the variable are plotted based on quartiles, the line separating the regions corresponding to second and third quartiles can be interpreted as an isogloss, dividing the Eastern United States where full *not* is relatively more common from the Western United States where *not* contraction is relatively more common. The use of ordinary kriging, coupled with a local spatial autocorrelation analysis, can therefore be used in this manner to plot isoglosses using a fully automated, replicable, and statistically justified procedure.

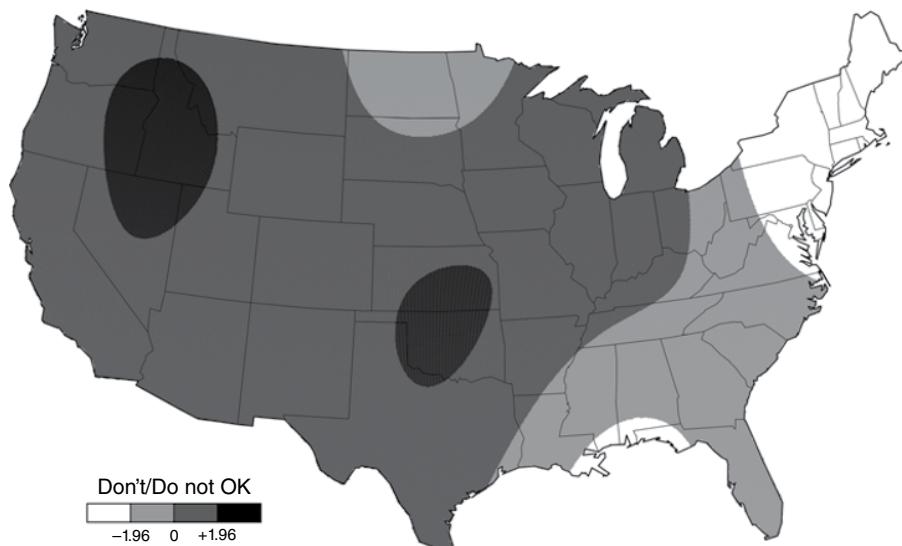


Figure 24.10 *Not contraction ordinary kriging.*

24.6 Conclusion

This chapter has provided an introduction to a range of common spatial statistics that have clear applications in dialectology, including methods for spatial autocorrelation analysis, variogram analysis, and spatial interpolation. Although these methods are relatively uncommon in dialectology, they constitute a powerful suite of techniques for the spatial analysis of linguistic variation, which overcome several limitations with traditional approaches to the analysis of regional linguistic variation. Specifically, spatial statistics allow for the analysis of individual linguistic variables to be conducted in an objective, standardized, and replicable manner. Spatial statistics are therefore especially useful when analyzing large datasets, where consistency and efficiency is often difficult to achieve through the manual analysis of dialect maps. Perhaps most important, spatial statistics can be used to prepare large dialect datasets for multivariate analysis, as is common in dialectometry, both by imputing missing values and by identifying underlying patterns of spatial clustering in the maps for individual linguistic variables so that the multivariate analysis can focus on the identification of common patterns of regional linguistic variation (see Grieve 2009, 2014; Grieve *et al.* 2011, 20012). There are also various other spatial statistics that have yet to be applied in dialectology, including additional methods for spatial autocorrelation and interpolation and techniques for point pattern analysis and spatial regression, which will hopefully be applied to the analysis of regional linguistic variation in future research.

Although spatial statistics are useful methods for the analysis of regional linguistic variation, it is important to remember that these statistics must be applied with care and do not replace the judgment of the dialectologist. In fact, like all methods for statistical analyses, the application of spatial statistics requires both the careful consideration of the data before the analysis begins and the careful interpretation of the results after the analysis is completed. Furthermore, applying spatial statistics generally requires that numerous decisions be made during analysis in order to maximize the accuracy of the results. The judgment of the dialectologist is therefore relevant at all stages of a spatial analysis. Nevertheless, applying spatial statistics allows for the decisions made by the dialectologist to be scrutinized and the effect

of making different decisions can be examined in a systematic way, leading to more reliable, unbiased, and defensible results. Finally, it is important to remember that despite the relatively definitive patterns identified by these methods, similar to the plotting of isoglosses in traditional dialect studies, the raw maps under analysis are rarely so clear. Spatial patterns identified by these statistics must therefore be interpreted accordingly—as simplifications of the reality that they represent. Such simplifications are often necessary to obtain a complete understanding of regional variation in a particular variety of language, but the underlying complexity of regional linguistic variation should never be forgotten.

REFERENCES

- Anselin, Luc. 1995. Local indicators of spatial association—LISA. *Geographical Analysis*, 27: 93–115.
- Asnaghi, Costanza. 2013. *An Analysis of Regional Lexical Variation in California English Using Site-Restricted Web Searches*. Unpublished Ph.D. Dissertation. University of Leuven.
- Bivand, Roger S., Edzer Pebesma, and V. Gómez-Rubio. 2008. *Applied Spatial Data Analysis with R*. Springer.
- Chun, Yongwan, and Daniel A. Griffith. 2013. *Spatial Statistics and Geostatistics: Theory and Applications for Geographic Information Science and Technology*. Sage.
- Cliff, Andrew David, and J. Keith Ord. 1969. The problem of spatial autocorrelation. In A. J. Scott (Ed.), *London Papers in Regional Science 1, Studies in Regional Science*, 25–55.
- Cliff, Andrew David, and J. Keith Ord. 1973. *Spatial Autocorrelation*. Pion.
- Cliff, Andrew David, and J. Keith Ord. 1981. *Spatial Processes: Models & Applications*. Pion.
- Cressie, Noel A. C. 1993. *Statistics for Spatial Data*. John Wiley & Sons.
- Fortin, Marie-Josée, and Mark R. T. Dale. (Eds.). 2005. *Spatial Analysis: a Guide for Ecologists*. Cambridge University Press.
- Fotheringham, A. S., Chris Brunsdon, and Martin Charlton. 2000. *Quantitative Geography: Perspectives on Spatial Data Analysis*. Sage.
- Geary, Roy C. 1954. The contiguity ratio and statistical mapping. *The Incorporated Statistician*, 5: 115–146.
- Getis, Arthur, and J. Keith Ord. 1992. The analysis of spatial association by use of distance statistics. *Geographical Analysis*, 24: 189–206.
- Grieve, Jack. 2009. *A corpus-based regional dialect survey of grammatical variation in written Standard American English*. Unpublished Ph.D. Dissertation. Northern Arizona University.
- Grieve, Jack. 2011. A regional analysis of contraction rate in written Standard American English. *International Journal of Corpus Linguistics*, 16: 514–546.
- Grieve, Jack. 2012. A statistical analysis of regional variation in adverb position in a corpus of written Standard American English. *Corpus Linguistics and Linguistic Theory*, 8: 39–72.
- Grieve, Jack. 2013. A statistical comparison of regional phonetic and lexical variation in American English. *Literary and linguistic Computing*, 28: 82–107.
- Grieve, Jack. 2014. A comparison of statistical methods for the aggregation of regional linguistic variation. In B. Szemrecsanyi & B. Wälchli (Eds.), *Aggregating Dialectology, Typology, and Register Analysis: Linguistic Variation in Text and Speech*. Walter de Gruyter.
- Grieve, Jack, Dirk Speelman, and Dirk Geeraerts. 2011. A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change*, 23: 193–221.
- Grieve, Jack, Dirk Speelman, and Dirk Geeraerts. 2013a. A multivariate spatial analysis of vowel formants in American English. *Journal of Linguistic Geography*, 1: 31–51.
- Grieve, Jack, Costanza Asnaghi, and Tom Ruette. 2013b. Site-restricted web searches for data collection in regional dialectology. *American Speech*, 413–440.
- Haining, Robert. 2003. *Spatial Data Analysis: Theory and Practice*. Cambridge University Press.
- Isaaks, Edward H., and R. Mohan Srivastava. 1989. *Applied Geostatistics*. Oxford University Press.
- Kretzschmar, William A. 1996. Quantitative areal analysis of dialect features. *Language Variation and Change*, 8: 13–29.

- Kretzschmar, William A. 2003. Mapping southern English. *American Speech*, 78: 130–149.
- Labov, William. 1972. *Sociolinguistic Patterns*. University of Pennsylvania Press.
- Lee, Jay, and Kretzschmar, William A. 1993. Spatial analysis of linguistic data with GIS functions. *International Journal of Geographical Information Science*, 7: 541–560.
- Moran, P. A. P. 1948. The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B*, 10: 243–251.
- Nerbonne, John. 2009. Data-driven dialectology. *Language and Linguistics Compass*, 3: 175–198.
- Ord, J. Keith, and Arthur, Getis. 1995. Local spatial autocorrelation statistics: distributional issues and an application. *Geographical Analysis*, 27: 286–306.
- Ripley, Brian D. 2005. *Spatial Statistics*. John Wiley & Sons.
- Sibler, Pius, Robert, Weibel, Elvira, Glaser, and Gabriela, Bart. 2012. Cartographic visualization in support of dialectology. In *Proceedings of AutoCarto 2012*, Columbus, Ohio.
- Szmrecsanyi, Benedikt. 2013. *Grammatical variation in British English dialects: a study in corpus-based dialectometry*. Cambridge University Press.
- Tamminga, Meredith. 2013. Phonology and morphology in Dutch indefinite determiner syncretism: Spatial and quantitative perspectives. *Journal of Linguistic Geography*, 1: 115–124.
- Wackernagel, Hans. 2010. *Multivariate Geostatistics: An Introduction with Applications*. Springer.

Section 3 – Data Introduction

CHARLES BOBERG

Dialectology is a field that, notwithstanding the importance of theory and methodology, has always been firmly rooted in its data: that is, in the facts of regional variation at different levels of linguistic structure in each language. It is therefore to that aspect of the field that we now turn in this last section of the handbook. Data on dialect variation are valuable not only to dialectologists, for whom they justify and motivate methods and theories, but also to a wide range of people outside the field, including both scholars and non-academics. They hold an obvious academic interest for closely allied disciplines like sociolinguistics, historical linguistics, and theoretical linguistics, but in many cases also for more distantly related concerns like anthropology, ethnography, history, literature, or film and cultural studies. Many members of the applied linguistics community, from language teachers to speech pathologists to dialect coaches, also share a professional interest in dialect variation. Beyond these vocational groups, dialect differences hold intrinsic interest for much of the general public: they can even be a subject of popular conversation and entertainment and some amateur dialectologists take great delight in noticing and commenting on—or imitating—how people from different regions speak. How, then, *do* the worlds' languages differ by region—according to the experts—and what have been the major achievements and concerns of the tradition of published research on each of them? These questions are the main subject of the 12 chapters that make up this section.

An exhaustive account of regional differences in all of the world's languages, or even a substantial proportion of them, would obviously take much more space than is available here. In planning this section, it was therefore necessary to be selective. Our first decision was to make this a section of 12 chapters, equal in that respect to the preceding sections, thereby striking what we thought would be the right balance between breadth—covering as many languages as possible—and depth—allowing something substantial to be said about those languages. Limits on the overall length of the volume, however, forced us to accept an even tighter limit—6,000 words—on the length of these chapters than was applied to the previous sections, but we felt that this was justified by the extra coverage it allowed. Our next decision concerned which languages to focus on. We were guided in this respect by two criteria that together provide a rough indication of relative prominence: number of speakers and quantity of published research. It seemed important to have some specific and coherent coverage, beyond incidental references in chapters on other topics, of both the world's most widely spoken languages and the most prolific language-specific traditions of dialect research. In some cases, coverage was increased—we hope at minimal

cost—by assigning chapters to groups of several genetically related languages that, to some extent, share a common research tradition. Authors of these chapters faced an even greater challenge than others in finding a balance among the languages they had to discuss and in fitting a great deal of material into a limited space. We trust our readers will agree with us that these authors have met that challenge admirably.

Notwithstanding the general need for concision, in one case, that of English, the multinational—indeed now global—nature of the language and the utterly enormous quantity of research were seen to justify two chapters. The first chapter in this section, Chapter 25 by Kevin Watson, therefore deals with British English and its close relations in the Southern Hemisphere, while the second, Chapter 26, which was my own responsibility, treats North American English; that is, English in the United States and Canada. Sadly, we were not able to include a treatment of the many other varieties of English spoken around the world today, but these have been well studied elsewhere, as mentioned in the general introduction to this book. Moving to Europe, we encounter the first of our multi-language chapters: Chapter 27, dealing with the continental European Germanic languages, written by Sebastian Kürschner. Though Danish, Dutch, and Swedish might easily have justified their own chapters, lack of space demanded that they be treated here in combination with German, which necessarily takes pride of place, given its dominant position both in the history of dialectology and in modern Europe. Readers interested in Dutch in particular will find several additional references to Dutch dialects in the preceding sections, while those interested in Danish are referred additionally to Tore Kristiansen on Sociodialectology (Chapter 6). The individual and more or less equal prominence of the major Romance languages seemed to demand independent treatment, so we have separate chapters on French (Chapter 28 by Damien Hall) and the dialects of Italy (Chapter 29 by Tullio Telmon), while Spanish is combined with Portuguese (Chapter 30 by John Lipski). Like the Germanic languages, the Slavic languages are given group coverage, with a primary focus on Russian, in Chapter 31, by Vladimir Zhobov and Ronelle Alexander.

An important consideration in planning this section was to broaden our perspective beyond the European, British, and North American—that is, “western”—examples that tend to dominate modern dialectological thought, by reserving the last five chapters of this section for some of the most prominent non-western languages in the world today. We hope that these chapters will hold a special interest for many of our readers, introducing them to new sets of data and distinct traditions of study, in some cases, like those of China or Japan, arguably longer-established and more prolific than those of Europe. If exposure to this greater diversity of data and research leads a few readers to think differently about dialectology, to form new hypotheses or undertake new studies, so much the better. It is in this spirit of potential cross-fertilization, as well as out of simple respect for the non-western majority of the world’s population, that we include here chapters on the dialects of Arabic (Chapter 32 by Enam Al-Wer and Rudolf de Jong); the Indic languages (Chapter 33 by Ashwini Deo); Chinese (Chapter 34 by Chaoju Tang); Japanese (Chapter 35 by Takuichiro Onishi); and Malay/Indonesian (Chapter 36 by Alexander Adelaar). Again, these choices were constrained by available space: chapters on Korean, Vietnamese, Javanese, Tamil, Persian, Turkish, and many other languages would have been highly desirable, but including all of these, or even a few of them, would have made satisfactory coverage of the languages we did include even more difficult to attain than it already was.

The nature of this section makes it less susceptible to general commentary than the previous sections, so I will be comparatively brief. There are few overarching theoretical or methodological themes to identify and discuss; rather, the section is notable for its diversity. This diversity was deliberately encouraged by allowing the contributors maximal freedom in structuring their chapters, instead of imposing a common format. As a result, the chapters vary widely in their emphases, reflecting a mixture of personal choices and varying national

or regional traditions and demonstrating the many different ways in which dialect study may be usefully approached. In some places, dialectology has a long tradition: the oldest example discussed in this section is the work of Yang Xiong in China 2,000 years ago, referred to by Tang as a "milestone for dialectal research not only in China but also in the world." In other places, it is comparatively recent: the research on Malay/Indonesian reviewed by Adelaar begins in the 1980s. Some cultures evidently cherish dialect variation and its analysis: Onishi discusses a geolinguistic research tradition in late twentieth-century Japan that produced 400 atlases and 30,000 maps, an astonishing quantity. Other cultures are evidently less comfortable with this topic: Al-Wer and De Jong observe that, "Departments of Arabic in Arab universities do not generally cater for research on Arabic dialects, which reflects widely held beliefs in the Arab world that the dialects are but a debased offshoot of standard Arabic and studying them is a sign of corruption." Most research on Arabic dialects is therefore done by non-Arabs, or at least by scholars working outside the Arab world, quite a different situation to that of research on English, French, or German, which has been a largely intra-national concern. Some chapters are necessarily more sociolinguistic in their emphasis, dealing with questions about national standards and language policy, whereas others are in a better position to focus on purely regional variation; we believe that this difference is a natural reflection of different dialect landscapes around the world.

One common theme that arises in several chapters is the difficulty of drawing clear boundaries between dialect regions and clear distinctions between dialects and languages, issues touched on in the general introduction to this book as well as in several chapters in previous sections. In this section, we see these issues addressed in the chapters on the Germanic languages, the dialects of Italy, Spanish and Portuguese, the Slavic languages, the Indic languages, and Chinese, among others. A particularly illustrative case is furnished by Zhobov and Alexander, who point out how non-linguistic considerations often trump mutual intelligibility as a criterion for deciding the status of dialect and language divisions in Eastern Europe: despite how ethnic populations of the former Soviet Union may feel, "The East Slavic languages – Russian, Ukrainian and Belarusian – constitute a dialect continuum, with a high degree of mutual understanding." Meanwhile, in the South Slavic region, the disintegration of Yugoslavia in the 1990s into several separate countries saw the split of Serbo-Croatian into three separate standards, with a fourth proclaimed in 2006, and even though Macedonian has been a fully independent language for over half a century, it is still perceived by many Bulgarian linguists as a dialect of Bulgarian, just as Serbs in the past viewed it as "Old Serbian." This is perhaps a good way of reinforcing the idea that the study of dialects is not necessarily just an academic subject detached from the daily concerns of ordinary people; rather, dialects are often a matter of heated dispute, fierce pride, or passionate delight for large sections of the general population.

25 Dialects of British and Southern Hemisphere English

KEVIN WATSON

25.1 Introduction

The main focus in this chapter is on dialects of British English, beginning with the classic dialectological studies that had a wide geographical scope before considering more recent work, which has tended to concentrate on smaller sets of localities. The chapter also discusses dialects of Australia, New Zealand, and South Africa, because of the family ties between English in these countries and English in Britain. The Southern Hemisphere varieties are considered together because they are all relatively young with similar linguistic systems that are quite different from dialects in North America, which was settled much earlier and which is covered in a separate chapter (Boberg, this volume). There has been an enormous amount of research on British and Southern Hemisphere English, which necessitates a selective approach in the space available. I have, therefore, chosen to focus on studies of phonological variation, which has arguably been the main concern of recent research in British sociodialectology.

During the development of English, Britain was invaded and reinvaded, and boundaries between settlers were drawn and redrawn (see e.g., Baugh and Cable 2013). Some of the main dialectal divisions we see in Britain today are the result of these settlement patterns, from as early as the fifth century when, to give an over-simplified example, the arrival of the Angles, Saxons, and Jutes pushed Celtic speakers to the edges of Britain (Scotland in the north and Wales and Cornwall in the west). Standard English began to develop in the fifteenth century, emerging from the southeast of England, which was the most populous area and included high status institutions such as Cambridge and Oxford universities, which conferred prestige on its dialect. Over time, linguistic innovations spread from this region, particularly London, which Wells (1982: 301) describes as “the most influential source of phonological innovation in England and perhaps in the whole English-speaking world.”

The story of English in the southern hemisphere is also one of invasion and settlement, and of the influence of London English. English was introduced to Australia in 1788 with the arrival of several hundred British convicts. Many of these early settlers came from London, but the population expanded rapidly and the transportation process continued, and by 1840, almost 700,000 English and Irish people from London, Lancashire, Dublin and elsewhere had arrived in New South Wales (Turner 1994, Kiesling 2004). English arrived a little later in New Zealand, its rapid growth promoted by the signing of the Treaty of Waitangi in 1840, which marked the beginning of British sovereignty and resulted in extensive migration from Britain. By 1858, the European population outnumbered the Māori, and by 1881, the number

of Europeans reached half a million (Bauer 1994, Gordon *et al.* 2004). As in Australia, most migrants to New Zealand were from the southeast region of England, but there was also a considerable number of people from Scotland, and a smaller number from the north of England. English was introduced to South Africa in 1795, but it was not until 1820 that the first large cohort of English-speaking settlers arrived at the Eastern Cape, who, again, were largely from southeast England but also came from the southeast Midlands, the West Country, Yorkshire, and Scotland. A second wave occurred from 1849–1851, when upper- and middle-class settlers, probably speaking an early form of Received Pronunciation, arrived in Natal (Branford 1994, Lass 2004: 370–371). The relative youth of the southern hemisphere varieties of English means there is nothing like the sort of regional variation we find in Britain or North America, but, as we will see, it is wrong to say there is none at all.

25.2 The Study of Dialects in Britain and the Southern Hemisphere

The first major dialectological survey of Britain, which focused on phonological variation, was Ellis's *The Existing Phonology of English Dialects* (EPED), published in 1889. Data was collected from almost 500 localities in England, Scotland, Wales, and Ireland. The EPED is an impressively extensive volume but it has not been taken up with the sort of vigor we might expect given its historical status. Instead, it has attracted criticism, not least because the distribution of localities is uneven, and Ellis's *palaeotype* transcription system is difficult to interpret (see Petyt 1980: 73, Maguire 2012). However, when some of Ellis's results have been systematically compared to later work, there are similarities, indicating the value of the EPED as a historical tool (see e.g., Jones 2002). Maguire (2012) successfully maps a subset of the EPED data, making it more accessible (see <http://www.lel.ed.ac.uk/EllisAtlas/>).

The second major study is Wright's *English Dialect Dictionary* (EDD), published as a multivolume work from 1898 to 1905. The main focus was on lexical variation, but the final volume, *English Dialect Grammar* (EDG; Wright 1905), focused on phonology and grammar. Wright based much of EDG on the results of postal questionnaires, and also relied heavily on Ellis's EPED but used a more standard IPA-like transcription system rather than Ellis's palaeotype. Like the EPED, EDG has limitations, many of which are noted by Wright (1905: vi), such as the fact that informants were not always natives of the locality in question. Despite this limitation (and others, see Wakelin 1972: 47 and Petyt 1980: 81), EDG is an important reference work in the dialectology of Britain. Work at the Universität Innsbruck has begun to digitize the EDD, and thereby to greatly increase its functionality (see <http://www.uibk.ac.at/anglistik/projects/edd-online/>).

The best-known dialect survey in England is the *Survey of English Dialects*, an ambitious project led by Eugen Dieth and Harold Orton (see Orton 1962). Dieth had been critical of Ellis's earlier work and set out to fill what he saw as notable gap in the dialectology literature. Between 1948 and 1961, nine fieldworkers visited 311 localities across England, and used a questionnaire consisting of over a thousand questions in nine different categories (e.g., the farm, social activities; see Orton 1962 for the full questionnaire). The localities were mostly rural but there was an impressive geographical spread, with the four “corners” of England represented by Lowick (Northumberland), Longtown (Cumbria), St Buryan (Cornwall), and Staple (Kent). Data was also collected from the Isle of Man. Fieldworkers sought older males who had not spent significant periods of time away from home (hence the term NORMs: non-mobile older rural males. See Chambers and Trudgill (1998: 29–31) for discussion of this methodology).

Results from the SED were published in multiple volumes. There are four volumes of Basic Material, which are lists of informant responses sorted by locality, and several other volumes have provided maps of SED data, either with a focus on only part of the country

(e.g., Kolb 1966, on northern England) or on a particular linguistic level (e.g., Orton and Wright's (1974) *A Word Geography of England*, on lexical variation). Orton, Sanderson, and Widdowson (1978) is the main source for England-wide maps of phonological, morphological, syntactic, and lexical variables.

In the 1970s, dialectological work in England slowed, at least in terms the appearance of new work that examined the geographical distribution of linguistic features. This was partly due to a shift in focus to a more socially informed dialectology, inspired by Labov's work in the USA (Labov 1966, see also Kristiansen, this volume), and the move away from the examination of NORMs toward the inclusion of a greater number of speakers exemplifying a much broader spectrum of social characteristics. An early attempt at collecting this sort of data is the Tyneside Linguistic Survey (TLS; see Allen *et al.* 2007). Although the TLS had different research questions from Labov's and did not aim to follow a Labovian methodology, there are parallels: speakers were audio recorded recounting oral histories, and were asked their opinions about local dialect. Detailed social information was also collected. Much of the TLS data has been lost over time, but 114 interviews have been saved and are now incorporated into the *Diachronic Electronic Corpus of Tyneside English* (<http://research.ncl.ac.uk/decte/>). The earliest study in England that followed Labov's methods closely is Trudgill's (1974) investigation of Norwich. Trudgill collected data from 60 informants, including males and females from a range of age and social groups. This is a marked departure from the traditional dialectological methodologies, and signaled the beginning of sociodialectological work in Britain. It also marked a shift away from a focus on the national picture. Instead we find a range of detailed studies of single localities, such as Knowles (1973) on Liverpool, Newbrook (1986) on West Wirral, and Petyt (1985) on West Yorkshire.

More recently, we have seen something of a resurgence of dialectological work in Britain. One large-scale enterprise is the *BBC Voices Project*, a collaboration between the University of Leeds and the BBC (see Upton and Davies 2013, and <http://www.bbc.co.uk/voices/>). Recordings of over 1,200 people were made in 2004–2005 by BBC radio journalists using an interview protocol based on the *Survey of Regional English* (SuRE) methodology (Llamas, this volume). In another project, still in progress (Maguire 2009 and <http://www.lel.ed.ac.uk/~wmaguire/survey/survey.html>), an online questionnaire is used to gather self-reported usage from informants across Britain, data from which are plotted and compared to existing dialectological results from the SED and elsewhere.

In Scotland, we began to see the published results of the *Linguistic Survey of Scotland* (LSS) in the 1970s. The LSS mainly focused on lexical and phonological variation, the former being investigated using postal questionnaires, the latter using trained fieldworkers in 250 localities across Scotland and northern England. The main publications are Mather and Speitel (1975, 1977, and 1986). As in England, there has not been a major dialectological survey across Scotland since the LSS, and instead the focus has been on socially informed dialectology, in, for example, Glasgow (Macaulay 1977) and Edinburgh (Romaine 1975). A recent cross-locality study is the *Accent and Identity on the Scottish/English Border* project (AISEB, see Watt, Llamas, and Johnson 2014), in which two localities in England (Carlisle and Berwick-upon-Tweed) are compared with two partner localities in Scotland (Gretna and Eyemouth), from the dual perspectives of the geographical distribution of certain key phonological features, such as the overtly realized coda /r/, and the social factors that may affect their usage (e.g., identity and attitudinal factors, see Llamas 2010). In Wales, an important study is the *Survey of Anglo-Welsh Dialects* (SAWD). The SAWD followed the SED methodologies quite closely, examining speech in rural localities in Wales and collecting data from older informants (see Parry 1979 and Penhallurick 1991). Large studies of Ireland are more recent, perhaps the largest being Hickey's (2004) *A Sound Atlas of Irish English*. This is an ambitious combination of traditional dialectology and socially-informed methods, offering over 1,500 recordings from speakers from each of the 32 counties in the island of Ireland.

Moving to the southern hemisphere, studies of Australian English (AusE), New Zealand English (NZE), and South African English (SAE) have been much less concerned with regional variation than those in Britain. Indeed, the lack of regional forms in these varieties is often reported, with variation instead said to be conditioned by age, gender, social class, and ethnicity. Each of these social variables has been studied across the southern hemisphere, to a greater or lesser extent, so in a way sociolinguistic work in these localities has been from its outset broadly sociodialectological in its approach. In Australia, the first large-scale study was that of Mitchell and Delbridge (1965), who surveyed a large sample of school students from across the country. Since then, there have been single-locality studies of Sydney (Horvath 1985), among others, but there have been no other countrywide studies, although one—*AusTalk*—is underway (see <https://austalk.edu.au/>). Considerations of regional variation in AusE have recently appeared, including Bradley (2004), which provides an overview, Horvath and Horvath (2002), which examines a wide geographical region (9 localities in Australia and New Zealand) but focuses on only one variable (/l/ vocalization—see below), and Cox and Palethorpe (1998, 2001), which examine vocalic differences in different areas of Sydney. In New Zealand, lay people feel strongly that there are many regional dialects (Nielson and Hay 2005), but linguistic work has focused instead on the transformation of NZE over time from a dialect mixture to a homogeneous focused variety (e.g., Trudgill 2004, Gordon *et al.* 2004). Exceptions are Bartlett (2003), a Labovian examination of the Southland region, widely said to be the only regional variety in New Zealand, and Marsden (2013), which examines coda /r/ in two rural communities in the lower and central North Island. The only study to date on the whole of New Zealand is Bauer and Bauer (2002), which used a postal questionnaire to examine lexical variation. In South Africa, it is common in the literature to see mention of, for example, “Cape English” and “Natal English” (see, e.g., Bekker 2012), but these labels are typically used as an indicator of social rather than regional varieties, with features of “Natal English” being perceived to be of relatively high status. Because of this social dimension, the focus of much early work was on “educated English” or “standard speech” (Lanham and Macdonald 1979), with little consideration of regional variation. It is only recently that geographical variation in South African English has come into focus, perhaps reflecting a growing sense that regional varieties are emerging: Bekker (2007) compares Johannesburg speakers with speakers from elsewhere, and Mesthrie *et al.* (2013, 2015) examine social and geographical variation across a large area (Cape Town, Durban, Johannesburg, Port Elizabeth, and Kimberly).

25.3 The Principal Linguistic Features of British and Southern Hemisphere Dialects

It is impossible in this short chapter to discuss every diagnostic feature of every variety of English, even if the focus is restricted to phonological variables. My approach is to broadly compare “north” with “south,” as follows: first I examine some of the phonological differences between English in Scotland and in England, and then between dialects in the north and south of England, also commenting where relevant on certain pockets of variation in the east and west of the country. Limited space prevents me from extending this analysis to Irish English; for the principal features of Irish varieties, see Hickey (2010). Finally, moving from the northern to the southern hemisphere, I illustrate some of the important phonological characteristics of AusE, NZE and SAE. In many of the comparisons, I use Received Pronunciation (RP; regarded as the standard accent of English in England) as a reference (see Upton 2008 for discussion of RP).

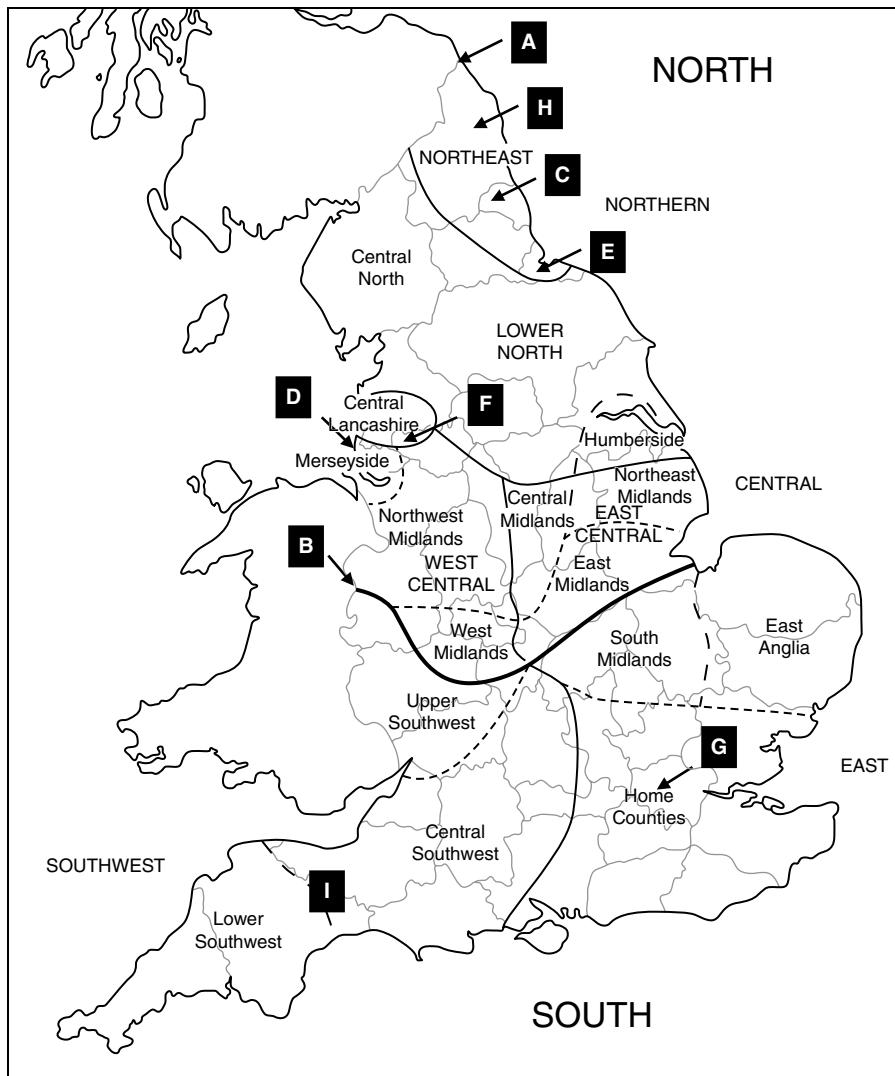


Figure 25.1 Trudgill's dialect boundaries of England (adapted from Trudgill 1999: 65). Letters A to I identify localities and boundaries mentioned in the main text.

The boundary between England and Scotland (see "A" in Figure 25.1) has been noted since the very earliest dialectological work. In the EPED, Ellis separates the "Lowland" dialect area (mainly localities in Scotland) from the "northern" dialect area (localities in northern England), but the isogloss (or, in Ellis's terms, *transverse line*) does not follow the political boundary exactly (e.g., parts of Northumberland and what is now Cumbria are assigned to the Lowland area even though they are English localities). Ellis (1889: 495) observes that the voiceless velar fricative, which he calls (kh), and rhotic (r), have "practically vanished" from the north of England, particularly in the east, "though on passing... the L line [Lowland], (kh) as well as (r) is strong." Of course, there are many other differences between the phonological systems of English in Scotland and England. For vowels, KIT is typically [ɪ] in Scottish English, but it can be lowered and/or centralized. TRAP/BATH/PALM have the same vowel, [ə], for many Scottish English speakers, unlike in RP where we

would expect TRAP [a], BATH [ɑ:] and PALM [ɑ:]. LOT/CLOTH/THOUGHT also have the same vowel in Scotland, [ɔ], whereas in (modern) RP we are likely to find LOT/CLOTH [ɒ] and THOUGHT [ɔ:]. Unlike RP FOOT [ʊ] and GOOSE [u:], Scottish English typically has [u] for both of these sets, which Wells (1982: 401) describes as the most important diagnostic feature of Scottish English. FACE and GOAT are usually the monophthongs [e] and [o] respectively in Scotland, but they can also be diphthongs for some speakers (see Schützler 2014). RP has the diphthongs [eɪ] and [əʊ] for these sets, but monophthongs are common in the north of England. Another well-known characteristic of Scottish English vowels is the so-called “Scottish vowel length rule” (SVLR), by virtue of which vowels are longer in certain environments (e.g., before voiced fricatives and /r/, before pauses, and before morpheme boundaries) such that the /i/ vowel is longer in “agreed” than “greed” (Stuart-Smith 2008: 58). As for consonants, the main distinction between Scottish English and English in most of England is the presence of the overtly realized coda /r/ in Scotland, as mentioned above, but this feature appears to be reducing in frequency, particularly in the west (e.g., Gretna, see Watt *et al.* 2014: 89). There is also phonetic variation in the realization of /r/, where [ɹ, r, ɾ] are all possible variants in Scotland. Other consonant features include the use of [m] in Scotland where modern RP and the rest of England are likely to have [w] in, for example, “whales,” and the use of [h] for /θ/ in, for example, “think” and “something,” which is unattested in England. These examples illustrate why the Scottish/English border has been said to host the “most numerous bundle of dialect isoglosses in the English-speaking world” (Aitken 1992: 895).

The boundary between northern and southern England is not an official one, but it is nevertheless an important psycho-social boundary for most English people. Its placement varies in the dialectological literature. Ellis’s *northern* dialect area covers “the entire North and East Ridings with some of the West Riding of Yorkshire, northern Lancashire, most of Cumberland and Northumberland, all of Westmoreland and Durham” (1889: 494) but the *Northern Counties* of the SED reach slightly further south, to include all of Lancashire and parts of modern-day Cheshire. The northern/southern England boundary is sometimes shifted even further south, with a line beginning at the Wash, in the east, at the boundary of East Anglia and Lincolnshire, and running west, just south of Birmingham to the Welsh border (see “B” in Figure 25.1). North of this line, FOOT and STRUT are homophones (e.g., “put” and “putt” have [ʊ]) but south of the line the two sets are distinct (e.g., “put” [ʊ] and “putt” [ʌ]/[ə]). The lack of the so-called FOOT/STRUT split is often used as the main criterion for defining the “linguistic north” (Trudgill 1999, Wells 1982). The distribution of the vowels in words like “last” is similar to that of FOOT/STRUT. In the north of England, the BATH lexical set usually patterns with TRAP (“path” [a], “pat” [a], “palm” [ɑ:]) but in the south of England BATH usually patterns with PALM (and, if the variety does not realize coda /r/, START: “path” [a:], “palm” [a:], “part” [ɑ:]). The phonetic quality of the BATH/PALM vowel changes across the south of England: in much of the southeast we find [a:], but in much of the southwest, [a:] or [ɛ:].

Other vocalic features of the north of England are not found across the whole northern region but in smaller geographical areas, and do as much to distinguish northern English accents from each other as they do to separate northern from southern varieties. FACE and GOAT, for example, as noted above, have diphthongs of various qualities in the south of England (see Watt and Altendorf 2008: 205–208) but monophthongs in much but not all of the north: diphthongs are common in the northeast (“C” in Figure 25.1), at least in older speakers (typically FACE [ɪə] and GOAT [ʊə] in Tyneside; see Watt and Milroy 1999) and in parts of the northwest (FACE [eɪ] and GOAT [əʊ] in Liverpool, “D” in Figure 25.1; see Watson 2007). The phonetic quality of NURSE, and the relationship between NURSE, NORTH, and SQUARE, also varies across the north of England. In much of the region, NURSE has [ɜ:], NORTH has [ɔ:] and SQUARE has [ɛə] (or [ɛ:]), broadly similar to the RP and southern English systems.

NURSE has a front vowel, [ɛ:], in Liverpool (Watson 2007; Watson and Clark 2013) and Middlesbrough ("E" in Figure 25.1, Beal, Burbano-Elizondo, and Llamas 2012: 32), and can have [ɔ:] in Newcastle ("C" in Figure 25.1), where it merges with NORTH for older working-class males (Watt and Milroy 1999). NURSE is merged with SQUARE in Liverpool and parts of Lancashire (e.g., St Helens) but the phonetic quality of the merger differs: in Liverpool the sets are merged to a front vowel (e.g., "her" // "hair" [ɛ:]) and in Lancashire ("F" in Figure 25.1) they are merged to a central vowel (e.g., "her" // "hair" [ɔ:]).

Unlike with vowels, very few consonant features clearly separate northern from southern English varieties, and instead are either widely spread across England or localized to small geographical areas. An example of the former is TH/DH-fronting, in which the interdental fricatives /θ, ð/ are realized as labiodentals [f,v]. This is extremely common across England and Scotland, and is thought to have begun in London ("G" in Figure 25.1) and diffused north (Kerswill 2003). The spread is continuing, and TH/DH-fronting is now attested in localities such as Liverpool which have been said to be resisting it. This is not to say there are no regionally restricted forms of TH/DH, however. In Liverpool, TH/DH-stopping (where /θ, ð/ are realized as [t,d]) is common. It is thought to have been introduced in that variety via contact with Irish dialects in the nineteenth century. A well-known example of a consonantal feature which is now found only in relatively localized areas is the overtly realized coda /r/. At the time of the SED, coda /r/ was realized in parts of the northeast (Northumberland ("H" in Figure 25.1) and slightly further south in Yorkshire), the northwest (Lancashire), along the south coast and in a relatively large area of the southwest ("I" in Figure 25.1). The phonetic quality of the /r/ also varied across the region: a uvular [ʁ] was common in Northumberland, an approximant [ɹ] was the typical form in the northwest and central part of the country, and the retroflex [ɾ] was common in the southwest. The geographical spread of dialects that realize coda /r/ has been reduced since the SED: it is now found in a much smaller pocket of the northwest, southwest and northeast, and usually only for older, male, working-class speakers (see e.g., Piercy 2007 on Dorset; Barras 2011 on East Lancashire). The phonetic quality of prevocalic /r/ also differs across the country: the majority realization is an approximant [ɹ], but a tap [ɾ] is common in Liverpool (Watson 2007), and a labiodental variant [v] is common in the southeast and increasingly further north, in Derby and Newcastle (Foulkes and Docherty 2002). With plosives, we find some realizations that are widespread across the country and others that are very localized. An example of the former is the realization of /t/ as [?], thought to be diffusing across the country. Other variants of /t/ are found only in the north of England. Examples include oral and glottal fricatives (e.g., [s, h]), found in Liverpool (which also has fricatives for /k/, Watson 2007; Clark and Watson 2016) and to a lesser extent Middlesbrough (Jones and Llamas 2008), and the approximant [ɹ], found in a restricted set of phonological environments in parts of Merseyside, Lancashire, Yorkshire, and parts of the northeast (see Clark and Watson 2011 and references therein).

Moving to the southern hemisphere, I begin a short discussion of AusE, NZE, and SAE by first commenting on the main differences between these Englishes, before noting some of the few regional differences within each variety. Arguably the most well-known feature of the southern hemisphere Englishes, present in NZE, SAE and, to a lesser extent, AusE, is the short front vowel shift, which sees TRAP and DRESS raised and (in New Zealand and South Africa) KIT centralized, resulting in, broadly, "had" [hɛd], "head" [hed], and "hid" [hɔd/ ɦid]. AusE KIT is raised rather than centralized, and this is often said to be the main distinguishing characteristic between it and NZE, even though TRAP and DRESS are also lower in AusE relative to NZE. In SAE, KIT is centralized but there are phonological environment effects: centralized [ɪ] is typical except when word-initial or when adjacent to velar consonants or [h], when [i] is likely. This system, according to Lass (2004: 375), is diagnostic of SAE. Like RP, AusE, NZE, and SAE have the FOOT/STRUT split, but STRUT can be fronter than in RP

(e.g., [ə]). Also similar to RP, BATH/PALM/START pattern together in NZE, SAE, and for many speakers of AusE, but there are phonetic differences: the vowel is front in NZE and AusE ([ɛ: ~ a:J]) and back and sometimes rounded in SAE ([ɑ: ~ ɒ:J]). Additionally, in AusE some BATH-class items have the vowel of TRAP, particularly when the vowel is followed by a nasal plus another consonant (e.g., *dance, plant*). NZE FLEECE is undergoing diphthongization, approximating [iɪ], but in SAE, FLEECE is a monophthong. NURSE is fronter and slightly higher in AusE, NZE and SAE than is typical in RP, and can be rounded, [œ:J]. Arguably the most notable characteristic of NZE diphthongs is the merger of NEAR and SQUARE, a well-established change in progress that differentiates NZE from AusE and that sees words like “bear, beer, hare, hear” produced with [ɪə] (Gordon and MacLagan 2001). In AusE and SAE, NEAR and SQUARE are distinct, and SQUARE can be monophthongal ([ɛ:J]). Phonologically, the consonantal systems of AusE, NZE, and SAE are very similar to that of RP, although there are phonetic differences. Generally speaking, AusE, NZE, and SAE do not realize coda /r/. In intervocalic position, /t/ is commonly a tap in AusE, NZE, and SAE, but in AusE and NZE we also find heavily aspirated and fricated variants (see Fiasson 2009 for NZE and Jones and McDougall 2009 for AusE). The glottal stop realization of /t/ is not as common in any of the southern hemisphere localities as it is in most of Britain, but there are reports of an increase of this variant in NZE (Holmes 1995). Coda /l/ is velarized and often fully vocalized in NZE and AusE, but in SAE /l/ vocalization is uncommon since /l/ tends to be non-velarized (clear) in coda position.

Finally, I consider regional variation within the southern hemisphere varieties. Bauer and Bauer (2002), in one of the few studies of this region with a wide geographical scope, posit three dialect areas in New Zealand: a northern region, which extends as far south as the central North Island; a central region, which extends to the middle of the South Island and includes Queenstown and Wanaka; and a southern region, which includes Otago and Southland. Bartlett (2003) presents evidence for the linguistic distinctiveness of the southern region or, more specifically, Southland, widely thought to be New Zealand’s only regional variety because of its overtly realized coda /r/. Bartlett (2003) shows that this feature is maintained in Southland, particularly in NURSE-class items, thus continuing to distinguish this region from others in New Zealand. Most recently, Marsden (2013) demonstrates the realization of coda /r/ in two towns within Bauer and Bauer’s central region—considerably further north than its traditional heartland. In Australia, the majority of studies divide AusE along social rather than geographical lines, following the early work of Mitchell and Delbridge (1965), identifying a continuum from “cultivated” English (the most prestigious), through to “general” and “broad” English (the least prestigious). It is therefore often difficult to tease out geographical effects. Cox and Palethorpe (1998, 2001), however, find differences between the Northern Suburbs and Western Suburbs of Sydney in the production of a range of vowels (e.g., in the Western Suburbs relative to elsewhere, /e, æ, ɔ, ɜ/ are raised, /ɔ, u/ are retracted and /u/ is fronted), even when social class is controlled, and argue that regional origin should be more tightly controlled in AusE phonetic work. A consonantal feature known to vary across Australia is /l/-vocalization, which is most frequent in South Australia (e.g., Adelaide), and least frequent in Brisbane and Melbourne (see Horvath and Horvath 2001, which also shows that /l/-vocalization is more frequent in New Zealand than any Australian locality). In South Africa, as noted above, it is also common to divide dialects along social rather than regional lines, following the Australian trichotomy: thus we have “conservative,” “respectable,” and “extreme” SAE (Lanham and Macdonald 1979, Lass 2004). However, recent work suggests that the northern suburbs of Johannesburg are emerging as a dialect area (see Bekker 2007 on fronted /s/), though it should be noted that this region is traditionally associated with wealth and prestige, again pointing to the difficulty of teasing apart geographical and social factors.

REFERENCES

- Aitken, A. J. 1992. Scottish English. In Tom McArthur (Ed) *The Oxford Companion to the English Language*, Oxford: Oxford University Press, 893–899.
- Allen, Will, Joan C. Beal, Karen P. Corrigan, Warren Maguire and Hermann L. Moisl. 2007. A linguistic ‘time capsule’: the newcastle corpus of tyneside English. In Beal, Joan C., Karen Corrigan & Hermann Moisl (Eds) *Creating and Digitizing Language Corpora*. Basingstoke: Palgrave. 16–48.
- Barras, William. 2011. *The sociophonology of rhoticity and r-sandhi in East Lancashire English*. Unpublished PhD dissertation, University of Edinburgh, UK.
- Barlett, Christopher. 2003. *The Southland variety of English: postvocalic /r/ and the BATH vowel*. Unpublished PhD Thesis, University of Otago.
- Bauer, Laurie. 1994. English in New Zealand. In Burchfield (Ed), 382–429.
- Bauer, Laurie and Winifred Bauer. 2002. *Playground Talk. Dialects and Change in New Zealand English*. Wellington, New Zealand: Victoria University of Wellington.
- Baugh, Albert, C. and Thomas Cable. 2013. *A History of the English Language*. 6th edition. London: Routledge.
- Beal, Joan, Lourdes Burbano-Elizondom and Carmen Llamas. 2012. *Urban North-East English: Tyneside to Teeside*. Edinburgh: Edinburgh University Press.
- Bekker, Ian. 2007. Fronted /s/ in general white South African English, *Language Matters*, 38(1): 46–74.
- Bekker, Ian. 2012. The story of South African English: a brief linguistic overview. *International Journal of Language, Translation & Intercultural Communication*, 1: 139–150.
- Bradley, David. 2004. Regional characteristics of Australian English: phonology. In Edgar Schneider, Kate Burridge, Bernd Kortmann, Rajend Mesthrie & Clive Upton (Eds). *A Handbook of Varieties of English: Volume 1 – Phonology*. Berlin/New York: Mouton de Gruyter. 645–655.
- Branford, William. 1994. English in South Africa. In Burchfield (Ed), 430–496.
- Burchfield, Robert. 1994. Ed. *The Cambridge History of the English Language: Volume 5 – English in Britain and Overseas, Origins and Development*. Cambridge: Cambridge University Press.
- Chambers, J. K. and Peter Trudgill. 1998. *Dialectology*. 2nd edition. Cambridge: Cambridge University Press.
- Clark, Lynn and Kevin Watson. 2011. Testing claims of a usage-based phonology with Liverpool English *t-to-r*. *English Language and Linguistics*, 15/3: 523–547.
- Clark, Lynn and Kevin Watson. 2016. Phonological leveling, diffusion, and divergence: /t/ lenition in Liverpool and its hinterland. *Language Variation and Change*, 28/1, 31–62.
- Cox, Felicity and Sallyanne Palethorpe. 1998. Regional variation in the vowels of female adolescents from Sydney. *Proceedings of the Fifth International Conference on Spoken Language Processing*, Sydney, 6: 2359–2362.
- Cox, Felicity and Sallyanne Palethorpe. 2001. The changing face of Australian English vowels. In David Blair & Peter Collins (Eds) *English in Australia*. Amsterdam: John Benjamins.
- Ellis, Alexander. 1889. *The Existing Phonology of English Dialects, Compared with that of West Saxon Speech*. New York: Greenwood Press.
- Fiasson, Romain. 2009. Intervocalic /t/ in New Zealand English. Paper presented at the La Phonologie de l’Anglais Contemporain conference, University of Toulouse, September 2009.
- Foulkes, Paul and Gerard Docherty. 2002. Another chapter in the story of /r/:'labiodental' variants in British English. *Journal of Sociolinguistics*, 4/1: 30–59.
- Gordon, Elizabeth, Lyle Campbell, Jennifer Hay, Margaret MacLagan, Andrea Sudbury, and Peter Trudgill. 2004. *New Zealand English: Its Origins and Evolution*. Cambridge: Cambridge University Press.
- Gordon, Elizabeth and Margaret MacLagan. 2001. Capturing a sound change: a real time study over 15 years of the NEAR/SQUARE diphthong merger in New Zealand English. *Australian Journal of Linguistics*, 21/2: 215–238.
- Hickey, Raymond. 2004. *A Sound Atlas of Irish English*. Berlin/New York: Mouton de Gruyter.
- Hickey, Raymond. 2010. *Irish English: History and Present Day Forms*. Cambridge: Cambridge University Press.
- Holmes, Janet. 1995. Glottal stops in New Zealand English: an analysis of variants of word-final /t/. *Linguistics*, 33/3: 433–463.

- Horvath, Barbara. 1985. *Variation in Australian English: the Sociolects of Sydney*. Cambridge: Cambridge University Press.
- Horvath, Barbara and Ronald J. Horvath. 2002. The geolinguistics of /l/ vocalization in Australia and New Zealand. *Journal of Sociolinguistics*, 6(2): 319–346.
- Jones, Mark, J. 2002. The origin of definite article reduction in northern English dialects: evidence from dialect allomorphy. *English Language and Linguistics*, 6(2): 325–345.
- Jones, Mark, J. and Carmen Llamas. 2008. Fricated realisations of /t/ in Dublin and Middlesbrough English: an acoustic analysis of plosive frication and surface fricative contrasts. *English Language & Linguistics*, 12/3: 419–443.
- Jones, Mark J. and Kirsty McDougall. 2009. The Acoustic character of fricated /t/ in Australian English: a comparison with /s/ and /ʃ/. *Journal of the International Phonetic Association*, 39/3: 265–289.
- Kerswill, Paul. 2003. Dialect levelling and geographical diffusion in British English. In Dave Britain and Jenny Cheshire (Eds) *Social Dialectology. In Honour of Peter Trudgill*. Amsterdam: Benjamins. 223–243.
- Kiesling, Scott, F. 2004. English input to Australia. In Hickey (Ed), 418–439.
- Kolb, Eduard. 1966. *Phonological Atlas of the Northern Region*. Bern: Francke.
- Kortmann, Bernd and Clive Upton (2008, Eds) *Varieties of English: the British Isles*. Berlin: Mouton de Gruyter.
- Knowles, Gerald. 1973. *Scouse: the urban dialect of Liverpool*. Unpublished PhD dissertation, University of Leeds, UK.
- Labov, William. 1966. *The Social Stratification of English in New York City*. Washington DC: Center for Applied Linguistics.
- Lanham, L. and Macdonald, C. (1979), *The Standard in South African English and Its Social History*. Heidelberg: Julius Groos Verlag.
- Lass, Roger. 2004. South African English. In Hickey (Ed), 363–386.
- Llamas, Carmen. 2010. Convergence and divergence across a national border. In Llamas, Carmen & Dominic Watt (Eds) *Language and Identities*. Edinburgh: Edinburgh University Press, 227–236.
- Macaulay, Robert. 1999. *Language, Social Class and Education: a Glasgow Study*. Edinburgh: Edinburgh University Press.
- Maguire, Warren. 2009. *Investigating phonological variation in contemporary English in the British Isles*. UK Language variation and Change conference, 1–3 September 2009, University of Newcastle. See <http://www.lel.ed.ac.uk/~wmaguire/survey/survey.html>.
- Maguire, Warren. 2012. Mapping the existing phonology of English dialects. *Dialectologia et Geolinguistica*, 20: 84–107.
- Marsden, Sharon. 2013. *Phonological variation and the construction of regional identities in New Zealand English*. Unpublished PhD thesis, Wellington, New Zealand: Victoria University of Wellington.
- Mather, James, Y. and Hans-Henning Speitel (Eds). *The Linguistic Atlas of Scotland*. London: Croom Helm. Volume 1: 1975, Volume 2: 1977, Volume 3: 1986.
- Mesthrie, Rajend, Alida Chevalier and Timothy Dunne. 2013. A study of variation in the BATH vowel among White Speakers of South African English in five cities. *Pennsylvania Working Papers in Linguistics*, 19/2: 130–140.
- Mesthrie, Rajend, Alida Chevalier and Timothy Dunne. 2015. A regional and social dialectology of the BATH vowel in South African English. *Language Variation and Change*, 27(1): 1–30.
- Mitchell, Alexander G. and Arthur Delbridge. 1965. *The Speech of Australian Adolescents*. Sydney: Angus and Robertson.
- Newbrook, Mark. 1986. *Sociolinguistic Reflexes of Dialect Interference in West Wirral*. Bern & Frankfurt am Main: Peter Lang.
- Nielsen, Daniel and Jennifer Hay. 2005. Perceptions of regional dialects in New Zealand English. *Te Reo*, 48: 95–110.
- Orton, Harold. 1962. *Survey of English Dialects: Introduction*. Leeds: E J Arnold and Son.
- Orton, Harold, Stewart Sanderson, and John Widdowson. 1978. *The Linguistic Atlas of England*. London: Croom Helm.
- Orton, Harold and Nathalia Wright. 1974. *A Word Geography of England*. New York: Seminar Press.
- Parry, David. 1979. The survey of Anglo-Welsh dialects. *Lore and Language*, 3(1): 9–14.
- Penhallurick, R. K. (1991) *The Anglo-Welsh Dialects of North Wales: a Survey of Conservative Rural Spoken English in the Counties of Gwynedd and Clwyd*. Frankfurt am Main: Peter Lang.
- Petyt, K. M. 1980. *The Study of Dialect: an Introduction to Dialectology*. London: Andre Deutsch.
- Petyt, K. M. 1985. 'Dialect' and 'Accent' in Industrial West Yorkshire. Amsterdam: John Benjamins.
- Piercy, Caroline. 2007. A quantitative analysis of rhoticity in Dorset: evidence from four

- locations of an urban to rural hierarchy of change. *Proceedings of the Fifth Postgraduate Conference in Linguistics, CamLing 2007*, 199–206.
- Romain, Suzanne. 1975. *Linguistic variability in the speech of some Edinburgh schoolchildren*. Unpublished MLitt thesis, University of Edinburgh, UK.
- Schützler, Ole. 2014. Vowel variation in Scottish Standard English. Accent-internal differentiation or Anglicisation? In Lawson (Ed), 129–152.
- Stuart-Smith, Jane. 2008. Scottish English: phonology. In Kortmann & Upton (Eds), 48–70.
- Trudgill, Peter. 1974. *The Social Differentiation of English in Norwich*. Cambridge: Cambridge University Press.
- Trudgill, Peter. 1999. *The Dialects of England*. 2nd edition. Oxford: Blackwell.
- Trudgill, Peter. 2004. *New Dialect Formation: the Inevitability of Colonial Englishes*. Edinburgh: Edinburgh University Press.
- Turner, George, W. 1994. English in Australia. In Burchfield (Ed), 277–327.
- Upton, Clive. 2008. Received pronunciation. In Kortmann & Upton (Eds), 237–252.
- Upton, Clive & Davies, Bethan, L. 2013. *Analysing 21st Century British English*. London: Routledge.
- Wakelin, Martyn. 1972. *English Dialects*. London: Athlone Press, University of London.
- Watson, Kevin. 2007. Liverpool English. *Journal of the International Phonetic Association*, 37/3: 351–360.
- Watson, Kevin and Lynn Clark. (2013). How salient is the nurse~square merger? *English Language and Linguistics*, 17(2), 297–323.
- Watt, Dominic and Ulrike Altendorf. 2008. The dialects in the South of England: phonology. Kortmann & Upton (Eds), 194–222.
- Watt, Dominic and Lesley Milroy. 1999. Patterns of variation and change in three Newcastle vowels: is this dialect levelling? In Foulkes and Docherty (Eds), 25–46.
- Watt, Dominic, Carmen Llamas, and Daniel Johnson. 2014. Sociolinguistic variation on the Scottish-English border. In Lawson (Ed), 79–102.
- Wells, John. C. 1982. *Accents of English*. 3 volumes. Cambridge: Cambridge University Press.
- Wright, Joseph. 1898–1905. *English Dialect Dictionary*. London: Frowe.
- Wright, Joseph. 1905. *The English Dialect Grammar*. London: Frowe.

26 Dialects of North American English

CHARLES BOBERG

26.1 Introduction

North American English is now almost 400 years old. It is the majority language of both the United States and Canada, though in the United States it has no national official status; in Canada, it shares official status as a national language with French. English-speaking settlement of North America began in the early seventeenth century, with the establishment of British colonies in present-day Virginia and Massachusetts, on the Atlantic coast. By the late seventeenth century, additional colonies had been founded at New Hampshire, Rhode Island, Connecticut, New York, New Jersey, Pennsylvania, Delaware, Maryland, and South Carolina, with the British taking over earlier Dutch settlements in several places. By the late eighteenth century, a continuous region of English-speaking settlement was consolidated along the eastern seaboard and began to expand inland, while British victory over France in the Seven Years' War opened Canada to English-speaking settlement as well, following initial footholds in Newfoundland and Nova Scotia. Shortly thereafter, the American War of Independence severed political links between Britain and its American colonies, while British dominion endured in Canada. As part of the growing cultural autonomy of the new United States, an independent linguistic standard for American English was pioneered by Noah Webster, with the publication of his *American Dictionary of the English Language* in 1828; this tradition is represented today by *Webster's Third New International Dictionary* (1961). During the nineteenth century, English-speaking settlement of the continent was completed in a westward expansion that eventually reached the Pacific coast and the more remote regions of the central and western interior, aided by the development of a transcontinental rail system. In 1867, Canada also attained independence and it too now has its own standard of English, represented by the *Canadian Oxford Dictionary* (Barber 1998), successor to the earlier *Gage Canadian Dictionary* (Avis *et al.* 1983), though the Canadian standard combines many features of the British and American standards, which also enjoy wide acceptance in Canada. By the beginning of the twenty-first century, English was spoken by almost 300,000 million people in the United States and almost 30 million in Canada (including second-language speakers; figures for native speakers are slightly lower).

In general, dialect variation in modern North American English is less apparent than in Britain or many non-English-speaking countries and the standard varieties of North American English are widely and increasingly used across the continent. Recognizable regional dialects are found mostly in earlier-settled regions along the Atlantic coast, including the American South, the metropolitan regions around Philadelphia and New York City, New

England, and Atlantic Canada. By contrast, most of the western half of the continent, including the United States west of the Mississippi River and Canada from British Columbia to Ontario, exhibits a remarkably uniform, homogeneous variety of English that is often difficult to distinguish from the standard varieties, especially among the European-origin majority of the population, even at lower social levels. This chapter will review the most important dialectological research on North American English and the main dialect divisions it has established. Given the enormous quantity of this research, the discussion will necessarily be highly selective.

26.2 The Study of North American English Dialects

Serious study of dialect variation in North American English began in the 1930s, when Hans Kurath and his colleagues planned a *Linguistic Atlas of the United States and Canada*, modeled on the French and Italo-Swiss atlases. Fieldworkers conducted and transcribed interviews with representatives of local speech in each community and the resulting data were analyzed and published in map form. The American atlas tradition is well reviewed in Atwood (1971) and Petyt (1980) and by Kretzschmar (this volume). An expensive and time-consuming method, it never managed to extend its fieldwork, much less its published results, to the entire continent, but did produce an invaluable set of publications on the eastern half of the United States. The most complete publications cover New England (Kurath *et al.* 1939–1943), the Gulf States of the South (Pederson, McDaniel, and Adams 1986–1993; Atwood 1962) and the Upper Midwest (Allen 1973–1976). Partial and summary data are available for several other regions: most notably for vocabulary (Kurath 1949), pronunciation (Kurath and McDavid 1961) and verb forms (Atwood 1953) along the entire eastern seaboard, but also for the North Central States (Marckwardt 1957; Shuy 1962) and several western regions (e.g., Colorado in Hankey (1960) and California in Bright (1971); see also Reed (1961) for the Pacific Northwest).

A second effort to map dialect variation across the United States, begun in the 1960s, did achieve national coverage, but its data were restricted to lexical variation and published mainly in dictionary rather than atlas form. The *Dictionary of American Regional English*, or *DARE* (Cassidy and Hall 1985–2012) makes only secondary use of maps; a more cartographic approach, together with an analysis of regional patterns, is offered by Carver (1987). An updated version of the *DARE* is now accessible on-line (Cassidy, Hall, and Von Schneidemesser 2013).

During the 1960s, sociolinguistics became the dominant framework of research on variation in North American English. While the main focus of sociolinguistics was social rather than regional variation, it was founded on studies of two traditional dialect regions by William Labov: Martha's Vineyard, an island off the Massachusetts coast (Labov 1963, 1972a); and New York City (Labov 1966, 1972a). Many subsequent sociolinguistic studies added to our knowledge of American dialect variation, from Callary (1975) in Illinois to Wolfram and Schilling-Estes (1997) on Ocracoke Island to Bailey *et al.* (1993) in Oklahoma. There have also been many studies of ethnic “dialects” in American English, especially those associated with the African American and Latino populations (Wolfram 1969, 1974; Labov 1972b; Fought 2003), much less commonly with Native American or First Nations indigenous groups (Leap 1993; Ball and Bernhardt 2008); in these studies, however, the association with place or region, the main concern of dialectology, is generally less direct than in studies of the European-American majority, which will be the focus of this chapter.

The melding of sociolinguistics and dialectology to create the hybrid field of sociodialectology, simultaneously examining social and regional variation as interrelated phenomena, produced a third effort to create a nationwide map of American English, beginning in the

1990s. As its title suggests, the resulting *Atlas of North American English*, or ANAE (Labov, Ash, and Boberg 2006), is truly continental in scope, for the first time granting Canada more than marginal status. It also introduced methodological innovations: the telephone and a tape recorder were used to gather data quickly and at low cost, while computerized acoustic analysis of the resulting recordings allowed for an unprecedented view of regional differences in phonemic inventory, in the phonetic qualities of vowels and in the progress of phonological and phonetic changes, such as mergers and chain shifts. The ANAE's analyses are largely restricted to phonetic and phonological variables, thereby complementing the lexical data of *DARE*. Within each region, statistical analysis of correlations between social categories and phonological variables supports conclusions about the social mechanisms of ongoing change.

Although Canadian English had been largely ignored by American dialectology prior to the ANAE, it was the subject of a separate tradition of dialect study in Canada that began in the 1950s. A major concern of early work was the alternation of American, British, and distinctively Canadian words, pronunciations, and grammatical usage in Canadian English, reflecting the country's settlement first by Loyalist refugees from the American Revolution, then by direct immigration from Britain, including Ireland, during the first half of the nineteenth century. Such alternations could be investigated with written questionnaires completed by respondents themselves, thereby increasing the speed and efficiency of data collection, but introducing limitations on the type and quality of data gathered (Boberg 2013; Chambers, this volume). This tradition began with Avis' study of speech differences along the Ontario-US border (1954–1956) and reached its summit in the *Survey of Canadian English*, Scargill and Warkentyne's massive survey of over 14,000 Canadians across the whole country, with data tabulated by province, generation, and sex (1972; see also Warkentyne 1973). It also produced a *Dictionary of Canadianisms on Historical Principles* (Avis et al. 1967), now updated and accessible on-line (Dollinger, Brinton, and Fee 2013). More recent contributions to this tradition are the *Dialect Topography* project (Chambers 1994; Chambers and Pi 2004) and the *North American Regional Vocabulary Survey* (Boberg 2005; 2010). By the 1980s, the sociolinguistic tradition spread to Canada, producing comprehensive studies of social variation in the speech of Vancouver (Gregg 1992) and Ottawa (Woods 1999). The phonetics of Canadian English were first examined by Gregg (1957); Boberg (2008, 2010) presents more recent acoustic data on regional phonetic variation within Canada.

The Canadian province of Newfoundland (and Labrador), which joined Canada in 1949 after over 300 years as a separate British colony, is home to highly distinctive varieties of English reflecting settlement from southwestern England and southeastern Ireland. These have been studied in a separate dialectological tradition, of which the cornerstone is a *Dictionary of Newfoundland English* (Story, Kirwin, and Widdowson 1982); a recent overview of Newfoundland English is provided by Clarke (2010).

26.3 Major North American English Dialects and Their Principal Features

The dialectological studies described above offer several taxonomies of North American English dialects. Perhaps the best known is that of Kurath (1949), which does not extend beyond the eastern seaboard of the United States, but divides that territory into three main dialect regions. Kurath saw these as having developed from three original centers of settlement: from Boston, a Northern region expanded to comprise New England, New York, and northern New Jersey; from Philadelphia, a Midland region expanded over southern New Jersey and most of Pennsylvania, West Virginia, and Ohio; and from early settlements at Richmond and Charleston, a Southern region expanded inland to include southern

Maryland and the eastern halves of Virginia and the Carolinas. Kurath's ternary view was challenged by Carver (1987), whose *DARE* data suggested a first-order binary division into North and South, approximately along the Ohio River, with Kurath's Midland region split into a Lower North subregion, north of the river, and an Upper South subregion south of it. The view that will be adopted here is based on that of the ANAE (Labov, Ash, and Boberg 2006), which presents a compromise, supported by acoustic phonetic data on vowel production: Kurath's division between North and Midland is retained, along with Carver's division between North and South, the latter including Kurath's Lower Midland. This division is shown, along with the other ANAE dialect regions, in Figure 26.1, reproduced from ANAE Map 11.15 (148).

The view presented in Figure 26.1 depends entirely on phonological and phonetic data. Despite the efforts of Kurath (1949) and Carver (1987) to demonstrate that clearly delimited dialect regions arise from sets of lexical variants with similar geographic distributions—a view criticized by Kretzschmar (1996)—the essentially asystematic nature of lexical variation is a difficult challenge for this approach, since regional boundaries for each variable tend to occur in different places. The international boundary between the United States and Canada, not surprisingly, is to some extent an exception to this, being aligned with a substantial set of fairly categorical lexical and phono-lexical differences (Avis 1954–1956; Allen 1959; Boberg 2005, 2010). Nevertheless, despite much valuable work on regional vocabulary (e.g., Johnson 1996; Boberg 2005), the low frequency of occurrence of most lexical variables, even when they are not obscure or obsolete, lessens their diagnostic utility, notwithstanding their intrinsic interest to dialectologists and students of settlement history and regional culture. Many members of the general public recognize that certain words vary from one region to another, but this sort of knowledge cannot be used either in projecting one's own regional identity or in assessing



Figure 26.1 Major dialects of North American English. Reproduced (with permission) from Labov, Ash, and Boberg (2006: 148).

that of an interlocutor if these particular words do not happen to arise in conversation. The ordinary person's ability to identify regional origins on the basis of speech cannot therefore depend mainly on lexical variables, regardless of their symbolic importance. As for grammatical variation, while regional syntactic and morphological patterns have been identified (see, e.g., Atwood 1953), such data have not so far generated an overall view of North American dialects.

It therefore seems safe to say that the most systematic and informative view of regional differences in North American English emerges from the analysis of phonological and phonetic data of the type presented in the ANAE, or in Thomas (2001). The ANAE is almost exclusively concerned with variation in vowel production: whereas many sociolinguistic studies have shown that consonantal variables like cluster simplification (especially /t,d/-deletion) display social variation *within* communities (Labov 1972a: 216–226), regional variation *among* communities is far more likely to involve vocalic variables. The ANAE's analysis reflects the view, derived from the structural dialectology of Weinreich (1954) and Moulton (1960, 1962), that the purely phonetic differences in vowel quality that the general public hears as regional "accents" are rooted in deeper, phonological differences of phonemic inventory and the underlying "structure" of regional vowel "systems." Changes in these systems are constrained by the structural principles governing vowel systems developed by Martinet (1955), particularly the concepts of fields of dispersion, margins of security, and maximum differentiation, which Labov (1991) develops into eight principles governing chain shifting and phonemic merger (see Gordon, this volume).

In order to discuss the regional differences identified by the ANAE, it will be necessary to use a set of symbols for transcribing the vowel phonemes of North American English. These are given in Table 26.1, together with the equivalent keywords from Wells (1982). Labov's system, which originates with Trager and Bloch (1941), divides the English vowels into four subsets: a primary division between short or lax vowels, with a single character indicating nuclear quality, and long or tense vowels, with a second symbol indicating the presence of a post-nuclear glide; and a secondary division of long vowels based on the direction of the glide, with /-y/ indicating front up-glides, /-w/ back up-glides, and /-h/ in-glides; pre-rhotic vowels require separate treatment (see Labov 1991: 7; ANAE: 11–15).

Regional patterns of chain shifting in American English were first explored by Labov, Yaeger, and Steiner (1972) but articulated more concisely by Labov (1991). Labov argues that a basic taxonomy of North American English dialects can be developed by considering two "pivot points," which entail the phonemic organization of the low-front and low-back corners of the vowel space:

1. whether the modern development of Middle English short /a/ is a split system, as in Standard British English, with short /æ/ (TRAP) distinct from long /æh/ (BATH), or a merged system, with a single quality for both sets; and
2. whether the low-back vowels, /o/ (LOT) and /oh/ (THOUGHT), are distinct or merged.

In the Mid-Atlantic region, including New York City and Philadelphia, both contrasts are maintained, with the long or tense members, /æh/ and /oh/, having risen to upper-mid position (BATH is [beəθ] and THOUGHT is [θoət]), whereas the short or lax members remain in low position (TRAP is [tɹæp] and LOT is [lɑt]). In the Inland North, the region of large industrial cities along the Great Lakes from Chicago through Detroit to western New York State, the low-back contrast is maintained instead by a forward shift of /o/, so that LOT is [lat] or even [læt] and THOUGHT is [θɔ:t]; this is made possible by the raising of /æ-æh/ as a single tense phoneme to upper-mid-front position, where TRAP and BATH are both [eə]. This configuration is the starting point for the "Northern Cities" chain shift, which includes unrounding of /oh/ to [o], backing of /ʌ/ (STRUT) toward [ɔ] and backing of /e/ (DRESS) toward [ʌ]. In the South, monophthongization of /ay/ (PRICE) to [a:] sets off a "Southern Shift," in which /ey/

Table 26.1 Broad transcription of English vowel phonemes, with keywords from Wells (1982). Keywords for /iŋ/ do not appear in Wells, who includes /iŋ/ in the GOOSE set.

Short/lax vowels (V)	Long/tense vowels			
	Front up-gliding (Vy)	Back up-gliding (Vw)	Monophthongal/ in-gliding (Vh)	Pre-rhotic (-r)
/i/ KIT	/iy/ FLEECE	/iŋ/ few, cue	/æh/ BATH	/iyr/ NEAR
/e/ DRESS	/ey/ FACE	/uŋ/ GOOSE	/ah/ PALM,	/eyr/ SQUARE
/æ/ TRAP	/ay/ PRICE	/ow/ GOAT	/oh/ THOUGHT, CLOTH	/ahr/ START
/o/ LOT	/oy/ CHOICE	/aw/ MOUTH		/owr/ FORCE
/ʌ/ STRUT				/ohr/ NORTH
/u/ FOOT				/uwr/ CURE
				/ɜ̄/ NURSE

(FACE) lowers to [ɛɪ] and /iy/ (FLEECE) lowers to [əɪ], while the nuclei of the long back vowels, /uŋ, ow, aw/ (GOOSE, GOAT, MOUTH), shift forward. The forward shift of /aw/ to [æʊ] initiates a “Back Upglide Shift” (ANAE: 254–256), in which /aw/ is reanalyzed as /æw/, allowing the low-back contrast to be maintained by the diphthongization of /oh/, which thereby joins the subsystem of back-upgliding vowels as a new /aw/: LOT is [lɒt] and THOUGHT, with a similar nuclear quality, is [θaʊt]. In most of the remainder of North America, including parts of New England, much of the Midland, the West, and Canada, we find what Labov (1991) labels the “Third Dialect,” in which both pivot points are single phonemes: TRAP and BATH are both [æ] or [a], with raising of /æ/ restricted to pre-nasal contexts; and LOT and THOUGHT are both [ɑ] or [ɒ], with the degree of rounding and advancement varying by region. Eastern New England is distinguished from the remainder of the Third Dialect regions by /r/-vocalization, which it shares, at least traditionally, with New York City and coastal parts of the South. This basic taxonomy—North, South, and Third Dialect—with distinctive local patterns centered on the major Mid-Atlantic cities, underlies the full set of major dialect regions shown in Figure 26.1. With their major features, these are as follows.

26.3.1 New England: Boston, Providence, Hartford, and Burlington

New England, in fact, contains four dialect regions, grouped together here to conserve space: a northeastern region, based around and north of Boston; a southeastern region around Providence, Rhode Island; a southwestern region around Hartford, Connecticut and Springfield, Massachusetts; and a northwestern region in Vermont, where the largest city is Burlington (Boberg 2001; Stanford, Leddy-Cecere and Baclawski 2012). These regions are formed by two important isoglosses: vocalization of /r/ (START as [sta:t], etc.) is still common in the East (Boston and Providence), but now rare in the West (Hartford and Burlington); and the low-back merger (LOT=THOUGHT) is found in the North (Boston and Burlington) but not in the South (Providence and Hartford). Where the low-back merger has occurred, it does not include /ah/ (PALM), which remains further front than /o/ (LOT), together with its allophone before /r/ (START). In traditional Boston speech, now recessive, some members of the BATH class (like *laugh* and *class*) remain distinct from /æ/, but rather than fronting and raising, as in the Mid-Atlantic region, they join the PALM class ([la:f, kla:s]), as in Southern British English, and FORCE and NORTH remain distinct ([foəs, nɒθ]). One feature that

unites all of New England is the conservative status of the back-upgliding vowels, /uw, ow, aw/, which display little or no shift toward the center of the vowel space.

26.3.2 The Mid-Atlantic: New York City, Philadelphia, and Baltimore

The Mid-Atlantic region also contains internal divisions, most importantly between its northern section, around New York City, which traditionally shared /r/-vocalization with New England, and its southern section, including Philadelphia and Baltimore, which did not. Labov (1966, 1972a) famously found that this difference was receding, as younger New Yorkers restored /r/ in their careful speech, a trend recently confirmed by Becker (2014). Another internal difference involves elements of the Southern Shift, which extend to Philadelphia, producing strongly fronted back-upgliding vowels and lowered /iy/ and /ey/ in non-pre-voiceless contexts, but not to New York. The Mid-Atlantic region is united by the pivot points discussed above: LOT and THOUGHT are distinct, with PALM merged with LOT; and TRAP and BATH are distinct, though New York and Philadelphia differ in class membership, with a wider range of environments included in the tense class in New York (for instance, *cab*, *bag*, *badge*, and *cash* are tense in New York but lax in Philadelphia; Labov 1972a: 73–75; ANAE: 173; Labov 2007: 354). A recent study by Labov, Rosenfelder and Fruehwald (2013) finds that many traditional features of Philadelphia phonology are now disappearing.

26.3.3 The South: Richmond, Charlotte, Atlanta, Nashville, Dallas, and Houston

Although monophthongization of /ay/ (PRICE) occurs before sonorants, as in *time*, *tile*, or *tire*, in some regions bordering the South, the ANAE (244–248) defines the South itself as the territory in which /ay/ (PRICE) is more generally monophthongal, including in final position and before voiced obstruents (*tie* and *tide* pronounced [ta:, ta:d]); monophthongization before voiceless obstruents, as in *tight*, is also restricted to the South but variable within it. As discussed above, monophthongization of /ay/ is the initiating condition for the Southern Vowel Shift (Labov 1991: 22–28; ANAE: 242–254), a set of changes in vowel quality that the general public identifies as a “Southern drawl”: once /ay/ exits the subsystem of front-up-gliding vowels (becoming a new /ah/), its former position in that subsystem is taken over by the lowering of /ey/ (so that Southern *day* sounds like Northern *die*); /iy/, in turn, falls toward the former position of /ey/ (*bee* sounds like *bay*). The short front vowels swap positions with their long counterparts, becoming long, or tense, and relatively more peripheral, with in-glides, so that *bid*, *bed* and *bad* become [biəd, beəd, bæjəd]. A separate set of shifts affects the back vowels, whose nuclei approach the center of the vowel space: this includes not only the long vowels, GOOSE, GOAT, and MOUTH, but also the short vowels, STRUT and FOOT, which, like the front short vowels, are lengthened. Parallel to the development of /ay/, /oy/ is partially monophthongized to become a new /oh/ (CHOICE as [tʃoəs]). The Southern Shift is found in its most advanced form in two interior regions of the South: a western region in Texas, from Lubbock to Dallas; and an eastern region centered on Chattanooga and Knoxville, Tennessee (ANAE 257). It is not traditionally found in the older coastal cities of the South, like Charleston or New Orleans, or in marginally Southern regions under northern influence, like Washington, DC, or southern Florida. Today it tends to be recessive, even in its core areas: whereas the fronting of long back vowels is well entrenched and has even spread to many regions outside the South, the front component of the Southern Shift, including the monophthongization of /ay/, is reversing in urban centers like Dallas, Houston, and Atlanta, particularly among younger, middle class people (Thomas 1997; Fridland 2001; Dodsworth and Kohn 2012). A more durable trait of Southern phonology is a merger of /i/ and /e/ before nasals, so that *pin* and *pen* both sound like *pin*.

26.3.4 The Inland North: Chicago, Detroit, Cleveland, and Buffalo

As discussed above, the Great Lakes region, from Milwaukee, Wisconsin, through Chicago, Detroit, Toledo, Cleveland, and Buffalo to Rochester, New York, is the domain of the Northern Cities Vowel Shift (Labov 1991: 14–20; ANAE: 187–215). Its most striking aspects are the mid-front position of TRAP and BATH, which are not distinct, and the advanced position of /o/ (LOT), merged with /ah/ (PALM), which has taken their place in the low-front quadrant of the vowel space, thereby preserving a stable phonemic contrast with /oh/ (THOUGHT). As a result, Inland Northern *hat* sounds to people from outside the region rather like *hay-at*, while *hot* sounds like *hat*. The mid vowels, DRESS and STRUT, are by contrast retracted (*deck* sounds like *duck* and *duck* like *dock*), while the back vowels, GOOSE and GOAT, have a conservative, unshifted, sometimes almost monophthongal quality much like that heard in New England, the most important source for the initial English-speaking settlement of this region. MOUTH also resists fronting, its nucleus remaining well behind that of PRICE, which is low-front. In all of these respects, the Inland North is starkly distinguished from the Midland region to its south, though McCarthy (2011) suggests that the initiating stages of the Northern Cities Shift—raising of /æ/ and fronting of /o-ah/—are no longer active changes in Chicago, while Driscoll and Lape (2015) observe a complete reversal of the shift in Syracuse, New York.

26.3.5 The Midland: Pittsburgh, Columbus, Cincinnati, Indianapolis, and St. Louis

It was pointed out above that the very existence of a distinct Midland region has been somewhat contentious in American dialectology. The North and South have a clear taxonomic status, but whether the Midland is of equal status or merely a transition zone, characterized by a gradual northward weakening of Southern features, is a matter of debate (Marckwardt 1957; Frazer 1978; Habick 1993; Thomas 2010). It certainly shares with the South a general fronting of back vowels, though monophthongization of /ay/, as already mentioned, is limited to pre-sonorant contexts. The Midland also shares many features with the West: whereas its border with the Inland North is well defined by the southern limit of the Northern Cities Shift (ANAE: 134, 207), its western edge, which the ANAE places in Nebraska and Kansas (134), is virtually imperceptible. Indeed, its future seems to lie in convergence with the West: as in the South, local patterns are eroding in favor of a Third Dialect phonology that is basically indistinguishable from Western speech. The low-back merger is already complete in a large territory around Pittsburgh and is now expanding among younger generations in other cities (ANAE 60–61), while the last vestiges of old split short /a/ systems are giving way to the western “nasal” system, with a single phoneme raised only but always before front nasals (in *band*, *family*, *hammer*, etc.). Nevertheless, the Midland traditionally contained a number of distinct local dialects based in its major cities: most notably Pittsburgh, with monophthongization of /aw/ (Johnstone and Kiesling 2008); Cincinnati, with a TRAP-BATH split related to that in the Mid-Atlantic region (Boberg and Strassel 2000; Labov 2007: 361–364); and St. Louis, with many northern features (Labov 2007: 372–382) but also with a unique set of contrasts before /r/, in which FORCE and NORTH are distinct but NORTH and START are merged (Murray 1986; ANAE: 49–53).

26.3.6 The West: Denver, Phoenix, Seattle, San Francisco, and Los Angeles

As already noted, the American West is really an extension of the Midland, with the distinct local patterns of that region lost in a homogeneous compromise dialect, which also shows influences of settlement from the North and South. What emerged from this process of

mixing and leveling is the basis of modern “General” or “Standard” American English, as heard from regionally unmarked characters in most of the films and television programs produced in Los Angeles today. The low-back merger is complete throughout the region and fronting of back vowels, particularly /ow/ (GOAT), while present, is less extreme than in the Midland or South. Short /a/ is a single phoneme with raising restricted to pre-nasal contexts and, in Seattle and the Pacific Northwest, before voiced velars (*bag*, *hang*, etc.), a feature also found in adjacent regions of the Upper Midwest and Canada (ANAE 182; Boberg 2008; Purnell 2009; Benson, Fox and Balkman 2011; Wassink 2016). Along the West’s eastern edge, several cities have a transitional status. The largest of these is Minneapolis-St. Paul, which combines the low-back merger of the West with the general raising of /æ-æh/ found in the Inland North, as well as northern resistance to the shifting of long up-gliding vowels, so that FACE and GOAT have an almost monophthongal quality, [e:] and [o:]. Much of the Great Plains region, including cities like Des Moines, Omaha, Kansas City, and Tulsa, displays a mixture of Midland and Western features, to the extent that these can be distinguished.

26.3.7 Canada: Vancouver, Edmonton, Calgary, Toronto, Ottawa, and Montreal

The relatively subtle regional transitions and widespread homogeneity of the American West extend across the international border into western and central Canada. In Ontario, the low-back merger and an unsplit, unraised /æ/ create a sharp contrast with the Inland North (Toronto versus Detroit or Buffalo; Boberg 2000), but this difference fades out further west (Calgary versus Denver; Vancouver versus Seattle). Canadian English, a thoroughly North American variety in most respects with only superficial vestiges of British influence, is noted for three features that distinguish it from other North American varieties. The best known is Canadian Raising, the shortening and raising of the nuclei of /aw/ and /ay/ before voiceless obstruents, so that *south* and *ice* are approximately [s3ʊθ] and [3ɪs] (Joos 1942; Gregg 1957; Chambers 1973; ANAE 221; Boberg 2008: 138–141, 2010: 149–151, 204–205), though raising of /ay/ is also heard in parts of the eastern United States. In the ANAE, the distinguishing feature of Canada is therefore the Canadian Shift, a lowering and retraction of TRAP and DRESS, apparently in response to the low-back merger, which allows /æ/ to shift into low-central position (Clarke, Elms, and Youssef 1995; ANAE 216–224; Boberg 2008: 135–138, 2010: 230). This change may now be underway in other Third Dialect regions as well, including California (Kennedy and Grama 2012). Finally, Canadian English displays a unique pattern of foreign (a) nativization, involving variable assignment of the /a/ vowel of foreign words to native English phonemes: where American English uses mostly /ah/ in words like *drama*, *façade*, *lava*, *macho*, *mantra*, *pasta*, and *saga*, and British English varies between /ah/ and /æ/, Canadians traditionally use mostly /æ/ (Boberg (2010: 137–140). Eastern Canada, especially Newfoundland, presents greater dialect diversity, but the speech of mainland urban centers like Halifax shares much with Ontario. In Montreal, English is a minority language in contact with French and exhibits many unique contact-related features, as well as considerable ethnic diversity (Boberg 2010: 182–184, 214–224), but the English of British-origin Montrealers is generally very similar to that of other Canadians.

REFERENCES

-
- | | |
|--|---|
| Allen, Harold Byron. 1959. Canadian-American speech differences along the middle border. <i>Journal of the Canadian Linguistic Association</i> 5/1: 17–24. | Allen, Harold Byron. 1973–1976. <i>The Linguistic Atlas of the Upper Midwest in Three Volumes</i> . Minneapolis: University of Minnesota Press. |
|--|---|

- Atwood, E. Bagby. 1953. *A Survey of Verb Forms in the Eastern United States*. Ann Arbor: University of Michigan Press.
- Atwood, E. Bagby. 1962. *The Regional Vocabulary of Texas*. Austin: University of Texas Press.
- Atwood, E. Bagby. 1971. The methods of American dialectology. In Harold Byron Allen and Gary N. Underwood (eds.), *Readings in American Dialectology* (New York: Appleton-Century-Crofts), 5–35.
- Avis, Walter S. 1954–1956. Speech differences along the Ontario–United States border. *Journal of the Canadian Linguistic Association* 1/1: 13–18 (Vocabulary); 1/1 (Regular Series) 14–19 (Grammar); and 2/2: 41–59 (Pronunciation).
- Avis, Walter S., et al. (eds.). 1983. *Gage Canadian Dictionary*. Toronto: Gage Educational Publishing.
- Avis, Walter S., et al. (eds.). 1967. *A Dictionary of Canadianisms on Historical Principles*. Toronto: W. J. Gage.
- Bailey, Guy, Tom Wikle, Jan Tillary and Lori Sand. 1993. Some patterns of linguistic diffusion. *Language Variation and Change* 5: 359–390.
- Ball, Jessica, and B. May Bernhardt. 2008. First Nations English dialects in Canada: Implications for speech-language pathology. *Clinical Linguistics & Phonetics* 22/8: 570–588.
- Barber, Katherine (ed). 1998. *The Canadian Oxford Dictionary*. Toronto: Oxford University Press.
- Becker, Kara. 2014. (r) we there yet? The change to rhoticity in New York City English. *Language Variation and Change* 26/2: 141–168.
- Benson, Erica J., Michael J. Fox, and Jared Balkman. 2011. The bag that Scott bought: The low vowels in northwest Wisconsin. *American Speech* 86/3: 271–311.
- Boberg, Charles. 2000. Geolinguistic diffusion and the U.S.–Canada border. *Language Variation and Change* 12: 1–24.
- Boberg, Charles. 2001. The phonological status of Western New England. *American Speech* 76: 3–29.
- Boberg, Charles. 2005. The North American Regional Vocabulary Survey: New variables and methods in the study of North American English. *American Speech* 80/1: 22–60.
- Boberg, Charles. 2008. Regional phonetic differentiation in Standard Canadian English. *Journal of English Linguistics* 36/2: 129–154.
- Boberg, Charles. 2010. *The English Language in Canada: Status, History and Comparative Analysis*. Cambridge, UK: Cambridge University Press.
- Boberg, Charles. 2013. Surveys: The use of written questionnaires in sociolinguistics. In Christine Mallinson, Becky Childs and Gerard Van Herk (eds.), *Data Collection in Sociolinguistics: Methods and Applications* (London: Routledge), 131–141.
- Boberg, Charles, and Stephanie Strassel. 2000. Short-a in Cincinnati: A change in progress. *Journal of English Linguistics* 28: 108–126.
- Bright, Elizabeth. 1971. *A Word Geography of California and Nevada*. *University of California Publications in Linguistics* 69. Berkeley and Los Angeles: University of California Press.
- Callary, Robert E. 1975. Phonological change and the development of an urban dialect in Illinois. *Language in Society* 4: 155–169.
- Carver, Craig. 1987. *American Regional Dialects*. Ann Arbor: University of Michigan Press.
- Cassidy, Frederic G., and Joan Houston Hall (eds.). 1985–2012. *Dictionary of American Regional English*. 5 vols. Cambridge, MA: Harvard University Press.
- Cassidy, Frederic G., Joan Houston Hall and Luanne Von Schneidemesser (eds.). 2013. *Dictionary of American Regional English*. Cambridge, MA: Belknap Press of Harvard University Press. <<http://daredictionary.com>> (accessed January 2014).
- Chambers, J.K. 1973. Canadian Raising. *Canadian Journal of Linguistics* 18/2: 113–135.
- Chambers, J.K. 1994. An introduction to dialect topography. *English World-Wide* 15/1: 35–53.
- Chambers, J.K., and Chia-Yi Tony Pi. 2004. *Atlas of Dialect Topography (On-Line)*. <<http://dialecttopography.chass.utoronto.ca>> (accessed January 2014).
- Clarke, Sandra. 2010. *Newfoundland and Labrador English*. Edinburgh: Edinburgh University Press.
- Clarke, Sandra, Ford Elms and Amani Youssef. 1995. The third dialect of English: Some Canadian evidence. *Language Variation and Change* 7: 209–228.
- Dodsworth, Robin, and Mary Kohn. 2012. Urban rejection of the vernacular: The SVS undone. *Language and Variation and Change* 24: 221–245.
- Dollinger, Stefan (ed.-in-chief), Laurel J. Brinton and Margery Fee (eds.). 2013. *DCHP-1 Online: A Dictionary of Canadianisms on Historical Principles Online*. Vancouver, BC: University of British Columbia. Based on Walter S. Avis et al. (1967). <<http://dchp.ca/DCHP-1/>> (accessed January 2014).
- Driscoll, Anna, and Emma Lape. 2015. Reversal of the Northern Cities Shift in Syracuse, New York. *University of Pennsylvania Working Papers in Linguistics* 21/2: 39–47 (Article 6). URL: <http://repository.upenn.edu/pwpl/vol21/iss2/6>. Accessed March 02, 2017.

- Fought, Carmen. 2003. *Chicano English in Context*. New York: Palgrave MacMillan.
- Fridland, Valerie. 2001. The social dimension of the Southern Vowel Shift: Gender, age and class. *Journal of Sociolinguistics* 5: 233–253.
- Frazer, Timothy. 1978. South Midland pronunciation in the North Central states. *American Speech* 53: 40–48.
- Gregg, Robert J. 1957. Notes on the pronunciation of Canadian English as spoken in Vancouver, B.C. *Journal of the Canadian Linguistic Association* 3/1: 20–26.
- Gregg, Robert J. 1992. The Survey of Vancouver English. *American Speech* 67/3: 250–267.
- Habick, Timothy. 1993. Farmer City, Illinois: Sound systems shifting south. In Timothy C. Frazer (ed.), "Heartland" English (Tuscaloosa, AL: University of Alabama Press), 97–124.
- Hankey, Clyde T. 1960. *A Colorado Word Geography*. *Publications of the American Dialect Society* 34. Tuscaloosa: University of Alabama Press.
- Johnson, Ellen. 1996. *Lexical change and variation in the southeastern United States, 1930–1990*. Tuscaloosa: University of Alabama Press.
- Johnstone, Barbara, and Scott F. Kiesling. 2008. Indexicality and experience: exploring the meanings of /aw/-monophthongization in Pittsburgh. *Journal of Sociolinguistics* 12: 5–33.
- Joos, Martin. 1942. A phonological dilemma in Canadian English. *Language* 18: 141–144.
- Kennedy, Robert, and James Grama. 2012. Chain shifting and centralization in California vowels: An acoustic analysis. *American Speech* 87: 39–56.
- Kretzschmar, William A., Jr. 1996. Quantitative areal analysis of dialect features. *Language Variation and Change* 8: 13–39.
- Kurath, Hans. 1949. *A Word Geography of the Eastern United States*. Ann Arbor: University of Michigan Press.
- Kurath, Hans, and Raven I. McDavid. 1961. *The Pronunciation of English in the Atlantic States*. Tuscaloosa, Ala.: University of Alabama Press.
- Kurath, Hans, et al. 1939–1943. *Linguistic Atlas of New England*. 3 vols. Providence: Brown University Press.
- Labov, William. 1963. The social motivation of a sound change. *Word* 19: 273–309.
- Labov, William. 1966. *The Social Stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.
- Labov, William. 1972a. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.
- Labov, William. 1972b. *Language in the Inner City: Studies in the Black English Vernacular*. Philadelphia: University of Pennsylvania Press.
- Labov, William. 1991. The three dialects of English. In Penelope Eckert (ed.), *New Ways of Analyzing Sound Change* (New York: Academic Press), 1–44.
- Labov, William. 2007. Transmission and diffusion. *Language* 83/2: 344–387.
- Labov, William, Sharon Ash and Charles Boberg. 2006. *The Atlas of North American English: Phonetics, Phonology and Sound Change*. Berlin: Mouton de Gruyter.
- Labov, William, Ingrid Rosenfelder, and Josef Fruehwald. 2013. One hundred years of sound change in Philadelphia: Linear incrementation, reversal, and reanalysis. *Language* 89/1: 30–65.
- Labov, William, Malcah Yaeger and Richard Steiner. 1972. *A Quantitative Study of Sound Change in Progress*. Philadelphia: U.S. Regional Survey.
- Leap, William L. 1993. *American Indian English*. Salt Lake City: University of Utah Press.
- Marckwardt, Albert H. 1957. *Principal and Subsidiary Dialect Areas in the North-Central States*. *Publication of the American Dialect Society No.* 27. Tuscaloosa, AL: University of Alabama Press.
- Martinet, André. 1955. *L'Économie des changements phonétiques*. Berne: Francke.
- McCarthy, Corrine. 2011. The Northern Cities Shift in Chicago. *Journal of English Linguistics* 39: 166–187.
- Moulton, William G. 1960. The short vowel systems of Northern Switzerland: A study in structural dialectology. *Word* 16: 155–182.
- Moulton, William G. 1962. Dialect geography and the concept of phonological space. *Word* 18: 23–32.
- Murray, Thomas E. 1986. *The Language of St. Louis, Missouri: Variation in the Gateway City*. Bern: Peter Lang.
- Pederson, Lee, Susan L. McDaniel and Carol M. Adams, eds. 1986–1993. *Linguistic Atlas of the Gulf States*. 7 vols. Athens, GA: University of Georgia Press.
- Petyt, K.M. 1980. *The Study of Dialect: An Introduction to Dialectology*. London: Andre Deutsch.
- Purnell, Thomas C. 2009. The vowel phonology of urban southeastern Wisconsin. *Publication of the American Dialect Society* 94/1: 191–217.

- Reed, Carroll E. 1961. The pronunciation of English in the Pacific Northwest. *Language* 37/4: 559–564.
- Scargill, Matthew Henry, and Henry J. Warkentyne. 1972. The Survey of Canadian English: A report. *English Quarterly* 5/3: 47–104.
- Shuy, Roger W. 1962. The Northern-Midland Dialect Boundary in Illinois. *Publications of the American Dialect Society* 38.
- Stanford, James N., Thomas A. Leddy-Cecere and Kenneth P. Baclawski. 2012. Farewell to the founders: Major dialect changes along the East-West New England border. *American Speech* 87/2: 126–169.
- Story, G. M., W. J. Kirwin and J. D. A. Widdowson (eds.). 1982. *Dictionary of Newfoundland English*. Toronto: University of Toronto Press.
- Thomas, Erik R. 1997. A rural/metropolitan split in the speech of Texas Anglos. *Language Variation and Change* 9: 309–332.
- Thomas, Erik R. 2001. *An Acoustic Analysis of Vowel Variation in New World English*. *Publication of the American Dialect Society No. 85*. Durham, NC: Duke University Press.
- Thomas, Erik R. 2010. A longitudinal analysis of the durability of the Northern-Midland dialect boundary in Ohio. *American Speech* 85: 375–430.
- Trager, George L., and Bernard Bloch. 1941. The syllabic phonemes of English. *Language* 17/3, 223–246.
- Warkentyne, Henry J. 1973. Contemporary Canadian English: A report of the *Survey of Canadian English*. *American Speech* 46/3–4: 193–199.
- Wassink, Alicia Beckford. 2016. The vowels of Washington State. *Publication of the American Dialect Society* 101/1: 77–105.
- Webster's Third New International Dictionary. 1961. Springfield, MA: Merriam-Webster, Incorporated.
- Weinreich, Uriel. 1954. Is a structural dialectology possible? *Word* 10: 388–400.
- Wells, J. C. 1982. *Accents of English*. Cambridge, U.K.: Cambridge University Press.
- Wolfram, Walt. 1969. *A Sociolinguistic Description of Detroit Negro Speech*. Washington, D.C.: Center for Applied Linguistics.
- Wolfram, Walt. 1974. *Sociolinguistic Aspects of Assimilation: Puerto Rican English in New York City*. Washington, DC: Center for Applied Linguistics.
- Wolfram, Walt, and Natalie Schilling-Estes. 1997. *Hoi Toide on the Outer Banks: The Story of the Ocracoke Brogue*. Chapel Hill, NC: University of North Carolina Press.
- Woods, Howard B. 1999. *The Ottawa Survey of Canadian English*. Kingston, ON: Strathy Language Unit, Queen's University.

27 Dialects of German, Dutch, and the Scandinavian Languages

SEBASTIAN KÜRSCHNER

27.1 Basic Facts

German and Dutch, like English, are West-Germanic languages. German is a major official language in Germany, Austria, Liechtenstein, Switzerland, and Luxembourg, covering a large section of central Europe; it is also an official language in Belgium, where it is spoken near the German border. Outside these countries it is further recognized as a regional official language in the South Tyrol region of Northern Italy, as a regional language along the eastern border of France and as a minority language along the southern border of Denmark. With more than 90 million speakers, German is one of the major languages of the European Union. Modern standard German descends from Old High German, the oldest known documents dating from around 800 AD. The standard language has also spread to regions traditionally characterized by Low German dialects, which are structurally different from High German and more closely connected to Dutch.

The standardization of German was heavily influenced by the invention of the book press in the fifteenth century and by the distribution of the High German bible translation by Martin Luther from the sixteenth century on. A spoken standard variety started developing in the eighteenth century but still displays much variation today. In Austria, and even more in Germany, dialects have been subject to regionalization, or regional leveling: Use and knowledge of local dialects has decreased strongly in recent decades, resulting in a tendency towards dialect loss in Northern Germany. In Switzerland, by contrast, we still find stable diglossia, with dialects used in almost all spoken communication except the most formal situations.

The traditional dialects of Dutch are found in the Netherlands and in Flanders, the northern part of Belgium. Together with Suriname, a former Dutch colony, these countries form the Dutch Language Union. In Europe, about 23 million speakers use Dutch as a first language. The Dutch language area also covers bordering regions of France and Germany, where Dutch is a minority language. Modern Dutch goes back to Low Franconian dialects, which were distinguishable from High German dialects by around the eighth century. The first Old Dutch texts we know of date from the tenth century. Just as with German, the start of language standardization is connected to the invention of printing in the fifteenth century. Standard Dutch is based mainly on Hollandic dialects, with some Brabantian elements.

A spoken standard based on the written language developed in the nineteenth and twentieth centuries and is referred to as ABN (*Algemeen Beschaafd Nederlands*, "Common Civilized Dutch").

Standard Dutch differs between the Netherlands and Belgium, and markedly non-standard speech is generally more common in Belgium than in the Netherlands, although both countries show a strong tendency toward dialectal regionalization. German and Dutch traditionally formed a dialect continuum, but growing influence of the standard languages and a general decrease in dialect use has produced a break in the traditional continuum along the national border.

Frisian, a third language of the West Germanic continuum, is spoken in parts of northern Germany and the Netherlands. West Frisian is spoken mostly in the Dutch province of Friesland and on adjacent islands, whereas East and North Frisian are spoken along the coast and Frisian islands of Germany. Munske (2001) provides a handbook of Frisian dialects. Space constraints prevent this chapter from giving further consideration to Frisian, or to other less commonly spoken West Germanic languages, like Afrikaans, Luxemburgish, and Yiddish.

Turning to the North Germanic or Scandinavian languages, Danish (with about 6 million speakers), Faroese (60,000), Icelandic (330,000), Norwegian (5 million), and Swedish (10 million) all descend from Proto-Norse. Danish, in addition to being the official language of Denmark, is used on the Faroe Islands, on Greenland and as a regional language in the Schleswig section of northern Germany, while Swedish, outside Sweden, has co-official status in Finland, where it is spoken by about 300,000 inhabitants along the south and west coasts.

Faroese and Icelandic are partially mutually intelligible, as are Danish, Norwegian, and Swedish. The latter proximity reflects, in addition to shared North Germanic ancestry, a common history in the medieval Hanseatic trading league (thirteenth to sixteenth century). In the main Scandinavian trade cities, intensive contact with and settlement by Low German tradesmen had a strong homogenizing influence on emerging standard languages, especially Copenhagen Danish and Stockholm Swedish. The resulting similarities, especially among standard varieties, allow mainland Scandinavians to communicate with each other using a receptive multilingual mode termed "semi-communication."

In Norway, there are two written standard languages, called *bokmål* "book language" and *nynorsk* "New Norwegian." *Bokmål* is a variety heavily influenced by Danish, reflecting Norway's status as part of the Danish kingdom until 1814. In a reaction against this influence, Ivar Aasen created *nynorsk*, in the mid-nineteenth century, on the basis of West Norwegian dialects. Today, however, most Norwegians (85–90%) claim to use *bokmål*, whereas only 10–15% use *nynorsk*.

There are considerable national differences in Scandinavian dialect use today. In Denmark, dialects went through heavy levelling during the twentieth century; today, the traditional dialects are hardly used. They were replaced by regional standard varieties differing mainly in prosody, while urban Copenhagen speech functions as the most common model for change (cf. Pedersen 2003). Leveling also occurred in Norway and Sweden, but with less drastic consequences (Kristensen and Thelander 1984). Communication in Norway is still polydialectal and no official form of a supraregional spoken standard language exists.

27.2 Main Data Collections and Sources

27.2.1 German

The history of dialectology has been heavily influenced by pioneering work on German dialects in the late nineteenth century. The major figure in this work was Georg Wenker, who undertook the collection of dialect data from all across the nineteenth-century German Empire.

Wenker's idea was to collect a comparable set of dialectal features from a large number of localities, giving a dense overview of regional variation. After a test phase with regionally limited projects, resulting in the first dialect atlas ever produced (*Sprach-Atlas der Rheinprovinz nördlich der Mosel sowie des Kreises Siegen* 1878), he was funded for a large-scale investigation beginning in 1887. To keep the task manageable, Wenker used an indirect methodology. He developed a set of 40 sentences in Standard German, the so-called "Wenkerrsätze," which included major phonological and morphological dialect variables. The sentences were then sent to schoolteachers all over the country, who were asked to translate them into local dialects with the help of their students. Wenker received around 45,000 completed forms from around 41,000 localities. These formed the basis of his *Sprachatlas des deutschen Reichs* (1888–1923), an atlas of hand-drawn maps displaying the regional variation in the data. Some of these maps were subsequently published in a simpler black and white format as the *Deutscher Sprachatlas* (1927–1956). The whole atlas has now been made available online by the Marburg research group and equipped with tools for comparison with other maps, modern dialect atlases, and other resources such as sound files (<http://www.diwa.info>, published in an extended version at <http://www.regionalsprache.de/>).

A further large-scale collection, focusing on lexical variation, was conducted with a similar methodology in the 1930s and 1940s, resulting in the *Deutscher Wortatlas* (DWA, 1951–1980).

Although Wenker's atlas provides a rich documentation of German dialects at the end of the nineteenth century, its main shortcomings quickly became clear. Due to its indirect method, the data had been collected in written form, so the phonetic information was poor. At the same time, a different atlas project was conducted in France, which was similar in geographical scope (all the dialects of France), but differed radically in methodology (see Hall, this volume). The *Atlas Linguistique de la France* (ALF, Gilliéron and Edmont 1902–1910) was collected in a direct manner, that is, the data come from face-to-face interviews (see Bailey, this volume). Since this method requires more resources, the number of locations surveyed (639) is much smaller than in Wenker's project, but the phonetic quality is much higher.

Subsequent research on German dialects was influenced by the evident advantages of the French method. Projects, therefore, became smaller, focusing on regional rather than national surveys, but the range of linguistic variables investigated with large questionnaires (*Fragebücher*) was much higher than in Wenker's atlas. The focus was on lexical and phonological criteria, but some morphological and syntactic variation was also covered. In the 1930s, a first project covering all German-speaking areas of Switzerland was begun, resulting in the *Sprachatlas der deutschen Schweiz* (SDS, 1962–1997). Further regional projects built on this methodology, spreading from Switzerland to the east and north. Space does not permit all of these to be discussed here, but they include: *Vorarlberger Sprachatlas mit Einschluß des Fürstentums Liechtenstein, Westtirols und des Allgäus* (VALTS); *Südwestdeutscher Sprachatlas* (SSA); *Sprachatlas von Oberösterreich* (SAO); *Thüringischer Dialektatlas* (ThDA); the six atlas projects forming the *Bayerischer Sprachatlas* (BSA), mapping the dialects of nearly the whole federal state of Bavaria; and several atlases as part of the series *Deutscher Sprachatlas: Regionale Sprachatlanten*. French dialectology added atlases of German dialects in Alsace (*Atlas Linguistique et Ethnographique de l'Alsace*) and Lorraine (*Atlas Linguistique et Ethnographique de la Lorraine germanophone*). Further projects are still in progress, resulting in most of the southern parts of the German language area being documented in atlases today.

In contrast with these "monodimensional" atlas projects, in which representative speakers for each location were sought mainly among older, non-mobile locals with rural occupations, the *Mittelrheinischer Sprachatlas* (MRhSA) added a social dimension by including a subsample of younger residents from each location, who commuted to nearby cities (and were thus mobile) and had manual instead of rural occupations. Some of the atlas' maps

compare the speech of the older and younger generations, allowing for the study of ongoing changes in apparent time. Two atlas projects within the *Bayerischer Sprachatlas* integrated even more social dimensions, in dealing with the metropolitan areas of Munich and Nuremberg (*Sprachregion München* (SRM) as part of the *Sprachatlas von Oberbayern*, and *Sprachregion Nürnberg* (SRN) as part of the *Sprachatlas von Mittelfranken*).

Beyond atlases, German dialects are also well documented in dialect dictionaries, following a tradition of compiling dictionaries for specific regions (*Territorialwörterbücher*), such as the *Badisches Wörterbuch* or the *Thüringisches Wörterbuch*. More than 30 such areal dictionaries have appeared or are still being published. There is also a tradition of dialect documentation in monographs. Especially at the end of the nineteenth and the first half of the twentieth century, many doctoral theses reported detailed data on the phonology of their authors' local dialects, seeking to test the Neogrammarian claim that sound laws have no exceptions. Much of this research, up to 1985, is catalogued in two bibliographies (Wiesinger and Raffin 1982; Wiesinger 1987). Schirmunski (2010 [1962]) provides a typological overview of German dialects.

Important handbooks on German dialects include the HSK volumes on dialectology (Besch *et al.* 1982–1983), sociolinguistics (Ammon *et al.* 2002–2006) and Language and Space (Auer and Schmidt 2010; Lameli *et al.* 2011). The latter are not specifically dedicated to German, but include many examples from German dialects.

27.2.2 Dutch

Large projects on the dialects of Dutch have often been carried out jointly by researchers from the Netherlands and Flanders. Dutch dialectology has also been in strong contact with the German and French traditions. Following the direct method of the ALF, the *Reeks Nederlandse Dialectatlassen* atlas project (RND, 1923–1982) elicited data in a direct manner, with dialect forms of 139 sentences collected from 1,956 sites, but regional comparability of the data is restricted by the long duration of the project, which produced 16 volumes. By contrast, the *Taalatlas van Noord- en Zuid-Nederland* (1938–1972) was collected in an indirect manner, like Wenker's data, with a focus on lexis. This atlas was continued as the *Taalatlas van het Nederlands en het Fries* (1981–1988).

The indirect method was also used to study dialect syntax, producing the two-volume *Syntactische Atlas van de Nederlandse Dialecten* (SAND, 2005–2008). Its conception in terms of level of linguistic analysis rather than region and its focus on syntax are quite distinct from the German tradition. The SAND is connected to two other atlas projects, which used direct elicitation to study other levels of structure: phonological variation in FAND (1998–2005, four volumes) and morphological in MAND (2005–2008, two volumes). Data from these projects are now available on-line from the Meertens Institute, the main center of Dutch dialectology in the Netherlands today (<http://www.meertens.knaw.nl/cms/en/collections/databases>). Apart from these large-scale projects, there are several additional atlases, including historical atlases and the cross-border *Taalatlas van Oost-Nederland en aangrenzende taalgebieden* (TON, 1957–1963), which includes dialects in Germany.

Lexicographical documentation has also been strong in Dutch dialectology. We can only mention three large-scale, regional projects here: the *Woordenboek van de Brabantse Dialecten* (WBD, completed in 2005); the *Woordenboek van de Limburgse Dialecten* (WLD, completed in 2008); and the *Woordenboek van de Vlaamse Dialecten* (WVD, being published since 1979).

Weijnen (1991) provides a comparative overview of the phonology of Dutch dialects. A bibliography of dialectal works can be found in De Schutter, Gerritsen, and Van Bree (1990). Recently, an up-to-date international handbook on language variation in Dutch also appeared in the series of handbooks on Language and Space (Hinskens and Taeldeman 2013).

27.2.3 Scandinavian Languages

There are only few works treating all the Scandinavian languages in a comprehensive manner. In one of them, which includes a number of maps, Bandle (2011 [1973]) attempts to establish the geographic divisions of North Germanic dialects. He identifies innovation zones and illustrates the geographical spread of innovations, forming dialect zones. Although his work has been criticized and his dialect divisions disputed, Bandle's book remains a valuable reference work on Nordic dialects.

A larger body of work treats the Scandinavian languages separately, although atlases like those of German and Dutch dialects are comparatively rare. One early atlas, *Kort over de danske folkemål med forklaringer* (1912), examines phonological, morphological, and lexical variation in Danish dialects, while southern Sweden was mapped more recently in the *Südschwedischer Sprachatlas* (1965–1970), but no comprehensive atlas exists of Norwegian, Swedish, Faroese, or Icelandic dialects.

Among lexicographic work, we should mention the dialect dictionary of Jutland Danish (*Jysk ordbog*) and that of the Danish islands (*Ømålsordbogen*). For Norwegian, *Trønderordboka* is a regional dialect project, which forms part of the basis of the *Norsk Ordbok* (Norwegian dictionary), representing Norwegian dialects and the written standard of *Nynorsk*. For Swedish, the *Ordbok över Sveriges dialekter* (1991–) and the *Ordbok över Finlands svenska folkmål* (1982–) provide supra-regional documentation, supplementing a couple of regional dictionary projects. Bandle *et al.* (2002, 2005) is a handbook on the history of the North Germanic languages with chapters on the development of dialectal variation.

27.3 Major Regional Dialect Divisions

In all of the languages considered in this chapter, dialects form a geographic continuum characterized more often by transition zones than by abrupt changes. The identification of dialect areas is therefore based on transitional connections among core areas, rather than on clear-cut regional divisions, and the attribution of dialects to particular regions depends on an interpretation by linguists, which in turn depends on methodological considerations. As a result, many differing dialect divisions exist for each language. There is no space to discuss all of these, so we will concentrate here on divisions that are currently widely agreed upon and presented in textbooks.

27.3.1 The Continental West-Germanic Continuum of German and Dutch Dialects

The dialects of German and Dutch are historically connected within a continental West-Germanic dialect continuum. Isoglosses reflecting the Second or High German Consonant Shift are regarded as the primary division of these dialects. In this shift, voiceless stops became fricatives or affricates and voiced stops became voiceless, as shown in Table 27.1. This change separates "High German" dialects, which became the basis of Standard German, from nearly all other Germanic languages (apart from High German-based Yiddish and Luxembourgish).

Map 27.1 shows the locations of the traditional dialects of German around 1900. Some of the regions shown are no longer part of modern Germany and no longer German-speaking; the northeastern Prussian regions, for instance, lost most of their German-speaking population to westward emigration after World War II. Nevertheless, the map indicates a major division corresponding to the geographic extent of the Second Germanic Consonant Shift, called the Benrath isogloss, which runs from west to east, just south of

Table 27.1 Examples of the Second German Consonant Shift in High or Standard German forms, compared to unshifted sounds in Low German and other Germanic languages.

	<i>High German</i>	<i>Low German</i>	<i>Dutch</i>	<i>Swedish</i>	<i>Icelandic</i>	<i>English</i>
<i>p>pf, f</i>	<i>Pfeffer</i>	<i>Peper</i>	<i>peper</i>	<i>peppar</i>	<i>pirpar</i>	<i>pepper</i>
<i>t>ts, s</i>	<i>Zunge</i>	<i>Tung</i>	<i>tong</i>	<i>tunga</i>	<i>tunga</i>	<i>tongue</i>
	<i>Wasser</i>	<i>Water</i>	<i>water</i>	<i>vatten</i>	<i>vatn</i>	<i>water</i>
<i>k>x</i>	<i>Buch</i>	<i>Book</i>	<i>boek</i>	<i>bok</i>	<i>bók</i>	<i>book</i>
<i>d>t</i>	<i>Vater</i>	<i>Vader</i>	<i>vader</i>	<i>fader (far)</i>	<i>faðir</i>	<i>father</i>

**Map 27.1** Traditional dialects of German around 1900. (Adapted from map at: http://www2.ku.edu/~germanic/LAKGD/Dialect_Regions_Germany.shtml.)

Essen, Magdeburg, and Berlin. This separates the northern-most or “Low German” group of dialects (*niederdeutsche Dialekte*), which do not show any effects of the Shift, from the central or “Middle German” dialects (*mitteldeutsche Dialekte*) and southern or “Upper German” dialects (*oberdeutsche Dialekte*), which show increasing effects of the Shift as one moves south. The Benrather line reflects postvocalic *k*-shift, for example, in different forms of the verb *machen*, the German cognate of English “to make.” North of the line we find the form *maken*, with a voiceless stop as in English; south of it we find *machen*, with a voiceless fricative. Low German shares unshifted stops with all Dutch dialects except some in Limburg, which

show the *k*-shift in a few words (e.g., *ik*, “I” > *ich*, the basis of the Uerdingen isogloss running north from the Benrath isogloss in the west). The western-most group of unshifted dialects is called Low Franconian and stretches over large parts of the Dutch-speaking area and part of Germany. Other dialect groups crossing the current border between the Netherlands and Germany are North Low Saxon dialects in the north, and Westphalian south of this area, both forming the western part of West Low German.

The dialects south of the Benrath line are further divided into two primary groups by the Germersheim isogloss, which separates northern, unshifted /p/ in *Appel*, “apple,” from southern, affricated /pf/ in *Apfel*, the form heard in Standard German. Between this line and the Benrath line, the Middle German dialects, spanning central Germany from Cologne to Leipzig, display only some of the subprocesses of the Shift, with much variation, especially in the west. South of the Germersheim isogloss, the Upper German dialects show nearly all sub-processes of the Shift, although non-postvocalic *k*-shift is restricted to the southern-most area, where Standard German *Kind* “child” > /kxint/.

While the main division of West Germanic dialects is thus based on a single important sound change, dialect groups within the main regions just discussed are divided based on several phonological and morphological criteria, which cannot be treated in any detail here (see Wiesinger 1987 and Hinskens and Taeldeman 2013 for detailed reviews). Low German dialects are further divided into Eastphalian dialects (West Low German), Mecklenburg-West Pommeranian dialects (northern East Low German), and Brandenburg dialects (southern East Low German). Middle German dialects are divided into a Western group of Middle Franconian, Rhine Franconian and Hessian dialects, and an Eastern group of Thuringian and Upper Saxonian dialects. The Upper German area is divided into Alemannic (including Swabian), East Franconian, and Bavarian dialects. While all of the large dialect groups just mentioned are found within Germany, the southwestern Alemannic dialects extend into Switzerland, Liechtenstein and western Austria, and the southeastern Bavarian dialects into the remainder of Austria.

Dutch dialects are usually divided into six sub-groups, classified mainly on the basis of phonological characteristics, as in the following classification by Hinskens and Taeldeman (2013): 1) a southwestern area including West Flemish, French Flemish, western Zeeland Flemish, and Zeelandish dialects; 2) an East Flemish area; 3) a Brabantine area; 4) a southeastern Limburg area; 5) a northwestern area including the dialects of Holland and Utrecht, with an important division between rural dialects and the urban dialects of Amsterdam, Den Haag, Rotterdam, and Utrecht; and 6) a northeastern area, in which Low Saxon dialects form a continuum with the North Low Saxon group of German dialects.

27.3.2 The Scandinavian Languages

A major division of the Scandinavian languages and dialects into western and eastern groups involves monophthongization of the Proto-Nordic diphthongs *ai, *au, and *ey. These vowels remain diphthongs in western Scandinavian varieties, including Faroese, Icelandic, and West Norwegian dialects, whereas eastern varieties, including Danish and some Swedish and Norwegian dialects, have monophthongs instead. For example, from Proto-Nordic *ai we get Icelandic *steinn*, “stone,” versus Swedish *sten*; from *au we get Icelandic *dauður*, “dead,” versus Swedish *död*; and from *ey we get Icelandic *reykur*, “smoke,” versus Swedish *rök*.

While our focus in this section will be on the more populous continental dialects, a few words on Faroese and Icelandic should be said. Both languages are based on West Nordic varieties spoken by Norwegian settlers. While Icelandic remains fairly homogeneous today, with very limited regional variation, Faroese has split up into many dialects that differ phonologically, with a principal division between northern and southern varieties.

In addition to the West-East divide, the continental North Germanic dialect continuum shows a North-South divide, resulting from a number of southern innovations: the southern zone comprises all of Denmark and parts of southern Sweden, whereas the more conservative northern zone includes the more densely populated areas of central Sweden. Southern varieties exhibit weakening of stops and strong effects of vowel apocope. The continuum also involves variation in the realization of a prosodic toneme contrast related to old mono- and bisyllabics. The contrast is realized with tones in most dialects of Norwegian and Swedish as well as some dialects of southern Denmark, but as a prosodic phenomenon called *stød*, a contrastive glottal stop, in the Northern Danish dialects.

Danish dialects are traditionally grouped into western, central, and eastern types. The western dialects are spoken on the Jutland peninsula; the central group, also called Insular Danish, on the main islands of Fyn and Sjælland (Zealand, which includes the capital, Copenhagen); and the eastern dialects only on the island of Bornholm, in the Baltic Sea, although they previously also included dialects of southern Sweden, which belonged to the Danish Kingdom until 1658 and was Danish-speaking before its transfer to Sweden. This division is based primarily on the degree of vowel reduction in unstressed syllables, which ranges from the preservation of full vowels in eastern dialects (as in Swedish), to reduction to schwa with facultative apocope in Insular Danish, to unrestricted apocope in western dialects (leaving exceptions aside). Jutland is further divided by morphological isoglosses, including pro- versus enclitic definiteness and diverging gender systems.

Norwegian and Swedish dialect divisions are based mainly on phonological and, secondarily, morphological variation, although space considerations prevent us from discussing these variables in detail. Norwegian dialects are classified into four large areas: East, West, North, and Trøndelag. For Swedish, six dialect areas are usually distinguished: South, Geatish, Svealand, Norrland, East (comprising Finnish and Estonian Swedish), and Gotland. The Baltic island of Gotland is perhaps a special case, being the historical home of Old Gutnish, a separate branch of Old Norse; modern Gutnish, however, is generally considered to be a dialect of Swedish and shares many features with East Swedish dialects.

27.4 Research Topics

German, Dutch, and Scandinavian dialectology has addressed many diverse research topics, reflecting the strong academic traditions and well-documented dialects in all of these countries. It is impossible to discuss all of these topics here, so we will focus on some of the most important methodological directions taken in previous research. We shall pass over traditional dialect mapping, already discussed in Section 27.2 above, and start with methodological innovations in the second half of the twentieth century, when other fields of linguistics, including pragmatics and sociolinguistics, began to influence dialectology, opening the field to a variationist paradigm that today is often referred to as "Language and Space" (as in the handbook by Auer and Schmidt 2010). The view of dialects as relatively stable, homogeneous systems, on which traditional atlases were based, became less acceptable to modern dialectologists. Another factor inspiring dialectologists was the obvious fact that dialects of all of the languages discussed here were changing rapidly in the twentieth and twenty-first centuries, prompting a search for the reasons for these changes.

Research addressing these issues took contact between dialects and surrounding varieties into account. It became clear that dialects were in constant contact not only with other dialects, but also with the standard language. In this view, the dialect speaker was seen to have a repertoire of dialectal and standard varieties at his disposal, which can either be cognitively separated (*diglossia*) or connected on a continuum ranging from dialect-related speech to standard-related speech (*diaglossia*). Projects adopting this perspective have now been

carried out in all the countries discussed here. Modern documentation projects like the MRhSA mentioned above tried to capture the regionalization of dialects, and new methods for eliciting speech at different levels between base dialect and standard have been tested. A theory of the dynamics of regional languages that models these processes was recently published as a new framework called “language dynamics” (*Sprachdynamik*, Schmidt and Herrgen 2011). There have been several approaches to the description of dialectal features shaping the regionally conditioned inner variability of standard languages (often referred to as *Umgangssprachen*), for example, the *Wortatlas der deutschen Umgangssprachen* (Eichhoff 1977–2000), and, most recently, the *Atlas zur deutschen Alltagssprache* (<http://www.atlas-alltagssprache.de/>).

Another research direction inspired by sociolinguistics considers dialectal variation from the point of view of the language users (“perceptual linguistics” or “folk linguistics,” see Preston, this volume), that is, their mental knowledge of dialectal variation, its connection to mental concepts of areal space and their evaluation of dialects. Studies in this tradition go back to the 1940s in the Dutch speech area, especially to Weijnen’s use of the so-called “pijltjesmethode” (“little arrow method”): participants are asked to use little arrows to indicate on a map which dialectal varieties of surrounding localities they consider to be close to their own varieties. Further work in perceptual dialectology included a recent project on German dialects at the University of Kiel. Current research focuses on the mental conceptualization and evaluation of dialects by linguistic laymen, including attitudes toward dialects.

Methodological progress has also been aided by the ever-growing use of computers for data processing. One such application is the use of algorithms for the automatic calculation of linguistic distances between dialects, mostly involving pronunciation, but also applied to morpho-syntactic and lexical variables. Such “dialectometric” analyses (see Heeringa and Prokic, this volume), although heavily dependent on the method of comparison, provide objective measures of linguistic variation and can be applied to geographical variation, resulting in maps that can be compared with linguists’ interpretations of the respective dialectal landscapes. Progress in computerized data processing has also improved the accessibility of both data and edited work, using the Internet: see, for example, the digitized edition of the Wenker atlas mentioned above, the *Digitaler Verbund der Dialektwörterbücher* (DWV; digitized dialect dictionaries), or the *Baydat* database (<http://www.baydat.uni-wuerzburg.de>), which collects all of the data from the large-scale atlas projects in Bavaria.

As for levels of dialect description, after a long focus on lexis, phonology and, to a lesser extent, morphology, syntactic variation is currently the focus of several projects, for example, an atlas of Swiss German dialect syntax, a project on the dialects of Hesse, and a Scandinavian Dialect Syntax Project (ScanDiaSyn, at <http://uit.no/scandiasyn>).

A few other research directions deserve at least a brief mention. Sociolinguistic dialectology has emphasized urban dialects and regarded varieties and styles as reservoirs of semiotic markers for social meaning. This kind of research also resulted in the re-interpretation of language histories in historical sociolinguistic studies. Still other linguistic subfields have included the study of dialect variation: for instance, onomastics (see the German atlas of family names, *Deutscher Familiennamenatlas* (DFA)), or research on prosody in everyday speech.

27.5 Future Research

Despite the diversity of approaches taken to studying dialects of the languages discussed in this chapter, it is possible to identify a few desiderata as profitable directions for future research. For example, although thorough documentation exists for many dialect areas,

it is lacking in others. For example, many regions in Scandinavia remain unmapped in this sense. At present, a collaborative project called *Sprachvariation in Norddeutschland* (SiN) is trying to provide a clearer picture of the changes affecting dialects in Northern Germany, resulting in the *Norddeutscher Sprachatlas* (NOSA, 2015–2017). For North Rhine-Westphalia, the *Dialektatlas mittleres Westdeutschland* (DMW) is currently prepared.

In well-documented areas, the data produced by large-scale projects over more than 100 years can now be used to document and interpret dialect change in real time. Internet-based instruments like those provided by the Marburg research group (see the digitized edition of the Wenker atlas and *Regionalsprache.de* above) are very helpful in this respect, and may be able to identify centers of innovation in dialectal contact. What is more, much of the data collected in regional atlas projects has not yet been analyzed, leaving many future opportunities for in-depth studies of German inflectional morphology and word-formation.

One question that is still largely unaddressed is the effect of dialect variation on intelligibility. Progress has recently been made in studying the intelligibility of closely related languages within the Germanic language family, but the intelligibility of dialects is still nearly undocumented. Some work has been done on the mutual intelligibility of Dutch regional varieties, and for German, the intelligibility between Moselle and Rhine Franconian dialects has been assessed (Schmitt 1992), but large-scale projects of this type are lacking.

Still further areas for future research can only be briefly mentioned. Studies in multilingualism provide a methodological framework that might be relevant to models of the competence of dialect and standard varieties. For example, comparisons of dialect contact and language contact are still lacking, and there has not been much research yet on the advantages that multilingualism in the sense of dialect and standard varieties provides, for instance by promoting higher degrees of intelligibility of other closely related varieties. A connection to the study of ethnolects, such as the prominent case of “Turkish-German,” is also relevant here (e.g., Auer 2013). The perspective that dialects can be learned as second varieties at any time of life has also only seldom been taken in dialectological research.

Connections to applied linguistics should also be mentioned. For instance, dialects have been considered a negative influence for students educated in the standard language for a long time, led by a strong standard ideology in many of the countries discussed here. In light of contemporary multilingual perspectives, this view could be reevaluated and new didactics developed, using dialect knowledge positively rather than banning dialects from schools.

REFERENCES

- Ammon, Ulrich, Norbert Dittmar, Klaus J. Mattheier, and Peter Trudgill (eds.) 2002–2006. *Sociolinguistics. An international handbook of the science of language and society*. 2nd edition. 3 vol. Berlin/New York: de Gruyter (HSK 3).
- Atlas Linguistique de la France.* (1902–1910). By Jules Gilliéron and Edmont Edmont (eds.). Paris: Champion.
- Atlas Linguistique et Ethnographique de l'Alsace.* (1969–1984). By Ernest Beyer, Arlette Bothorel-Witz, Raymond Matzen, Marthe Philipp, and Sylviane Spindler (eds.). Paris: Éd. du Centre Nationale de la Recherche Scientifique.
- Atlas Linguistique et Ethnographique de la Lorraine germanophone.* 1977. By Marthe Philipp, Arlette Bothorel-Witz, and Guy Levieuge (eds.). Paris: Éd. du Centre Nationale de la Recherche Scientifique.
- Auer, Peter. 2013. Ethnische Marker zwischen Varietät und Stil. In *Das Deutsch der Migranten*, edited by Arnulf Deppermann, 9–40. Berlin: de Gruyter.
- Auer, Peter, and Jürgen Erich Schmidt (eds.) 2010. *Theories and methods (Language and space, vol. 1)*. Berlin/New York: de Gruyter (HSK 30.1).

- Badisches Wörterbuch.* (1925–). By Ernst Ochs, Karl Friedrich Müller, Gerhard W. Baur, Rudolf Post, and Tobias Streck. Lahr: Schauenburg & München: Oldenbourg.
- Bandle, Oskar. 2011. *Die Gliederung des Nordgermanischen*. Reprint of the original ed. 1973. Tübingen/Basel: Francke.
- Bandle, Oskar, Kurt Braunmüller, Ernst Håkon Jahr, Allan Karker, Hans-Peter Naumann, Ulf Telemann, Lennart Elmevik, and Gunn Widmark (eds.) 2002, 2005. *The Nordic languages. An international handbook of the history of the North Germanic languages*. 2 vol. Berlin/New York: de Gruyter (HSK 22).
- Bayerischer Sprachatlas.* (1997–). Consisting of six sub-projects: I. Sprachatlas von Bayerisch-Schwaben. II. Sprachatlas von Mittelfranken. III. Sprachatlas von Unterfranken. IV. Sprachatlas von Nordostbayern. V. Sprachatlas von Niederbayern. VI. Sprachatlas von Oberbayern. Heidelberg: Winter.
- Besch, Werner, Ulrich Knoop, Wolfgang Putschke, and Herbert E. Wiegand (eds.) (1982–1983). *Dialektologie. Ein Handbuch zur deutschen und allgemeinen Dialektforschung*. 2 vol. Berlin/New York: de Gruyter (HSK 1).
- Deutscher Familiennamenatlas.* (2009–). By Konrad Kunze and Damaris Nübling (eds.). Berlin/New York: de Gruyter.
- Deutscher Sprachatlas.* (1927–1956). By Ferdinand Wrede, Walther Mitzka, and Bernhard Martin (eds.). Marburg: Elwert.
- Deutscher Sprachatlas. Regionale Sprachatlanten.* (1961–1972). By Ludwig Erich Schmitt, Karl Kurt Klein, Reiner Hildebrandt, and Kurt Rein (eds.). Consisting of five atlases. Marburg: Elwert.
- Deutscher Wortatlas.* (1951–1980). By Walther Mitzka. Gießen: Schmitz.
- De Schutter, Georges, Marinel Gerritsen, and Cor van Bree (eds.) 1990. *Dialectsyntax*. (Taal en Tongval 3).
- Fonologische atlas van de Nederlandse dialecten.* (1998–2005). By Jan Goossens, Johan Taeldeman, Geert Verleyen, and Chris de Wulf. Gent: Koninklijke Academie voor Nederlandse Taal- en Letterkunde.
- Hinskens, Frans, and Johan Taeldeman. (eds.) 2013. *Dutch (Language and space, vol. 3)*. Berlin/New York: de Gruyter (HSK 30.3).
- Jysk Ordbog.* (1970–). By Peter Skautrup et al. Published online at www.jyskordbog.dk.
- Kort over de danske folkemål med forklaringer* (1912). By Bennike, Valdemar, and Marius Kristensen. København: Gyldendal.
- Kristensen, Kjeld, and Mats Thelander. 1984. On dialect levelling in Denmark and Sweden. *Folia linguistica* 18, 223–246.
- Lameli, Alfred, Roland Kehrein, and Stefan Rabanus (eds.) 2011. *Language mapping (Language and space, vol. 2)*. Berlin/New York: de Gruyter (HSK 30.2).
- Mittelrheinischer Sprachatlas.* (1994–2002). By Günter Bellmann, Joachim Herrgen, and Jürgen Erich Schmidt. Tübingen: Niemeyer.
- Morfologische atlas van de Nederlandse dialecten.* (2005–2008). By Georges de Schutter, Ton Goeman et al. Amsterdam: Amsterdam University Press.
- Munske, Horst Haider (ed.) 2001. *Handbook of Frisian studies*. Tübingen: Niemeyer.
- Norddeutscher Sprachatlas.* (2015–2017). By Michael Elmentaler & Peter Rosenberg. Hildesheim: Olms.
- Norsk Ordbok* (1966–). Oslo: Samlaget.
- Ordbok över Finlands svenska folkmål.* (1982–). Helsinki: Svenska litteratursällskapet i Finland.
- Ordbok över Sveriges dialekter.* (1991–). Uppsala: Institutet för språk och folkminnen.
- Pedersen, Inge Lise. 2003. Traditional dialects of Danish and the de-dialectalization 1900–2000. *International Journal of the Sociology of Language* 159, 9–28.
- Reeks Nederlandse Dialectatlassen.* (1925–1967). By Edgard Blancquaert and Willem Pee (eds.). Antwerpen: De Sikkel.
- Schirmunski, Viktor M. 2010. *Deutsche Mundartkunde. Vergleichende Laut- und Formenlehre der deutschen Mundarten*. Reprint of the original ed. 1962. Frankfurt a. M.: Lang.
- Schmidt, Jürgen Erich and Joachim Herrgen. 2011. *Sprachdynamik. Eine Einführung in die moderne Regionalsprachenforschung*. Berlin: Schmidt.
- Schmitt, Ernst Herbert. 1992. *Interdialektale Verstehbarkeit*. Stuttgart: Steiner.
- Sprachatlas der deutschen Schweiz.* (1962–1997). By Rudolf Hotzenköcherle (ed.). Bern: Francke.
- Sprach-Atlas der Rheinprovinz nördlich der Mosel sowie des Kreises Siegen.* (1878). By Georg Wenker. Marburg (hand-drawn maps).
- Sprachatlas des deutschen Reichs.* (1888–1923). By Georg Wenker. Marburg (hand-drawn maps).
- Sprachatlas von Mittelfranken.* (2003–2014). By Horst Haider Munske, and Alfred Klepsch (eds.). Heidelberg: Winter.
- Sprachatlas von Oberbayern.* (2008–2011). By Ludwig M. Eichinger (ed.). Heidelberg: Winter.

- Sprachatlas von Oberösterreich.* (1998–). Linz: Adalbert-Stifter-Institut des Landes Oberösterreich.
- Sprachregion München.* 2005. By Kurt Rein (ed.). (*Sprachatlas von Oberbayern*, supplementary issue). Heidelberg: Winter.
- Sprachregion Nürnberg.* 2004. By Alexander Mang. (*Sprachatlas von Mittelfranken*, vol. 6). Heidelberg: Winter.
- Südschwedischer Sprachatlas.* (1965–1970). By Sven Benson. Lund: Gleerup.
- Südwestdeutscher Sprachatlas.* (1989–2011). By Hugo Steger, Volker Schupp, and Eugen Gabriel (eds.). Marburg: Elwert.
- Syntactische atlas van de Nederlandse dialecten.* (2005–2008). By Sjef Barbiers, Hans Bennis et al. Amsterdam: Amsterdam University Press.
- Taalatlas van het Nederlands en het Fries.* (1981–1988). By Gesinus G. Kloeke et al. Leiden: Brill.
- Taalatlas van Noord- en Zuid-Nederland* (1938–1972). By Gesinus G. Kloeke et al. Leiden: Brill.
- Taalatlas von Oost-Nederland en aangrenzende taalgebieden.* (1957–1963). By Klaas Heeroma. Assen: Van Gorcum.
- Thüringischer Dialektatlas.* (1961–1965). By Hermann Hücke. Berlin: Akademie-Verlag.
- Thüringisches Wörterbuch.* (1966–2006). By Karl Spengenberg, Wolfgang Lösch, and Susanne Wiegand. Berlin: Akademie-Verlag.
- Trønderordboka.* 2007. By Tor Erik Jenstad and Arnold Dalen. Bergen: Fagbokforlaget.
- Vorarlberger Sprachatlas mit Einschluß des Fürstentums Liechtenstein, Westtirols und des Allgäus.* (1985–2006). By Eugen Gabriel (ed.). Bregenz: Vorarlberger Landesregierung.
- Weijnen, Antonius A. 1991. *Vergelijkende klankleer van de Nederlandse dialecten.* S'Gravenhage: SDU.
- Wiesinger, Peter, and Raffin, Elisabeth. 1982. *Bibliographie zur Grammatik der deutschen Dialekte. Laut-, Formen-, Wortbildungs- und Satzlehre. 1800 bis 1980.* Bern/Frankfurt a. M.: Lang.
- Wiesinger, Peter. 1987. *Bibliographie zur Grammatik der deutschen Dialekte. Laut-, Formen-, Wortbildungs- und Satzlehre. 1981 bis 1985 und Nachträge aus früheren Jahren.* Bern/Frankfurt a. M. etc.: Lang.
- Woordenboek van de Brabantse Dialecten.* (1967–2005). By Antoon A. Weijnen, Jan van Bakel et al. Assen: Van Gorcum.
- Woordenboek van de Limburgse Dialecten.* (1983–2008). By Antoon A. Weijnen, Jan Goossens et al. Assen: Van Gorcum.
- Woordenboek van de Vlaamse Dialecten.* (1979–). By Magda Devos, Hugo Ryckeboer et al. Gent: Academia Press.
- Wortatlas der deutschen Umgangssprachen.* (1977–2000). By Jürgen Eichhoff. Bern/München: Saur.
- Ømålsordbogen.* (1992–). Odense: Universitets-Jubilæets danske Samfund.

ONLINE RESOURCES

- Elspaß, Stephan, and Robert Möller (eds.) (2003–). *Atlas zur deutschen Alltagssprache.* <http://www.atlas-alltagssprache.de/>
- Schmidt, Jürgen Erich, and Joachim Herrgen (eds.) (2001–2009). *Digitaler Wenker-Atlas (DiWA).* Erste vollständige Ausgabe von Georg Wenkers "Sprachatlas des Deutschen Reichs". 1888–1923 handgezeichnet von Emil Maurmann, Georg Wenker und Ferdinand Wrede. Marburg:

- Forschungszentrum Deutscher Sprachatlas. <http://www.diwa.info/>
- Schmidt, Jürgen Erich, Joachim Herrgen, and Roland Kehrein (eds.) (2008–). *Regionalsprache.de (REDE).* Forschungsplattform zu den modernen Regionalsprachen des Deutschen. Marburg: Forschungszentrum Deutscher Sprachatlas. <http://www.regionalsprache.de/>

28 Dialects of French

DAMIEN HALL

28.1 Basic Facts

“Dialect of French” has at least two meanings. In most academic texts, it refers to a regionally or socially differentiated variety of French, comprehensible to a speaker of Standard (Reference) French, and whose historical origin can be traced to the language of medieval Paris, not to one of the sister varieties spoken in other regions of France (Section 28.1.2). In this definition (used in this chapter), these sister varieties are not “dialects of French.” Particularly in France itself, however, what counts as “(a dialect of) French” is often unclear. Many lay people and some academics may call closely related Romance varieties spoken in France “dialects” (as opposed to “languages”), and may also say they are dialects of French.

28.1.1 Types of Variety and Differences in Terminology in France

The continuum of indigenous Romance varieties in France extends from French/*français* (High) to *patois* (Low), with *français régional* “regional French” and probably *dialecte* “dialect” between them, as well as other terms (Lodge 1993, Müller 1985). There is nothing “official” or standardized about these terms. Some are unambiguous, but others are certainly more defined by social and discourse context.

As of 2013, French was the sole official language of 14 countries and was a co-official language in a further 15 countries (Leclerc 2015; see also Rossillon *et al.* 1995). Where it is official, French is often a speaker’s only language, particularly in Europe and in towns. In comparison with speakers of other varieties, speakers of French may be seen as urbane and sophisticated. At the other end of the continuum lie the *patois*. *Patois* usually refers to the indigenous Romance varieties of the Northern Gallo-Romance area that did not develop into Standard French (although the term is occasionally used for other non-standard varieties of France, whether Romance or not). An indicative list of Northern Gallo-Romance *patois* is Picard, Norman, Angevin, Poitevin, Champenois, Berrichon, Bourguignon, and Lorrain in France, and Walloon in Belgium, although other names and other subdivisions can be found. *Patois* are usually rural varieties, now spoken only by older people, although there is the important exception of the Picard variety *Ch’ti* or *Ch’ti-mi*, spoken in the industrialized area around Lille (Armstrong and Pooley 2010, 169–175). *Patois* cannot be used in wider society, where they have no official status, and they may not be mutually intelligible with standard French. Despite the pejorative connotations that the term *patois* can have, academic linguists working in French (not least Gilliéron and Edmont 1902–1912 in the *Atlas Linguistique de France* (ALF)) have also used *patois* to refer to these varieties in a neutral way, and many of their speakers also do so, with no belittling intended; but the stigma generally associated with the term has meant that it is often avoided (Séguy 1973b).

Somewhere between *français* and *patois* falls *dialecte*. Like *patois*, it has two distinct connotations which may well not overlap (Knecht 1997). On the one hand, *dialecte* can be used in the non-evaluative sense of a regional form of some overarching language, usually without official status. On the other hand, in France, the term is often used by lay people in a way that is not clearly different from the evaluative sense of *patois*. To avoid this ambiguity, some Francophone linguists now use the neutral term *variété*, much as Anglophones may use “variety.”

Partially because of the different connotations of *dialecte*, dialectological terminology used in French (particularly in France) can be different from that used in English. Even a word-for-word translation of the title of this handbook might have a different reading. While “dialectology” in academic texts in English will usually have a linguistic sense (see Hickey, this volume), the primary sense of *dialectologie* in French is usually an ethnographic, documentation-related sense (Chaurand 1972, Hall 2013): the documentation of the indigenous non-French varieties of France. The (socio)linguistics of the varieties documented are secondary at best. Since the beginning of the French dialectological tradition (the very first page of the *Notice* to the ALF), documentation of indigenous varieties has been considered best done by collecting words for things. All this is in striking contrast to the dominant Anglophone academic dialectological tradition, where “now it would be shocking [...] if someone tried to keep apart [sociolinguistics and dialectology]” (Chambers and Trudgill 1998, xiii).

28.1.2 The Typology of French

French is a Romance language, descended from Latin through the Western Romance and Gallo-Romance subfamilies (Posner 1996, Greub and Chambon 2008). Lower in the *Stammbaum*, however, there is controversy about two questions: whether French itself is descended from a single medieval variety of Paris and its hinterland; and whether the modern forms of that Paris variety’s sisters (Norman, Picard, etc.) are forms of French.

The first question is about the origins and nature of the medieval variety of Paris: whether it was a regional indigenous Romance variety, which has been called *francien*, or a koiné composed of elements of the varieties spoken in areas adjoining Paris. Lodge (2004) gives much detail, arguing for a spoken koiné; *contra*, Grübl (2013) contends that modern French could also have arisen at least in part from a koinéization of regional written varieties. On the second question, this chapter will consider Norman, Picard, and so on only in passing, because we take the view that they are not forms of French—and resolving the question is the subject of a separate part of the literature. We do note, however, that not all scholars agree with this view.

The typology of the varieties of French spoken outside Europe, where French is not indigenous, is a different question. There, the debate is not about the path from Latin to the modern varieties, but about which indigenous varieties of France were spoken by the original colonizers (thus becoming the substrate of the modern “colonial” varieties). There has been a considerable amount of taxonomic research in Canada and the other French-speaking communities of North America, and it is generally accepted that the plurality of original French-speaking migrants to those territories came from the West of France. Much research on this theme is reviewed in Mougeon and Beniak (1994).

28.2 Major Regional Dialect Divisions and Linguistic Variables

Almost all the main linguistic studies are themselves on particular regional dialects; we therefore describe regional dialect divisions before describing the main studies.

28.2.1 France

The most important dialect division in France runs from West to East in a rough arc, separating varieties jointly referred to as *langue d'oil* in the North from those classified as *langue d'oc* in the South, and separating the Northern and Southern accents of Standard French. The Southern accent is the marked one, as the accent used in most major population centers of Northern France is considered standard (Lyche 2010). Figure 28.1 shows the regional varieties of France mentioned in Section 28.1.1 above. In the map, bold lines separate the linguistic sub-families of Gallo-Romance, and also separate Gallo-Romance from other languages spoken in France. The southernmost bold line between regional varieties (that is, the southern boundary of the *langue d'oil* and *Francoprovençal* areas) is also the division between the Northern and Southern accents of Standard French. (*Francoprovençal*, is the preferred spelling of that variety; it appears with a hyphen in the map for cartographical reasons: Kasstan 2016.)

Two main factors differentiate the standard/Northern accent and the Southern accent: vowel systems (both oral and nasal), and treatment of unaccented schwa. Standard/Northern French has at least 10 (variably) contrastive oral vowels, plus schwa, and usually three (variably) contrastive nasal vowels. The main variables are that some Northern speakers contrast /a/ and /ɑ/ (*patte* “paw” [pat] and *pâte* “paste” [pat]), whereas most merge them to /a/ (*patte* and *pâte* both [pat]); and some Northern speakers contrast /œ/ and /ɛ/ (*brun* “brown” [bʁœ̃] and *brin* “sprig” [bʁɛ̃]), whereas most merge them to /ɛ/ (*brun* and *brin* both [bʁɛ̃]). Most Northern speakers pronounce schwa only when it is necessary to avoid disfavored clusters of three or more consonants (the *loi des trois consonnes* “law of three consonants”). Not all clusters are disfavored. Thus, *il nous retrouve* “he meets up with us” can be pronounced [ilnuʁtʁuv] or [ilnuʁɛtʁuv] by a Northern speaker, even though the first realization contains a sequence of three consonants, because the second and third consonants of the sequence are the permissible cluster [tʁ]; on the other hand, *appartement* “apartment” must be pronounced [apartɛ̃mã] and not *[apartmã], because, without schwa, the word contains the sequence [ɛtm], and neither [ɛt] nor [tm] is permissible. This rule also applies across word-boundaries in fluent speech.



Figure 28.1 The indigenous non-French varieties of France.

By contrast, many Southern speakers pronounce schwa in any syllable ending in <e>; thus, *il nous retrouve* must be pronounced [ilnu:kət̪kuva]. The minimal vowel system of Southern French (Coquillon and Durand 2010) has seven contrastive oral vowels and usually four nasal vowels. The word “contrastive” is emphasized, because the *loi de position* “law of position” applies to both front and back mid-vowels: “open vowels in closed syllables and close vowels in open syllables.” Open vowels are also found in open syllables where the following syllable contains schwa (e.g., *creuse* “hollow” (f.sg.) [krœzə]), suggesting that these syllables may be closed underlyingly. Southern speakers may have a four-way nasal-vowel contrast, so that *brun* and *brin* are always pronounced differently; but a stereotyped Southern French accent in fact has few nasal vowels overall, as they are replaced with the corresponding oral vowel plus a nasal consonant. Thus, *je plante une rose* “I plant a rose” would be pronounced [ʒɔplāty̪nøz] by a Northern speaker, but [ʒɔplant̪t̪ynøzə] by a stereotypical Southern speaker. Southern speakers with these stereotypical features are becoming rarer, however, as the Northern accent spreads South (Armstrong and Pooley 2010).

Among consonants, rhotic variation is highly salient when it occurs. Most speakers in France now use the uvular fricative [χ], though the uvular trill [r] (“*r grasseyé*”) can still be heard. Alveolar taps [ɾ] and trills [r], once standard in French, are severely recessive (Durand and Rossi-Gensane 2010).

In Francophone Europe, most non-phonological linguistic variables—like expression of negatives and of futurity—are not geographically but socially distributed (Tuaillet 1983). One celebrated geographical example, however, is the *passé surcomposé* “double-composed past tense,” said to be a feature of regional French in the Francoprovençal-substrate area of Eastern France and Western Switzerland (see de Saussure and Sthioul 2012).

Armstrong and Pooley (2010) survey recent and classic studies of finer variation within France. Within the North, we can particularly single out *Ch’ti* or *Ch’ti-mi* around Lille (the name for the local accent of French as well as one of the names for the local variety of Picard). Studies within the South are fewer, but are also detailed by Armstrong and Pooley (2010).

28.2.2 Canada

The main difference between Canadian French (CanFr) and French French (FrFr) is between their vowel systems. (This makes the controversial—Kircher 2012—assumption that we consider FrFr as the “basic variety” to which others should be compared.)

Several major differences are apparent (Walker 1984). CanFr has phonemic vowel-length for many qualities of vowel: for example, *mettre* “put” (infinitive) [mɛ:t̪] is robustly contrasted with *maitre* “master” [mɛ:t̪]. In FrFr, this distinction is now very recessive. CanFr also displays phonetic lengthening of the “intrinsically long vowels” (/ø o œ/ and the nasal vowels) in any context, and of any vowel when followed by the *consonnes allongantes* “lengthening consonants” /v z ʒ ʁ/. Long stressed vowels in CanFr can develop a homorganic glide, for example, *pâte* /pat/ > FrFr [pa(:t̪)], CanFr [pa“(:t̪)]. CanFr also laxess non-lengthened high vowels in closed final syllables, and in certain pretonic syllables: *pipe* “pipe” /pip/ regularly > [pip], *plume* “feather” /plym/ > [plym], *coupe* “cup” /kup/ > [kup], *vulgaire* “vulgar” /vyl'gɛr/ > [vyl'gɛr]. Front close-mid and open-mid vowels are distinct in all positions for most speakers of CanFr, whereas many FrFr speakers merge them to close-mid when stressed: *fée* “fairy” /fe/ > FrFr and CanFr [fe], but *fais do* /fe/ > FrFr often [fe], CanFr [fe].

Just as in FrFr, rhotic variation in CanFr is salient. [χ] is the dominant prestige pronunciation, but, unlike FrFr, [r] is also a significant variant in CanFr (Sankoff and Blondeau 2007 and references therein). Also highly salient to FrFr speakers is CanFr affrication of /t d/ before high front vowels (which also occurs in FrFr, but not as widely).

Quebec French (though not the French of Acadia, i.e., the Atlantic Provinces) has an important syntactic-semantic feature that is much less prevalent in FrFr (Roberts 2013 and references therein): future tense variation according to verb polarity, whereby the inflected future is used with non-negated verbs (1):

- (1) *il faudra faire un tour*
it be-necessary.FUT.3 PS make.INFIN a tour

"You will have to do a tour."

and the periphrastic future with negated ones (2).

- (2) *ça ne va plus se reproduire*
that NEG go.PRES.3 PS anymore itself do-again.INFIN

"That won't happen again."

CanFr also has an interrogative particle, *-tu* [ty]/[ty] (Vecchiato 2000; compare *-ti* in Northern France), which can be placed after the verb in any declarative sentence to make it interrogative ((3) and (4)).

- (3) *il vient*
he come.PRES.3 PS

"He's coming."

- (4) *il vient-tu?*
he come.PRES.3 PS-INTERROGATIVE

"Is he coming?"

Other non-phonological variables in CanFr vary sociolinguistically and stylistically rather than geographically.

At a more fine-grained level, CanFr is usually divided into Quebec, Ontario and the West on the one hand, and Acadia on the other. Most descriptions of CanFr that do not specify a region describe Quebec French; space prevents a detailed description here but, for Acadian French, see Flikeid (1984) and King and Ryan (1989).

28.2.3 Other Parts of the French-Speaking World

There are fewer detailed linguistic studies of French outside Europe and North America, though see Section 28.3.3.2. For Africa, some studies challenge the initial assumption that there is only one "African French." For Pacific French, personal observation reveals rhotic variation: both [r] and [ɾ] are present, and Tahitian speakers report that [r] is the majority variant there, though it may not be the majority variant throughout French Polynesia.

28.3 Dialects of French: Main Studies and Sources

It can be difficult to conclusively separate sources dealing with varieties of French from those dealing with regional indigenous Romance varieties closely related to French (Section 28.1). The following section will make a separation, but it will become clear that not all authors make the same one.

28.3.1 Atlases

28.3.1.1 Linguistic Atlases of Romance Varieties in French-Speaking Territories

One might think that the main study and source for dialects of French would be the *ALF* (Gilliéron and Edmont 1902–1912). In fact, however, the introductory text for the *ALF* states that it is a collection of Gallo-Romance *patois* (see also Brun-Trigaud, Le Berre, and Le Dû 2005). The *ALF*, therefore, covers minority *langue d'oïl* varieties, but not French itself; *langue d'oc* varieties; and Corsican varieties.

Similarly, mainly lexical differences are documented in the *Nouvel Atlas Linguistique de la France par régions* “New Linguistic Atlas of France by region” (Dauzat 1939, Séguy 1973b, Simoni-Aurembou 2004), which later became the *Atlas Linguistiques et Ethnographiques de la France par regions* (*ALFR*), though some atlases also include grammatical information. *ALFR* will finally include 25 atlases of indigenous minority regional varieties of France (not all of them Romance varieties), if they can be published.

The French tradition of documentation of such varieties has also been adopted elsewhere. In Switzerland, the online *Atlas Linguistique Audiovisuel du Francoprovençal Valaisan* “Audiovisual Linguistic Atlas of Valais Francoprovençal” is being constructed (Boeri *et al.* 1993–2015); in Belgium, the *Atlas Linguistique de la Wallonie* “Linguistic Atlas of Wallonia” has reached Volume 17 (Baiwir 2011).

28.3.1.2 Linguistic Atlases of French in France

No geographical studies of France cover Standard French in the way that, say, Labov, Ash, and Boberg (2006) covers North American English (though see Section 28.5). This might simply be because, for political reasons, the majority attitude in France is that Standard French varies little, and that any variation that does exist is comparatively insignificant.

28.3.1.3 Linguistic Atlases of French Outside France

The largest atlas-type resource on any variety of French outside France is the *Atlas Linguistique de l'Est du Canada* (*ALEC*) (Dulong and Bergeron 1980), covering French-speaking Eastern Canada, though coverage outside Quebec is sparse. Like many French atlases, *ALEC* privileges the lexicon, and mainly uses older men as informants; analysis of *ALEC*'s entries can still show the main differences perceived today between European and CanFr, however, where two or more *ALEC* entries are phonologically similar enough. *ALEC* contains no maps: it lists questions and transcriptions of responses.

To our knowledge there are no linguistic atlases of the French of other parts of the world, although there are many atlases of their indigenous languages.

28.3.2 Dictionaries and Grammars

28.3.2.1 The Académie Française

Many believe that the *Académie Française* “French Academy” has power of codification and ratification over French, but in fact no French-speaking territory has an official dictionary. The *Académie* does issue occasional edicts, some of which attempt to modernize spelling and other aspects of French; however, though these edicts have moral authority for some, they do not have legal force (Schiffman 2006). The *Académie* does have a dictionary (see *Académie Française* n.d.), but its current edition dates from 1935, though a new one is mostly complete as of 2017. The most widely consulted dictionaries in the French-speaking world are *Le Robert* (Robert and Rey 2001) and the “*Larousse*” (see the *Larousse* website: *Larousse* 2014), which are more frequently updated than the *Académie* dictionary and therefore more descriptively authoritative.

The Académie Française also had a grammar, but disowned it in 1932 for being overly prescriptive—a fact which also goes against the predominant image of the Académie as a rule-giver. The most widely used grammar in the French-speaking world is now Grevisse and Goosse (2016).

28.3.2.2 *Regional Dictionaries and Grammars*

Regional dictionaries and grammars have been published in many French-speaking areas, particularly Canada. As has been said, Quebec French is usually taken to be the CanFr norm, and since the 1960s there has been a debate in Canada about the possibility and nature of a Quebec/CanFr norm separate from the international French norm (Bigot 2011). Bergeron (1997) and DesRuisseaux (1990) are dictionaries of the features which distinguish Quebec French from FrFr; Cormier (1999) is a dictionary of the other main Canadian variety, Acadian French, with an extensive historical and lexical introduction.

Specific dictionaries and grammars are scarcer for other parts of the Francophone world. The *Base de données lexicographiques panfrancophone* “Pan-francophone lexicographic database” (AUF/TLFQ 2014) gives regional lexical items for 20 French-speaking countries and territories. Among major print dictionaries, we can cite Delcourt (1998–1999) for Belgian French, Thibault and Knecht (2004) for Swiss French, and Valdman and Rottet (2009) for Louisiana (Cajun) French. A wide range of dictionaries of African and Pacific varieties of French are also available—as of 2015, many are published by Édicef but also freely downloadable. Lexilogos 2002–2017 contains a comprehensive catalogue.

28.3.3 *Linguistic Descriptions and Surveys*

28.3.3.1 *Descriptions and Surveys of French Worldwide and Within France*

The Programme “*Phonologie du Français Contemporain*” “‘Phonology of Contemporary French’ Program” (PFC; Durand, Laks, and Lyche 2009) is now the major descriptive project in French phonological variation. Researchers worldwide are invited to record speakers using a common interview protocol, and to analyze the data for a common set of features. Anonymized recordings can be uploaded to the PFC website (PFC 2004–2016), where researchers who have contributed to the database can use them, and some are made available to the general public. A number of research and teaching books based on the material have now appeared (e.g., Detey *et al.* 2010). PFC recordings have been made on all the continents where French is spoken, though the greatest concentration is in France.

Even so, there are still not many structural surveys of varieties of FrFr compared to the number on English in particular. This is possibly because of the predominance of interactional (macro-) sociolinguistics and not variationist (micro-) sociolinguistics in France (Gadet 2003). There are some classic studies of French phonological structure, however: Martinet (1945) and Walter (1982) can be singled out. Martinet (1945) is a self-report study of upper-middle-class speakers from all over France, remarkable for the conditions it was done under (a prisoner-of-war camp), and the first study of its kind. Walter (1982) gives systematic accounts of the phonology of speakers from not only France but also Belgium and Switzerland. Slightly different is Carton *et al.* (1983): its title mentions “accents” but, given the age of the informants, there may be doubt about whether this book actually documents regional accents of French, or regional indigenous minority varieties. The study can still be recommended for its rigor, though: it is done from recordings, not from interviewer transcription, and there is some comparative phonological analysis. Hall (2013) gives an overview of microsociolinguistic surveys in the North of France; the urban areas around Lille and Paris have received particular attention. For the South of France, many references are given by Armstrong and Pooley (2010, 186ff).

28.3.3.2 Descriptions and Surveys of French Outside France

In Francophone Europe outside France, Belgium and Switzerland have been most extensively covered. The most comprehensive recent study of Belgian French is Hambye (2005), which also gives references to other significant work on Belgian varieties, such as that of Michel Francard. Many other references can be found in Armstrong and Pooley (2010) and Detey *et al.* (2010).

Switzerland holds a special place in variationist linguistics, as the location of the first real-time trend study, carried out in the early twentieth century in the village of Charmey. The study was on Francoprovençal, and so lies outside the scope of this chapter, but more information can be found in Labov (1994). The largest single body of studies of the French of Switzerland can be found among PFC publications (Detey *et al.* (2010), PFC 2004-8).

A common finding in studies of Belgian and Swiss French is that, counter to the common French perception of a single “Belgian accent” or “Swiss accent,” each country has more than one accent. Impressionistically, however, Belgian and Swiss French do have features in common. Most salient is their shared non-FrFr vocabulary (*septante* “seventy” and *nonante* “ninety” for standard French *soixante-dix* and *quatre-vingt-dix*; certain Swiss cantons also have *huitante* and *octante* for standard *quatre-vingts* “eighty”). Belgian and Swiss French are also both characterized as being “slower” than standard French: possibly for this reason, much research on intonation in French has been carried out using data from these countries (Simon 2012, Auchlin *et al.* 2004).

Outside France, the best studied variety is CanFr. The standard works on its phonology are Walker (1984) and Ostiguy and Tousignant (1993). Both state that they describe Quebec French, but their descriptions are also valid for Acadian French. For the grammar of Quebec French, a standard work is Léard 1995. Sociolinguistically, CanFr has an extremely rich literature, particularly on Montreal and Ottawa-Gatineau: it starts with the work of Gillian Sankoff and collaborators for Montreal, and with Shana Poplack and collaborators for Ottawa-Gatineau. Much of Sankoff’s work on French has been based on a corpus of Montreal French with interviews recorded in 1971, 1984, and 1995 (Vincent, Laforest, and Martel 1995). The work encompasses variables from the phonological (changes in rhotic realization) to the morphosyntactic (negation: Sankoff and Blondeau 2007 and references therein). Poplack and her collaborators’ work on French focusses on morphosyntactic variation (Poplack 2015) and relations between French and other languages present in the communities investigated (Poplack and Dion 2012).

Sociolinguistic work on other varieties of French is much rarer. In Africa, it is important to note that the role of French in society is usually different from its role in Europe and Canada, because French is not official or even a national language in all the countries where it is spoken. It is therefore a second language for many of its African speakers—or at least they are equally fluent in French and some other language. Queffélec (2000) is a wide-ranging survey of French in African societies, and PFC publications (PFC 2004-16, Detey *et al.* 2010) contribute data analysis.

28.4 Themes of Past and Current Research

The scope of this section is different from the scope of the corresponding section in other chapters of this book, in order to give a short account of the themes from French dialectological research since the end of the nineteenth century, which have influenced dialectology worldwide. For further details, the reader is referred to Hafner (2006).

28.4.1 Past Research

Many dialectological advances have originated in work on French. ALF and its methodological influence are discussed above. In dialectological theory, Paris (1888) gives the first clear statement that in any linguistic territory (of the same mother language) there are no

clear unitary dialects or dialect borders, but instead a gradual change between dialects. This is true because “dialects” consist of collections of individual linguistic features that never all change together, and so distinct dialect boundaries are seldom created, although politics and geography can impose them. In saying this, Paris launched a debate that is still current about the way in which linguistic change propagates across territories and among speakers (Labov 2007). Paris (1888) is also often cited as the inspiration for the *Atlas Linguistiques de la France par régions* (see Section 28.3). Finally, Joret (1883) uses isoglosses drawn from a dialectological survey of Normandy to segment the territory in which the indigenous minority Romance variety Norman is spoken: as far as we are aware, the first use of purely linguistic criteria to do this.

Because of these forerunners, for some time French was the language of choice for dialectological studies of many other languages, particularly in Europe. Pop (1950) gives a wide-ranging account—and it is noteworthy that Pop was Rumanian, though he latterly worked in francophone Belgium.

28.4.2 Themes of Current Research

A number of advances in sociolinguistic theory have been made using Montreal French data. Many important early papers are collected in Sankoff (1980); a crucial recent publication is Sankoff and Blondeau (2007), adding lifespan change to the taxonomy of relationships between a community and its language. The other major sub-discipline where Regional French data has been used is phonology, and we have mentioned the PFC project (Durand, Laks, and Lyche 2009).

The regional indigenous minority varieties of France have also been used for theoretical work. Montreuil (e.g., 2013) has done much work in theoretical phonology using both the Regional French of Normandy and Norman; Julie Auger and her collaborators have applied Picard data to problems in general sociolinguistics (Villeneuve and Auger 2013) and morpho-syntax (Auger and Villeneuve 2008). Since the early 1970s, much work in dialectometry has been done using Norman data, particularly by Hans Goebl (Goebl, this volume), and that period also saw dialectometrical work on and in the *Atlas Linguistique de la Gascogne* (Séguy 1973a).

28.5 Future Research

A recurring theme of this chapter has been the presence in France of two distinct kinds of non-standard linguistic variety: regional varieties of French itself on the one hand, and regional indigenous minority varieties on the other. There has been much research on the regional indigenous minority varieties, and that fact, coupled with political unwillingness to fully accept regional varieties of the standard language, has meant that much of what might be called the basic research on regional varieties of French is still to be done.

At present, therefore, many of the next few years’ pressing research themes seem to be adding detail on topics about which existing research is relatively scarce. Hall (2013) gives details of the ongoing *Towards A New Linguistic Atlas of France* research project, which is making a survey of phonetic and phonological variation across the large cities of the Northern third of France, in the style of Labov, Ash, and Boberg (2006). This study is intended to complement and deepen the research in the same area mentioned in Section 28.3.3.1. Similar research could also be carried out comparatively easily within other regions of France and Canada. In other regions of the French-speaking world, deeper phonological research of this kind could face accessibility problems, but the research would be desirable if it could be done.

REFERENCES

- Académie Française. n.d. "Académie Française." Accessed March 6 2017. <http://www.academie-francaise.fr>
- Armstrong, Nigel, and Tim Pooley. 2010. *Social and Linguistic Change in European French*. Basingstoke: Palgrave Macmillan.
- Auchlin, Antoine, Laurent Filliettaz, Anne Grobet, and Anne Catherine Simon. 2004. "(Én)action, experientiation du discours et prosodie." *Cahiers de Linguistique Française*, 26: 217–249. Accessed March 6 2017. http://clf.unige.ch/files/4314/4102/7607/11-Auchlin_nclf26.pdf
- AUF/TLFQ 2014 = Agence universitaire de la Francophonie and Trésor de la Langue Française au Québec. 2014. "Base de données lexicographiques panfrancophone." Accessed March 6 2017. <http://www.bdlp.org/>
- Auger, Julie, and Anne-José Villeneuve. 2008. "Ne-deletion in Picard and in regional French: Evidence for distinct grammars." In *Social Lives in Language – Sociolinguistics and multilingual speech communities: Celebrating the work of Gillian Sankoff*, edited by Miriam Meyerhoff and Naomi Nagy, 223–247. Amsterdam: Benjamins.
- Baiwir, Esther. 2011. *Atlas Linguistique de la Wallonie Tome 17*. Liège: Presses Universitaires de Liège.
- Bergeron, Léandre. 1997. *Dictionnaire de la langue québécoise*. Montréal: Éditions Typo. 2nd edition.
- Bigot, Davy. 2011. "De la norme grammaticale du français parlé au Québec." *Arborescences*, 1: 1–18. DOI: 10.7202/1001939ar.
- Boeri, Gisèle, Federica Diémoz, Magda Jezioro, Raphaël Maître, Aurélie Reusser-Elzingre, et al. 1993–2015. "Atlas Linguistique Audiovisuel du Francoprovençal Valaisan." Accessed March 6 2017. <https://unine.ch/islc/alaval>
- Brun-Trigaud, Guylaine, Yves Le Berre, and Jean Le Dû. 2005. *Lectures de l'Atlas Linguistique de la France de Gilliéron et Edmonton*. Paris: Comité des travaux historiques et scientifiques.
- Carton, Fernand, Mario Rossi, Denis Autesserre, and Pierre Léon. 1983. *Les Accents des Français*. Paris: Hachette. Available online (accessed March 6 2017): <http://accentsdefrance.free.fr/>
- Chambers, Jack, and Peter Trudgill. 1998. *Dialectology*. Cambridge: Cambridge University Press. 2nd edition.
- Chaurand, Jacques. 1972. *Introduction à la Dialectologie Française*. Paris: Bordas.
- Coquillon, Annelise, and Jacques Durand. 2010. "Le français méridional: éléments de synthèse." In Sylvain Detey et al. (eds), 185–197.
- Cormier, Yves. 1999. *Dictionnaire du français acadien*. Anjou: Fides.
- Dauzat, Albert. 1939. "Un nouvel Atlas linguistique de la France." *Le Français Moderne* 7(2): 97–101.
- de Saussure, Louis, and Bertrand Sthioul. 2012. "The Surcomposé Past Tense." In *The Oxford Handbook of Tense and Aspect*, edited by Robert I. Binnick, 586–610. Oxford: Oxford University Press.
- Delcourt, Christian. 1998–9. *Dictionnaire du français de Belgique*. Bruxelles: Le Cri. 2 vols.
- DesRuisseaux, Pierre. 1990. *Dictionnaire des expressions québécoises*. Montreal: Bibliothèque Québécoise.
- Detey, Sylvain, Jacques Durand, Bernard Laks, and Chantal Lyche, eds. 2010. *Les Variétés du Français Parlé dans l'Espace Francophone*. Paris: Ophrys.
- Dulong, Gaston, and Gaston Bergeron. 1980. *Le Parler Populaire du Québec et de ses Régions Voisines: Atlas linguistique de l'Est du Canada*. Montreal: Office de la Langue Française.
- Durand, Jacques, Bernard Laks, and Chantal Lyche. 2009. *Phonologie, Variation et Accents du Français*. Paris: Hermès.
- Durand, Jacques, and Nathalie Rossi-Gensane. 2010. "Conversation à Douzens (Aude): Retour sur les deux guerres mondiales." In Sylvain Detey et al. (eds), 2010, DVD, 95–106.
- Flikeid, Karin. 1984. *La Variation Phonétique dans le Parler Acadien du Nord-Est du Nouveau-Brunswick*. New York: Peter Lang.
- Gadet, Françoise. 2003. "Is there a French theory of variation?" *International Journal of the Sociology of Language*, 160: 17–40.
- Gilliéron, Jules, and Edmond Edmonton. 1902–1912. *Atlas Linguistique de France, avec une Notice et une Table, ainsi qu'un Supplément*. Paris: Champion.
- Greub, Yan, and Jean-Pierre Chambon. 2008. "Histoire des dialectes dans la Romania: Galloromania." In *Romanische Sprachgeschichte*, edited by Gerhard Ernst, Martin-Dietrich Glessgen, Christian Schmidt and Wolfgang Schweickard, 2499–2520. Berlin: Walter de Gruyter.

- Grevisse, Maurice, and André Goosse. 2016. *Le Bon Usage*. Louvain-la-Neuve: De Boeck. 16th edition.
- Grübl, Klaus. 2013. "La standardisation du français au Moyen Âge: point de vue scriptologique." *Revue de Linguistique Romane*, 77: 343–383.
- Hafner, Jochen. 2006. *Ferdinand Brunot und die nationalphilologische Tradition der Sprachgeschichtsschreibung in Frankreich*. Tübingen: Gunter Narr.
- Hall, Damien. 2013. "The Linguistic Geography of the French of Northern France: Do we have the basic data?" *Language and Linguistics Compass*, 7(9): 477–499. DOI: 10.1111/lnc3.12046. Accessed March 6 2017. <http://onlinelibrary.wiley.com/doi/10.1111/lnc3.12046/full>
- Hambye, Philippe. 2005. "La prononciation du français contemporain en Belgique: variation, normes et identités." PhD diss., Université Catholique de Louvain. Accessed March 6 2017. <http://hdl.handle.net/2078.1/4883>; <https://dial.uclouvain.be/pr/boreal/object/boreal:4883>
- Joret, Charles. 1883. *Des Caractères et de l'Extension du Patois Normand*. Paris: Vieweg.
- Kasstan, Jonathan. 2016. Denomination and the 'problem' of Francoprovençal. In: *Le nom des langues IV. Nommer des langues romanes*, edited by Jean-Michel Éloy. Leuven: Peeters.
- King, Ruth, and Robert Ryan. 1989. "La phonologie des parlers acadiens de l'Île-du-Prince-Édouard." In *Le Français Canadien Parlé Hors Québec*, edited by Raymond Mougeon and Édouard Beniak, 245–259. Quebec: Presses de l'Université Laval.
- Kircher, Ruth. 2012. "How pluricentric is the French language?" *Journal of French Language Studies*, 22(3): 345–370. DOI:10.1017/S0959269512000014.
- Knecht, Pierre. 1997. "Dialecte." In *Sociolinguistique*, edited by Marie-Louise Moreau, 120–124. Sprimont: Mardaga.
- Labov, William. 1994. *Principles of Linguistic Change, Volume 1: Internal factors*. Malden: Blackwell.
- Labov, William. 2007. "Transmission and Diffusion." *Language*, 83(2): 344–387.
- Labov, William, Sharon Ash, and Charles Boberg. 2006. *The Atlas of North American English*. Berlin: Mouton de Gruyter.
- Larousse. 2014. "Dictionnaires de français." Accessed March 6 2017. <http://www.editions-larousse.fr/collections/langue-francaise>
- Léard, Jean-Michel. 1995. *Grammaire Québécoise d'Aujourd'hui*. Montreal: Guérin.
- Leclerc, Jacques. 2015. "La francophonie dans le monde." Accessed March 6 2017. <http://www.axl.cefan.ulaval.ca/francophonie/francophonieacc.htm>
- Lexilogos. 2002-17. "Dictionnaire francophone." Accessed March 6 2017. http://www.lexilogos.com/francophonie_dictionnaires.htm
- Lodge, R. Anthony. 1993. *French: From dialect to standard*. London: Routledge.
- Lodge, R. Anthony. 2004. *A Sociolinguistic History of Parisian French*. Cambridge: Cambridge University Press.
- Lyche, Chantal. 2010. "Le français de référence: éléments de synthèse." In Sylvain Detey *et al.* (eds), 143–166.
- Martinet, André. 1945. *La Prononciation du Français Contemporain*. Paris: Droz.
- Montreuil, Jean-Pierre. 2013. "Assimilation and opacity in Cotentin and Island Norman: the derivational perspective." *Language Sciences*, 39: 178–188.
- Mougeon, Raymond, and Édouard Beniak, eds. 1994. *Les Origines du Français Québécois*. Sainte-Foy: Laval.
- Müller, Bodo. 1985. *Le Français d'Aujourd'hui*. Paris: Klincksieck. French translation of Müller, Bodo. 1975. *Das Französische der Gegenwart*. Heidelberg: Carl Winter.
- Ostiguy, Luc, and Claude Tousignant. 1993. *Le Français Québécois: Normes et usages*. Montreal: Guérin.
- Paris, Gaston. 1888. "Les Parlers de France." *Revue des Patois Gallo-Romans*, 2: 161–175. Accessed March 6 2017. <http://gallica.bnf.fr/ark:/12148/bpt6k92984n/f161.image>; r=revue%20des%20patois%20gallo-romans.langFR
- PFC 2004-16 = Phonologie du Français Contemporain. 2004-2016. "Le projet PFC." Accessed March 6 2017. <http://www.projet-pfc.net/>
- Pop, Sever. 1950. *La Dialectologie*. Gembloux: Duculot.
- Poplack, Shana. 2015. "Norme prescriptive, norme communautaire et variation diaphasique." In *Variations diasyntétiques et leurs interdépendances*, edited by Kristen Kragh and Jan Lindschouw. Newcastle upon Tyne: Cambridge Scholars Publishing.
- Poplack, Shana, and Nathalie Dion. 2012. "Myths and facts about loanword development." *Language Variation and Change*, 24: 279–315. DOI:10.1017/S095439451200018X.

- Posner, Rebecca. 1996. *The Romance Languages*. Cambridge: Cambridge University Press.
- Queffélec, Ambroise. 2000. "Le français en Afrique noire francophone." In *Histoire de la Langue Française 1945-2000*, edited by Gérald Antoine and Bernard Cerquiglini, 797–837. Paris: Centre National de la Recherche Scientifique.
- Robert, Paul, and Alain Rey. 2001. *Le grand Robert de la langue française*. Paris: Le Robert.
- Roberts, Nicholas. 2013. "The Influence of Linguistic Factors on the Expression of Futurity in Martinique French." *Newcastle and Northumbria Working Papers in Linguistics*, 19(1): 138–151.
- Rossillon, Philippe, Françoise Cauquil, Claude Couvert, and Yves Rey-Herme. 1995. *Atlas de la Langue Française*. Paris: Bordas.
- Sankoff, Gillian. 1980. *The Social Life of Language*. Philadelphia: University of Pennsylvania Press.
- Sankoff, Gillian, and Hélène Blondeau. 2007. "Language Change Across the Lifespan: /r/ in Montreal French." *Language*, 83(3): 560–588.
- Schiffman, Harold. 2006. "Language Policy and Linguistic Culture." In *An Introduction to Language Policy*, edited by Thomas Ricento, 111–125. Malden: Blackwell.
- Séguy, Jean. 1973a. "La Dialectométrie dans l'Atlas Linguistique de la Gascogne." *Revue de Linguistique Romane*, 37: 1–24.
- Séguy, Jean. 1973b. "Les Atlas Linguistiques de la France par Régions." *Langue Française*, 18: 65–90. DOI: 10.3406/lfr.1973.5631.
- Simon, Anne-Catherine. 2012. *La Variation Prosodique Régionale en Français*. Leuven: De Boeck.
- Simoni-Aurembou, Marie-Rose. 2004. "Les Atlas Linguistiques de la France par régions (1939–1970)." *Flambeau*, 30: 1–22.
- Thibault, André, and Pierre Knecht. 2012. *Dictionnaire Suisse Romand*. Geneva: Zoé. 3rd edition.
- Tuaillon, Gaston. 1983. "Régionalismes grammaticaux." *Recherches Sur le Français Parlé*, 5: 227–239.
- Valdman, Albert, and Kevin Rottet. 2009. *Dictionary of Louisiana French*. Jackson: University Press of Mississippi.
- Vecchiato, Sara. 2000. "The *Ti/Tu* Interrogative Morpheme in Québec French." *Generative Grammar in Geneva* 1: 141–163. Accessed March 6 2017. <http://www.unige.ch/lettres/linge/syntaxe/journal/>
- Villeneuve, Anne-José, and Julie Auger. 2013. "'Chtileu qu'i m'freumereu m'bouque i n'est point coér au monne': Grammatical variation and diglossia in Picardie." *Journal of French Language Studies*, 23: 109–133.
- Vincent, Diane, Marty Laforest and Guylaine Martel. 1995. "Le corpus de Montréal 1995: Adaptation de la méthode d'enquête sociolinguistique pour l'analyse conversationnelle." *Dialangue*, 6: 29–46.
- Walker, Douglas C. 1984. *The Pronunciation of Canadian French*. Ottawa: University of Ottawa Press.
- Walter, Henriette. 1982. *Enquête Phonologique et Variétés Régionales du Français*. Paris: Presses Universitaires de France.

29 Dialects of Italy

TULLIO TELMON

29.1 Historical and Sociolinguistic Background

29.1.1 *Italian and Latin Dialects*

It is widely known that Italy has only recently been politically united. Lesser known is that in 1861, when political unification came about, the number of Italians who could speak Italian was minimal. De Mauro (1964) estimates that only 2.5% of the population spoke Italian, whereas the remaining 97.5% spoke only their local dialect. By saying “spoke Italian,” we mean, “able to use, even orally, linguistic norms that have been strongly consolidated, especially in written form, and that have structural bases deriving from Florentine/Tuscan literature.” This double restriction, both diamesic (restricted to writing) and diatopic (restricted to Tuscan forms), explains De Mauro’s otherwise surprising data, which reveal, on one hand, the high degree of illiteracy at this time and, on the other, the small population of Tuscany (and Rome) relative to the rest of Italy.

What we call “Italian” is really just one of the many geographically based forms that Latin, both spoken and vulgar, gradually reached in the Italian peninsula. From Rome, Latin spread to the many non-Latin populations of Italy at different times and in different modes and intensities. These reflect the different relations that each of the various subordinate populations had with the Romans who conquered them, as well as their different languages, which included other Italic languages (e.g., Osco-Umbrian), other non-Italic but Indo-European languages (e.g., Celtic dialects and Greek), and non-Indo-European languages (e.g., Etruscan). The entire process began during the last centuries of the Roman Republic (first to second centuries BC) and continued until the fifth to seventh centuries AD, when spoken Latin ceased to be “vulgar Latin” and instead became “neo-Latin Vernacular.” Forms of neo-Latin Vernacular, which include the Tuscan dialects that became modern Italian, were thus numerous and diverse. This diversification, further modified over time, has led to the modern-day “Italian dialects.” In this definition of “Italian dialects,” the adjective “Italian” is used as a geographical reference (“the dialects of Italy”, as in the title of this chapter) and not in the sense of linguistic genealogy. It is more correct to refer to varieties of speech in Italy as a whole as “vulgar neo-Latin dialects”: Venetian, Tuscan, Calabrian, Piedmontese, and so on, are really “Latin dialects” or “dialects of Latin” (closely related Romance languages), rather than “Italian dialects” in the historical-linguistic sense, since the modern use of “Italian” to mean “the Italian language” really designates only one of these many vulgar neo-Latin dialects (that spoken originally in central Italy).

29.1.2 Temporal Continuity of the Dialects Belonging to the First Dialectization

The first dialectization of Italy refers to the acquisition of Latin by populations speaking different languages (Etruscan, Celtic, Venetic, Umbrian, Oscan, Samnitian, Sicilian, and so on), producing the neo-Latin dialects. This led to a dialect landscape that the current specialized texts date from the eleventh century to today (cf. Grassi, Sobrero, and Telmon 1997 and 2003 or Marcato 2007). It was followed by a second dialectization that reflects the influence of the Tuscan-based Italian national language, which produced new Italian regional modulations that we might call Italian dialects. The first dialectization has generally been seen as a subject for research in history and archaeology, whereas synchronic dialectology is supposed to address only the study of the regional Italian languages produced by the second dialectization, but this division has been blurred for sociolinguistic reasons. In order to quantify the number of dialect speakers in Italy at the time of its unification, De Mauro relied principally on rates of illiteracy: illiterate people who were not Tuscan could not be considered Italian speakers. As a consequence, two sociological classes emerge: illiterate and literate subgroups of the populations speaking dialect and Italian. A large number of other attributes might be implicitly linked to both of these: above all, illiterate or dialect speakers belong to the lower social classes, whereas literate and Italian speakers belong to the higher social classes. Dialectophony, then, came to be seen sociologically as the main marker of a clear social inferiority, whereas learning Italian was seen as the key to emancipation and social success. This was an important component of the ideology supporting a shared national language; as a sign of acquired dignity and in order to achieve social redemption, it was fundamentally important for people to repudiate their dialects, which were perceived as shameful.

Today, however, with the process of Italianization on the verge of success, dialects are still in use. Ironically, the triumph of Italianization has produced a new regard for dialects. Now that language is no longer a matter of stigma, since most people speak Italian and are equal from the linguistic point of view, the ability to speak dialect as well as Italian has assumed a positive social value. Being able to quote or use a typical dialect expression is considered a "*quid pluris*" that only someone who has had the chance to keep his own dialect can manage in a diglossic (or dilalic) manner.

Some public opinion surveys now reflect these trends. Until the 1970s and 1980s, the dominant trend was a progressive growth of italophony, with a complementary decrease in dialectophony. From the beginning of the present century, however, this inverse relation in the use of Italian and dialect came to an end. In particular, new research has found:

- a) a slowing down in the decrease in dialect speakers;
- b) a slowing down in the growth of monolingual Italian speakers; and
- c) an increase in the number of people using both codes.

Nevertheless, from the sociolinguistic point of view of "language death," it remains unclear whether social dignity alone is enough to guarantee the survival of the "(Latin) dialects of Italy," in the absence of a more robust pattern of intergenerational transmission.

29.1.3 Political, Demographic, and Cultural Status

The paradigm of modernization applied to the diglossia of coexisting codes has involved the progressive neglect of the most unusual linguistic codes, which are replaced by the most widely diffused codes. Such neglect is led tendentially by interruption of the transmission of the local language as a mother tongue. This occurred widely over the twentieth century. If we consider the example of Piedmont, we clearly see that between the two world wars the

dialect of Turin, which was the most prestigious linguistic variety of the region and was becoming more and more a regional *koiné*, tended to spread and compete with the local varieties, even undermining them sometimes. After the Second World War, however, the dialect of Turin was obliged to succumb to the powerful expansion of Italian, which was transmitted by mass media and made more authoritative by compulsory schooling and the diffusion of bureaucratic language.

In spite of this, some public opinion polls (the most recent from 2006) display a very strong resistance among the dialects belonging to the first dialectization. According to research on home language use carried out by ISTAT in 2006 on a sample of 24,000 families (cf. <http://www.istat.it/>), 9.6 million Italians today speak their local dialect (almost) exclusively, 27.3 million use only Italian and 20 million use both codes. If we combine the latter group with those who use only dialect, it is apparent that more than 50% of the Italian population has a certain competence in dialect. These figures obviously represent a national average; in detail, the research shows that dialect users are mostly old people living in small towns. From the diatopic point of view, the research carried out by ISTAT states that, with the exception of Tuscany and partially of Latium (regions where the perception of the differences between Italian and dialects is minimal), the most intense italicization took place in the northwestern regions (Aosta Valley, Piedmont, Liguria, Lombardy). By contrast, the northeastern regions (Veneto, Friuli, Trentino) and the southern Italian regions show the highest levels of dialectophony. Using the ISTAT data, we have produced a chart showing the contrasts between the maximum level of italicization (Tuscany), of dialect retention (Veneto) and of diglossia (Apulia), with the changes recorded between 2000 and 2006, which we provide here as Figure 29.1.

As can be clearly seen in Figure 29.1, the six-year time lapse between the two surveys reveals that italicophony grew by 1.4% on a national scale (the highest growth was in the province of Bolzano: 4.1%). Again on a national scale, the decrease of dialectophony, at 3.1%, is a bit more marked.

As regards the regions, Tuscany presents the highest level of italicophony (which continues its expansion over the six years) and the lowest level of dialectophony (which declines between the two surveys). As might be predicted, since Tuscan dialects were the starting point for Italian, Tuscans scarcely perceive the vernacular aspects of their speech. It is highly probable that the scant 4.1% of Tuscans who reported using dialect, which decreased to 2.8%

	Only or mainly Italian		Only or mainly dialect		Both Italian and dialect		Other languages	
	2000	2006	2000	2006	2000	2006	2000	2006
Italy	44,1	45,5	19,1	16,0	32,9	32,5	3,0	5,1
Tuscany	83,0	83,9	4,1	2,8	10,1	8,8	2,2	4,0
Veneto	22,6	23,6	42,6	38,9	29,8	31,0	3,9	6,0
Apulia	31,6	33,0	17,7	17,3	49,8	47,9	0,4	0,9
Bolzano/Bozen	21,1	25,2	1,8	1,5	5,7	4,1	70,0	65,5

Figure 29.1 Percentage of Italian, dialect, diglot, and other languages speakers in Tuscany, Veneto, Apulia, and around the province of Bolzano in 2000 and 2006 (cf. ISTAT).

in 2006, is confined to the edges of the region, including Lunigiana in the north and Maremma or Aretino in the south. Veneto is the region where dialects flourish the most and Apulia is where the use of language and dialects together (both alternated and mixed) is highest. The region (or better, the province) where we can observe the highest percentage of foreign languages (specifically German, which is safeguarded through some special international agreements) is Sudtirol/Alto Adige (province of Bolzano), with a percentage of 70% in 2000, which in 2006 decreased to 65.5%.

Returning to the general data from the survey, regarding changes over time it should be emphasized that between 2000 and 2006 the use of Italian has decreased in Aosta Valley, Lombardy, Emilia Romagna, Umbria, and Basilicata; during the same period the use of dialect has increased in Umbria, Basilicata, and Sardinia and the use of both codes (even if slightly decreased on a national scale) has increased in Aosta Valley, Veneto, Emilia Romagna, Umbria, Molise, Campania, Calabria, and Sicily, thereby compensating for the decreasing exclusive use of dialect.

In order to complete the picture of the status of the primary dialects, it is necessary to consider that besides the dialects derived from Latin (which we have been focusing on until now), Italy is also home to dialects derived from autochthonous and foreign linguistic minorities. Unlike the Latin-derived dialects, some non-Latin minority languages are protected by law (Law 482 of 19.12.1999). This selective treatment produces notable asymmetries in the use of minority varieties: for instance, the Grecanic of Corigliano d'Otranto, a protected minority language in Apulia, might be used during a town council meeting, yet the same might not be true of the "primary" dialects of Veneto, since, though they are also not "dialects of Italian" in the sense being used here, Law 482 does not apply to them.

In a few cases there are local standardization traditions within certain regions. The case of Turin, where the local dialect was evolving into a regional *koiné* in the mid-twentieth century before the widespread diffusion of Italian, was mentioned above. In most places, however, the local languages display a total dialectal fragmentation, with a different local language for each site. Nonetheless, in the most important Italian cities (other than Florence), some signs of local literary dialect traditions can be observed: Goldoni for Venice; Carlo Porta for Milan; Edoardo Ignazio Calvo or Brofferio for Turin; Gioacchino Belli for Rome; Giovan Battista Basile and more recently Salvatore Di Giacomo for Naples; and Giovanni Meli and more recently Ignazio Buttitta for Palermo and Sicily.

29.2 Sources

Graziadio Isaia Ascoli is generally recognized as the founder of dialectology in Italy. In 1873 he founded the *Archivio Glottologico Italiano*, which is still active. He also wrote the *Saggi ladini*, which can be considered the first scientifically founded description of a dialectal set; apart from the detailed research it involved, it provided a precious and enlightening methodological guide, which was helpful to later generations of scholars. During the century following Ascoli's death in Milan in 1907, Italian dialectologists sought to build on his work, in order to achieve a fuller knowledge of the Italian dialectal landscape. The following subsections point out some highlights of this tradition, classified by type of publication.

29.2.1 Atlases

The linguistic atlas tradition is well represented in Italy by two major works: the *Atlante italo-svizzero* (AIS; Jaberg and Jud 1928–1940), also known by its German title, *Sprach- und Sachatlas Italiens und der Südschweiz*, covering Italy and Italian-speaking Switzerland; and the

more recent *Atlante Linguistico Italiano* (ALI; Bartoli, Matteo *et al.* 1997 et ss.), conceived and started by Matteo Bartoli and then carried forward by a number of other scholars, including its current director, Lorenzo Massobrio.

Unlike other countries, Italy has never produced any national plan for regional atlases, but many such works have nevertheless arisen from individual efforts, such as the ASLEF for Friuli (Pellegrini 1972–1986); the ALT for Tuscany (Giacomelli 2000); or the ALD and ALD2 for the Ladin-speaking regions of north-eastern Italy and southern Austria (Goebel 1985–1998, 1999–2013). We should also mention the ALEPO for Piedmont (Canobbio and Telmon 2004 ss.), which is focused on Gallo-Romance, Occitan Alpine, and Franco-Provençal in north-western Italy; the ALS for Sicily (Ruffino and D'Agostino n.d.), which, after gathering almost 400 survey responses, has been publishing some copious dialectal and sociolinguistic materials in the form of a dictionary-atlas; and the ALBa for the Basilicata region of southern Italy (Del Puente 2008 ss.). As regards Calabria, two projects deserve mention. The first is the ALECal, (Trumper and Maddalon n.d.), whose phonetic, grammatical, and ethnolinguistic data are being published on a multimedia CD ROM, together with those of other countries joining the project. The second is the ASiCa (Krefeld n.d.), which focuses on syntactic data. A wider view of Italian geosyntax can be found in the ASIt (Benincà n.d.). Publication of atlases has been greatly facilitated in recent years by the web, which resolves many economic problems and material encumbrances. Recent atlas projects taking advantage of these benefits include the ALCam for Campania (Radtke n.d.) and the VIVALDI (Bauer and Kattenbusch n.d.), which offers acoustic materials from different regions of Italy.

29.2.2 *Bibliographies*

The *Bibliografia della linguistica italiana* (Hall 1958), produced by an American scholar, is the most important bibliographic reference work for scholars working on Italian dialects; it was followed by three *Decennial Supplements* (Hall 1969, 1980, 1988). More recently, the *Rivista Italiana di Dialettologia* (RID; see below) offers a bibliographical update of at least two Italian regions in each issue.

29.2.3 *Historical Grammars*

The most complete historical grammar of Italian is by the German author Gerhard Rohlfs, who was one of the collectors of the AIS and was thus able to analyse the historical evolution of Italian dialects with a large number of illustrative examples (Rohlfs 1949–1953, 1967). Historical grammars related to single dialects are less numerous, but two good examples are the *Grammatica storica* of Altamura (Loporcaro 1988) and the *Grammatica diacronica del Napoletano* (Ledgeway 2009).

29.2.4 *Dictionaries*

A major dictionary project still in progress is the *Lessico etimologico italiano* (LEI; Pfister and Schweikard 1979 ss.). At present, 12 volumes and many pamphlets have been published, covering the letters A (volumes I–III), B (vol. IV–VIII) and partially C (vol. XII); as regards D, only five pamphlets have been published, as well as seven for Germanisms. Overall, the project will include about 30 volumes and is supposed to be completed by the end of 2032. The *Vocabolario dei dialetti della Svizzera italiana* (VDSI; Lurà 1952) displays remarkable features and well-organized editing; it recently published its sixth volume and the first pamphlet (*covertón – crená*) of the seventh. The *Lessico dialettale della Svizzera italiana* (LSI; Lurà 2004), in five volumes, documents the lexical heritage of the *Centro di dialettologia e di etnografia*.

Relating to specific regions, we should mention the *Dizionario dei dialetti dell'Abruzzo e del Molise* (DAM; Giammarco n.d.), whose volume VII was published posthumously; and the *Vocabolario Siciliano* (VS; Piccitto, Tropea and Trovato), which has produced its fifth volume.

In regard to dictionaries of this type, my own review of 117 dialectal dictionaries published during the last decade shows that only 17 were produced by authors who are “professional dialectologists,” so scholarly quality in this category of publication is sometimes variable, in terms of completeness of linguistic and ethnographic information, lexicographic accuracy, precision in transcription and spelling, and bibliographic care. The best examples in these terms are the *Dizionario etimologico della Val Tartano* (Bianchini and Bracchi 2003); the remarkable *Dizionario etimologico e storico Tabarchino* (Toso 2004), still limited to the entries between *a* and *cùzò* of the first volume; and the exemplary *Dizionario del dialetto di Montagne di Trento* (Grassi 2009).

29.2.5 Scientific Journals

In addition to the *Archivio Glottologico Italiano*, many other scientific journals have appeared, but a large number of them did not succeed in moving with the times and ceased publication. At present, the most long-lived and important survivors are *L'Italia dialettale: Rivista di Dialettologia Italiana*, founded in Pisa by Clemente Merlo in 1924 and later directed by Tristano Bolelli and, since 2001, Franco Fanciullo; and *Rivista Italiana di Dialettologia: Lingue, Dialetti, Società* (RDI), founded in 1977 by a group of young dialectologists (Gaetano Berruto, Francesco Bruni, Lorenzo Coveri, Annibale Elia, Fabio Foresti, Luciano Giannelli, Glauco Sanga, Alberto Sobrero, Tullio Telmon, Arianna Uguzzoni, and Alberto Zamboni) and aimed at promoting innovative theories and methods in a discipline that seemed, at the time, to be unconcerned with new linguistic theories and interdisciplinarity and their effects on linguistic education. It is also important to mention some other journals, such as *Bollettino dell'Atlante Linguistico Italiano* (BALI), whose first issue was published in 1930; the *Nouvelles du Centre d'Études Francoprovençales René Willien*, published biannually by a research centre in Saint-Nicolas (Aosta) since 1978; the *Bollettino del Centro di Studi Filologici e Linguistici Siciliani*, published since 1961; the *Contributi di Filologia dell'Italia mediana*, directed by Enzo Mattesini and Ugo Vignuzzi; and the *Bollettino Linguistico Campano*, directed by Nicola De Blasi and Rosanna Sornicola since 2002.

29.3 Classification

29.3.1 Dialect Variation: A Practical Example

In order to get to a clearer view of the Italian dialect situation, it may be helpful to consider a practical example of dialect variation. To this end, we refer to chart 1678 of AIS, showing regional forms of the phrase, *Questa donna non mi piace* (“I don’t like this woman”; lit. “This woman doesn’t please me”). Selected data from this chart are reproduced here as Figure 29.2, which shows 13 locations distributed throughout northern and central Italy and Sicily where informants invert the “default” subject-verb order (SV) to VS, including Standard Italian, shown in the first line (*Non mi piace questa donna*). Among these, only the first three—545 Subbiano, in Tuscany; 575 Trevi, in Umbria; and 845 Calascibetta, in Sicily—display the same syntax as standard Italian, although two of them show lexical variation instead, replacing *donna* with *femmina*. The remaining nine locations display important syntactic differences from Standard Italian.

	<i>non</i>	<i>la</i>				<i>non</i>	<i>a me</i>	<i>mica</i>		<i>lì</i>
Standard Italian			non	mi	piace				questa donna	
545 Subbiano			um	me	pjèshe				questa dònna	
575 Trevi			non	me	pjage				questa	
									fémmina	
845 Calascibetta			um	mi	pjaci				shta fimmmina	
139 Galliate					pjàs		<i>a mmi</i>	<i>mmi</i>	sa dònna	
187 Zoagli	<i>a</i>	nu	me	güsta					quèsta dònna	
250 Bienate		<i>la</i>		me	piaz	<i>nu</i>	<i>mi</i>		sta dona	<i>kì</i>
275 Cast. D'A.		<i>la</i>		me	pias	<i>nò</i>			kla dònna	
331 Stenico	<i>no</i>	<i>la</i>		me	pias				kola dònna	
367 Grado	<i>no</i>	<i>la</i>		me	pjaze				quela dònna	
378 Montona	<i>no</i>	<i>la</i>		me	piaze				sta dònna	
427 Ferrara		<i>la</i>	n	am	piazh			<i>brizha</i>	kla dònna	<i>lì</i>
436 Nonantola		<i>la</i>	n	um	piezh			<i>brizha</i>	kla dònna	<i>lì</i>

Figure 29.2 Survey locations where informants invert the subject-verb order of *Questa donna non mi piace* (AIS 1678). In gray, we indicate the “standard” word order; in italics, the added elements, whose basic forms appear as column headings in the top row.

Even where the syntactic structure is similar to that of Standard Italian, we find divergence at other levels of structure: nasal assimilation in [umme] and [ummi]; palatalization of /a/ in ['pjɛʃe]; Sicilian vocalism in ['fimmmina]; and different morphological forms, such as third-person -i in ['pjatʃi]. The syntactically divergent forms are widely distributed, occurring in Piedmont (Galliate: “*piace a me mica questa donna”); Liguria (Zoagli: “*ella non mi gusta questa donna”); Lombardy (Bienate: “*ella mi piace non mi questa donna qui” and Castiglione d’Adda: “*ella mi piace non questa donna”); Veneto (Stenico and Grado: “*non ella mi piace quella donna”); the Istrian peninsula (Montona: “*non ella mi piace questa donna”); and Emilia Romagna (Ferrara, Nonantola: “*ella non mi piace mica questa donna lì”).

As regards the default SV-order syntax, this is found in 188 of the 359 records listed in *AIS* chart 1678, just over half. Geographically, they are distributed over the whole country: 9 in Italian Switzerland and Italian Grigioni; one in Liguria; one in Lombardy; 26 in Veneto, Trentino and Friuli Venezia Giulia; one in Emilia and Romagna; 40 in Tuscany, Umbria and Marche; 28 in Lazio and Abruzzi and Molise; 44 in Campania, Basilicata, Apulia and Calabria; 18 in Sicily (almost everywhere except Calascibetta); and 20 in Sardinia. In the allohotot areas, we find the following forms:

- 715 Faeto (FG), (French Provencal): [sta ffennë më pja pa]
 748 Corigliano d'Otranto (LE) (Griko): [si gjí'neka ε 'mmu pja't[εj]]
 760 Guardia Piemontese (CS) (Provencal): [sta 'fyménë i mi 'pje pa]
 792 Ghorio di Roghudi (RC) Grecanic: [tuti ji'neka de 'mmu pja't[εj]].

While the South and the islands show a close correspondence in syntactic structures, even more than Tuscany, there we instead find differences in phonetics and morphology, for example, in Làconi, on Sardinia (NU): *kusta vémina nommi 'bra:ZiðI* [sic]. This shows several Sardinian features, such as rhotacism of initial /pl-/; conservation of final -et (cf. Latin PLACET) and addition of final /-i/. Other variable forms are found in Palermo: *shta fimmmina ûmmi pjashi*; in Centrache in Calabria: *ssa hìmmmina nommi pjacia*; and in Bari: *këssa fémènë nommë pjashë*.

In the northern regions, where the structural distance from standard Italian negation is wider, we find some structures which are partially similar to German *diese Frau gefällt mir nicht* [lit. "this woman pleases me not"], with the personal pronoun after the verb and the negative adverb at the end, as in Lenz, in the Swiss Grigioni: *kwèla dòna plezh a mo bec*. Above all, however, we find the French simplified structure (without the first negative adverb): *cette femme (ne) me plait pas*. This occurs mainly in Piedmont, Lombardy, and Emilia-Romagna. Some examples are:

- 135 Pettinengo (BI): *sta dòna qui m pjaz mìja*
 142 Bruzolo (TO) (French Provencal): *sla fumna a m pjas pa*
 150 Sauze di Cesana (TO) Provencal: *ekéla fenna ma plaj pa*
 155 Turin: *sa fumna a m pjaz nèn;*
 244 Selino (BG): *sta fümna là m pjäs migä;*
 261 Milan: *quèla dòna lì la me pjas nò;*
 456 Bologna: *shta dòna què la n mè pjèzh brizha.*

The few examples we have reported above may suggest other divisions, motivated by the variable forms of the final adverb, which developed from different Latin sources: from Latin PASSUM in French Provencal and Alps Provencal in the western Piedmont to Latin NEC ENTEM in Turin and the central Piedmont; from Latin MICAM in the eastern Piedmont and largely in Lombardy to Latin NON in Milan and surrounding area; to the Vulgar Latin *BRISARE in Emilia; and so on.

29.3.2 Efforts at Classification

The preceding discussion suggests that the major typological division of Italian dialects in terms of syntax is between Standard Italian and Northern dialectal varieties. From the phonological point of view, however, all dialectal varieties (including those of Tuscany) differ from standard Italian and, above all, among themselves. In fact, there is some truth to the declaration that in Italy, the number of primary dialects (those deriving from vulgar Latin) is equal to the number of municipalities, although it must be recalled that such differences are

always partial and gradual. Achieving a broader level of classification is therefore a difficult challenge, but can be attempted if we focus on the most marked differences. For instance, the above analysis of the sentence “questa donna non mi piace” pointed to some larger conceptual groupings, such as “Northern dialects,” “Central dialects”, “Southern dialects,” and, we might add, “extreme Southern dialects.”

In his 1880 entry on Italian in the *Encyclopaedia Britannica*, Ascoli adopted a classification of Italian dialects that is quite similar to the present situation. His division comprised three major linguistic types:

- 1) a group of dialects derived from Neo-Latin systems that are not peculiar to Italy;
- 2) a group of dialects separate from the Italian system that do not reflect any Neo-Latin system different from that of Italy;
- 3) a group of dialects that are clearly different from Italian and Tuscan, but that form a special system of Neo-Latin dialects together with the Tuscan.

In the first group Ascoli included French Provençal and Provençal dialects of the Aosta Valley and of the western Piedmont, as well as the Ladin dialects of the Dolomites and Friuli. In the second, he included the Sardinian dialects and the so-called “Gallo-Italic” dialects, because of the Gallic substratum of the populations who lived in Piedmont, Liguria, Lombardy, and Emilia. In the third, he included the dialects of Veneto, Corsica, Sicily, and Southern and Central Italy. This third group, moreover, was seen to form together with Tuscan “a special system of Neo-Latin dialects,” as was observed in the above example. In fact, the dialects of Piedmont, Lombardy, and Emilia Romagna (and to a lesser extent Liguria) display some very complex syntactic structures in the negative sentence that always differ from Tuscan (and standard Italian), as can be seen in the sequence “Questa donna + non + mi + piace.” By contrast, the Triveneto or Tre Venezie (Venezia Euganea, Venezia Giulia, and Venezia Tridentina), Tuscany and all of Umbria, Marche, Latium, Abruzzi, Campania, Basilicata, Apulia, Calabria, Sicily, and Sardinia all repeat the structure of Tuscan with a great compactness, which shows a conservative disposition (though Latin would have preferred a sequence of pronoun, negative particle, and verb, rather than of negative particle, pronoun, and verb, as in modern Italian). This basic split is also demonstrated by the large number of negative adverbs, both new and old, which in the Northern dialects have been placed after the verb, as discussed above.

According to Clemente Merlo (1924), dialect differences in Italy are due to the varying effects that pre-existing languages had on Latin while it was expanding throughout the peninsula. For this reason, the Tuscan dialects would be the result of Etruscans’ Latin language learning, the Northern dialects would show signs of a Celtic substratum, the Central-Southern dialects of an Italic (but non-Latin) substratum, those of Liguria that have a mixture of Celtic and Ligurian features a pre-Indo-European substratum, and those of Lunigiana, Corsica, and Sicily-Calabria-Salento still other, Mediterranean, substrata.

Later dialect classifications focused not on historical origins of the differences but on synchronic comparative analysis. For example, Rohlfs (1967), observing the recurring locations of some isoglosses and using geolinguistic methods, established the existence of a “La Spezia–Rimini line” and a “Rome–Ancona line.” To the north of the first line, he noticed features like degemination, palatalization, lenition, and sonorization, falling final unstressed vowels, and so on, which, as well any other internal subdivision, give the Northern dialect region a certain homogeneity and exclude all dialects below the line, that is to say those of Tuscany, Marche, Umbria, and Northern Latium. These dialects, in turn, are separated from those south of the Rome–Ancona line by other features. In the South, for example, we find: metaphony; vowel apocope; assimilation of /-nd-/ and /-mb-/ to /-nn-/ and /-mm-/; use of *stare* and *tenere* as auxiliaries instead of *essere* and *avere*; post-positioning of the possessive

adjective and merged forms of family nouns, such as *frateme* for *mio fratello*; and so on. Considering all of these factors, we can distinguish three main dialect groups, Northern, Central-Tuscan, and Southern, whose most important features are not significantly different from those established by Ascoli.

Nonetheless, it is important to remember that each of these major regional groupings comprises a large amount of variation among the dialects it contains. Within the South, for instance, we find several important internal divisions. “Outer Southern” dialects, which include Salento (the province of Lecce), Aspromonte in Calabria, and Sicily, all have a system of five stressed vowels, compared to the Tuscan system of seven, as well as cacuminal (or retroflex) phones. Cacuminal consonants (as well as unusual glottal consonants found in some other Italian dialects) are also present in Sardinian dialects. Together with several other characteristics (e.g., the conservative syntactic character already mentioned), the distinctness of the Sardinian phonological system has led some dialectologists to classify Sardinian dialects as a separate group.

Within the North, we may distinguish the dialects of Veneto. Even if northern dialects show a large number of common traits, as we have noticed above in commenting on negation patterns, there are many peculiarities that make the dialects of Veneto closer than other northern dialects to those of Tuscany. Phonetically, for example, we may cite a higher resistance of unstressed vowels to lenition and the absence of palatalization of the consonantal group -CT-, which becomes -t(t)- in the dialects of Veneto and Tuscany, whereas it becomes -jt- or -c- in the dialects of Piedmont, Lombardy, and Emilia. Latin NOCTEM, for example, becomes *næjt* in Turin, *noc* in Milan, and *noc* in Bologna. The dialects of Liguria, instead, are matched with those of Veneto and Tuscany: in Genova we find *næte*, in Venice *nòte*, and in Florence *nòtte* (cf. AIS, chart 342).

29.3.3 Classification: A Visual Summary

To give the reader a more concrete visual image of the complex classificatory categories that have been set forth in the preceding paragraphs, it seems useful to include here a map of Italy, in which the various dialect regions are represented by different types of hatching (Figure 29.3). This shows the Gallo-Italic region in the North; the Tuscan-Central (“Mediana”) region in the middle of the peninsula (extending down to Rome); the main Southern (“Meridionale”) region below that; and an extreme Southern region comprising the heel and toe of the “boot”, as well as Sicily.

Considering this map, it should be pointed out that the northern political borders of Italy do not coincide with linguistic boundaries. Within Italy we find Occitan groups and Franco-Provençal in the West; Romansh, German, and Ladin in the North; and Friuli and Slovenian in the East. Outside Italy we find the dialects of the Swiss Ticino region, classifiable under the Lombard type and belonging linguistically to Italy. It is also important to note the divisions within the map’s main regions. Beginning in the North, we find an important divide between the Gallo-Italic dialects of the Northwest (in Piedmont, Lombardy, Liguria, and Emilia-Romagna) and the Venetian dialects of the Northeast. Within the Central region, the map emphasizes the strongly autonomous character of Tuscan dialects, which are linked to those of Corsica, reflecting that island’s long political dependence on the Tuscan maritime republic of Pisa; these are together distinct from the other Central dialects of Umbria, Marche, and northern Lazio. Finally, in the Southern Area we find the dialects of southern Lazio and Abruzzo, Molise, Campania, Puglia, Lucania, and Northern Calabria. Although associated by vertical hatching with this southern group, the Sardinian dialects are generally considered by dialectologists to form a distinct group, especially including Campidanese, Logudorese, and Nuorese (in Gallura and Sassari we find more “continental” features).



Figure 29.3 The dialect groupings in Italy (Source: Adapted from Grassi, Sobrero, and Telmon 1997, p. 82).

REFERENCES

AIS – Jaberg, Karl, Jud Jakob. (1928–1940), *Sprach- und Sachatlas Italiens und der Südschweiz*, 8 vols., Zofingen, Ringier u. C. <http://www3.pd.istc.cnr.it/navigais-web>

ALBa – Del Puente, Patrizia. (2008 et ss.), *Atlante Linguistico della Basilicata*, Rionero in Vulture.
ALCam – Radtke, Edgar. n.d., *Atlante Linguistico della Campania*, www.alcam.de/alcamframeset.htm

- ALD – Goebel, Hans. (1985–1998), *Atlante Linguistico del Ladino Dolomitico*, Salzburg.
- ALD2 – Goebel, Hans. (1999–2013), *Atlante Linguistico del Ladino Dolomitico*, Part II, <http://ald.sbg.ac.at/ald-i/index.php>
- ALECal – Trumper, John and Maddalon, Marta. n.d., *Atlante Linguistico e Etnografico della Calabria*, Cosenza. <http://www.linguistica.unical.it/linguist/pubblicazioni/alecal.htm>
- ALEPO – Canobbio, Sabina and Telmon, Tullio. (2004 et ss.), *Atlante Linguistico ed Etnografico del Piemonte Occidentale*, 3 vol published until now, Ivrea and Alessandria.
- ALI – Bartoli, Matteo, and Alii. (1997 et ss.), *Atlante Linguistico Italiano*, 8 voll. published until now, Roma, Istituto Poligrafico and Zecca dello Stato. Redaction at the University of Turin, with the supervision of L. Massobrio.
- ALS – Ruffino, Giovanni and D'Agostino Mari. (1989 et ss.), *Atlante Linguistico di Sicilia*, Palermo.
- ALT – Giacomelli, Gabriella. 2000. *Atlante Lessicale Toscano*, Roma, Lexis. http://serverdbt.ilc.cnr.it/altweb/RT_ALT-WEB_home.htm
- Ascoli, Graziadio Isaia. 1873. *Saggi ladini*, «Archivio Glottologico Italiano» 1, pp. 1–556.
- Ascoli, Graziadio Isaia. 1880. *Italy. Language, «Encyclopaedia Britannica» XIII*, pp. 491–498.
- Ascoli, Graziadio Isaia. (1882–1885), *L'Italia dialettale*, «Archivio Glottologico Italiano» 8, pp. 98–128.
- ASiCa – Krefeld, Thomas. n.d. *Atlante Sintattico della Calabria*, <http://asicawgi.uni-muenchen.de>
- ASIt – Benincà, Paola. n.d. *Atlante Sintattico d'Italia*, <http://asit-cnr.unipd.it>
- ASLEF - Pellegrini, Giovan Battista. 1972–1986. *Atlante Storico-Linguistico-Etnografico del Friuli, Padova-Udine*.
- Bianchini, G., Bracchi, Remo. 2003. *Dizionario etimologico della Val Tartano*, Sondrio.
- DAM – Gianmarco, Ernesto. (1965–2003), *Dizionario dei dialetti dell'Abruzzo e del Molise*, 7 vol., Roma.
- De Mauro, Tullio. 1964. *Storia linguistica dell'Italia unita*, Bari, Laterza.
- Grassi, Corrado. 2009. *Dizionario del dialetto di Montagne di Trento, San Michele all'Adige*.
- Grassi, Corrado, Sobrero, Alberto A., and Telmon, Tullio. 1997. *Fondamenti di dialettologia italiana*, Roma - Bari, Laterza.
- Hall, Robert A. Jr. 1958. *Bibliografia della linguistica italiana*, 3 voll., Firenze, Sansoni.
- Hall, Robert A. Jr. 1969. *Bibliografia della linguistica italiana. Primo supplemento decennale*, Firenze, Sansoni.
- Hall, Robert A. Jr. 1980. *Bibliografia della linguistica italiana. Secondo supplemento decennale*, Pisa.
- Hall, Robert A. Jr. 1988. *Bibliografia della linguistica italiana. Terzo supplemento decennale*, Pisa.
- Jaberg, Karl and Jud, Jakob. 1928. *Der Sprachatlas als Forschungsinstrument. Kritische Grundlegung und Einführung in den Sprach- und Sachatlas Italiens und der Südschweiz*, Halle (Saale), Niemeyer (trad. it. AIS. *Atlante linguistico ed etnografico dell'Italia e della Svizzera meridionale*, Milano, Unicopli, 1987, 2 vol., vol. 1°, *Atlante linguistico come strumento di ricerca. Fondamenti critici e introduzione*, edited by G. Sanga).
- Ledgeway, Adam. 2009. *Grammatica diacronica del napoletano*. Tübingen: Max Niemeyer Verlag.
- LEI – Pfister, Max and Schweikard, Wolfgang. (1979 ss), *Lessico Etimologico Italiano*, Marburgo.
- Loporcaro, Michele. 1988. *Grammatica storica del Dialetto di Altamura*, Pisa.
- LSI – Lurà, Franco. 2004. *Lessico dialettale della Svizzera italiana*, 5 vols, Bellinzona.
- Marcato, Carla. 2007. *Dialetto, dialetti e italiano*, Bologna, il Mulino (1a ed. 2002).
- Merlo, Clemente. 1924. *L'Italia dialettale*, «L'Italia dialettale» 1, pp. 12–26.
- RID – Lurà, Franco. 2013. *Repertorio italiano-dialetti*, Bellinzona.
- Rohlfs, Gerhard. (1949–1953), *Historische Grammatik der Italienischen Sprache und ihrer Mundarten*, Bern, A. Francke, 3 vol. (trad. it. *Grammatica storica della lingua italiana e dei suoi dialetti*, Torino, Einaudi, 1966–1969, 3 vol.).
- Rohlfs, Gerhard. 1967. *L'Italia dialettale (dal Piemonte alla Sicilia)*, «Nuovi argomenti», 5, pp. 22–27.
- Toso, Fiorenzo. 2004. *Dizionario etimologico e storico Tabarchino*, Vol.1, Recco- Genova.
- VDSI- Lurà, Franco. (1952 ss), *Vocabolario dei dialetti della Svizzera italiana*, Bellinzona.
- VIVALDI – Bauer, Roland and Kattenbusch, Dieter. n.d. *Vivaio Acustico delle Lingue e dei Dialetti d'Italia*, Berlino, <http://www2.hu-berlin.de/vivaldi/>
- VS – Piccitto, Giorgio, Tropea, Giovanni, and Trovato, Salvatore (1977–2002), *Vocabolario Siciliano*, Palermo.

30 Dialects of Spanish and Portuguese

JOHN M. LIPSKI

30.1 Basic Facts

30.1.1 *Historical Development*

Spanish and Portuguese are closely related Ibero-Romance languages whose origins can be traced to the expansion of the Latin-speaking Roman Empire to the Iberian Peninsula; the divergence of Spanish and Portuguese began around the ninth century. Starting around 1500, both languages entered a period of global colonial expansion, giving rise to new varieties in the Americas and elsewhere. Sources for the development of Spanish and Portuguese include Lloyd (1987), Penny (2000, 2002), and Pharies (2007). Specific to Portuguese are features such as the retention of the seven-vowel system of Vulgar Latin, elision of intervocalic /l/ and /n/ and the creation of nasal vowels and diphthongs, the creation of a “personal” infinitive (inflected for person and number), and retention of future subjunctive and pluperfect indicative tenses. Spanish, essentially evolved from early Castilian and other western Ibero-Romance dialects, is characterized by loss of Latin word-initial /f/, the diphthongization of Latin tonic /ɛ/ and /ɔ/, palatalization of initial C+L clusters to /ʎ/, a complex series of changes to the sibilant consonants including devoicing and the shift of /ʃ/ to /χ/, and many innovations in the pronominal system.

30.1.2 *The Spanish Language Worldwide*

Reference grammars of Spanish include Bosque (1999a), Butt and Benjamin (2011), and Real Academia Española (2009–2011). The number of native or near-native Spanish speakers in the world is estimated to be around 500 million. In Europe, Spanish is the official language of Spain, a quasi-official language of Andorra and the main vernacular language of Gibraltar; it is also spoken in adjacent parts of Morocco and in Western Sahara, a former Spanish colony. In the Americas, Spanish is the official language throughout South America except Brazil, Suriname, French Guiana, and Guyana; in the Caribbean nations of Cuba, Puerto Rico, and the Dominican Republic; and in Mexico and all of Central America except for Belize. Unofficially, it is widely used in Belize, Haiti, Aruba, the US Virgin Islands, and in the United States, where the nearly 45 million speakers make the US a strong contender for second place among the world’s Spanish-speaking countries; Canada is home to nearly half a million Spanish speakers. Spanish is also residually present in the Philippines and the Mariana Islands. Spanish is the third most widely used language on the Internet (after English and Chinese).

30.1.3 The Portuguese Language Worldwide

Reference grammars of Portuguese include Cunha and Cintra (1984), Perini (2002), and Thomas (1974). Worldwide, the number of Portuguese speakers is estimated to be between 215 and 250 million, with the higher number including second-language speakers. Beyond Portugal, Portuguese is the official language of Brazil, Cape Verde, Guinea-Bissau, São Tomé and Príncipe, Angola, Mozambique, and East Timor. It still has some vitality in Goa (India) and is spoken natively in much of northern Uruguay and in the northeastern Argentine province of Misiones. Portuguese also has co-official status in Macau and Equatorial Guinea.

30.1.4 Mutual Intelligibility

At the level of educated speech, all varieties of Spanish are highly mutually intelligible. At the colloquial or vernacular level, differences in pronunciation and vocabulary and to a lesser extent morphosyntax result in considerable divergence, and although such differences are easily resolved in face-to-face encounters, they may represent obstacles to comprehension in passive listening situations. Differences between the Portuguese varieties of Portugal (extending to Lusophone Africa and Asia) and those of Brazil are often considerable, and Brazilians in particular frequently experience difficulty in understanding spoken European Portuguese, although easily comprehending the written language.

In their written forms, Spanish and Portuguese share a high degree of mutual intelligibility, once simple transpositions are mastered. In areas along the Spanish-Portuguese border and along the borders between Brazil and Spanish-speaking nations, mutual intelligibility of spoken Spanish and Portuguese is facilitated by familiarity, but regional and social varieties of the two languages often diverge to the point of limited mutual comprehension away from border regions. Although Spanish and Portuguese were once end points of a dialect continuum containing intermediate varieties such as Leonese and Extremeno, only (recently standardized) Galician enjoys contemporary vitality within the continuum.

30.2 Spanish: Main Sources

30.2.1 General

Moreno Fernández (2005) surveys recent corpora of oral Spanish. The most extensive multi-genre resource is the *Corpus del Español* searchable database (www.corpusdelespanol.org). Another research tool is the *Corpus de Referencia del Español Actual* (CREA) of the Real Academia Española (corpus.rae.es).

The first major attempt to consolidate transcriptions of spoken Spanish (and Portuguese) was the *Estudio Coordinado de la Norma Lingüística Culta de las Principales Ciudades de Iberoamérica y de la Península Ibérica*, known as the *Norma Culta* project (PILEI 1971–1973). Several of the transcriptions have been consolidated on a CD-ROM collection (Samper Padilla *et al.* 1998). The original recordings, mostly made with reel-to-reel tape recorders, have never been centrally archived or digitized; although copies of some of the recordings can be readily located, in most cases the published transcriptions cannot be compared with the original recordings.

30.2.2 European Spanish

Sources of information about dialect variation in Spain include Moreno Fernández (2009), Zamora Vicente (1967), and the articles in Alvar (1996a). Numerous traditional monographs have also been published on the traditional speech of single villages or rural sectors, usually

entitled, *El habla de* “the speech of ...”. Gibraltar Spanish is covered by Kramer (1986), Levey (2008), and Lipski (1986).

In the atlas tradition, Spain is represented by the *Atlas lingüístico de la Península Ibérica* (ALPI, 1962). A digital version is now freely available to scholars (www.alpi.ca; see also Heap 2008). Regional atlases include Alvar (1978) for the Canary Islands; Alvar (1995) for Cantabria; Alvar *et al.* (1973) for Andalucía; Alvar *et al.* (1979-1983) for Aragón, Navarra, and Rioja; and García Mouton and Moreno Fernández (n.d.) for Castilla-La Mancha. Norma Culta transcriptions for Spanish cities include Esqueva and Cantarero (1981) for Madrid; Fernández Juncal (2005) for Salamanca; Gómez Molina and Albelda (2001) for Valencia; Pineda (1985) for Seville; and Salvador Salvador and Águila Escobar (2006) for Granada. Recent corpora of transcribed interviews include García Marcos (2000) for Almería; and Lasarte Cervantes (2008) and Vida Castro (2007) for Málaga.

30.2.3 Latin American Spanish

Basic references on Spanish in Latin America include Cotton and Sharp (1988), Lipski (1994, 1996), Moreno de Alba (1988), Noll (2009), Zamora and Guitart (1988), and Alvar (1996b). Latin American pronunciation is surveyed in Canfield (1962, 1981), and Spanish varieties in the United States are described in Lipski (2008a).

Linguistic atlases are available for Colombia (Instituto Caro y Cuervo 1981), Costa Rica (Quesada Pacheco 2010), Mexico (Lope Blanch 1990), Nicaragua (Chavarría Ubeda 2010), and New Mexico and southern Colorado (Bills and Vigil 2008). Norma Culta transcriptions for Spanish American cities include Caravedo (1989) for Lima; Gutiérrez Marrone (1992) for La Paz; Heras Poncela (1999) for Guadalajara; Lope Blanch (1971, 1979) for Mexico City; Martorell de Laconi, Hondrogianis, and Soto (2000) for Salta, Argentina; Morales and Vaquera de Ramírez (1990) for San Juan, Puerto Rico; Rabanales and Contreras (1979) for Santiago; Rodríguez Cadena (2009) for Barranquilla, Colombia; Rosenblat and Bentivoglio (1979) for Caracas; and Universidad Nacional de Buenos Aires (1987) for Buenos Aires.

30.2.4 Spanish in Africa and Asia

For the Spanish of Equatorial Guinea, see Lipski (1985, 2000) and Quilis and Casado-Fresnillo (1995); for that of the Philippines, see Lipski (1987) and Quilis and Casado-Fresnillo (2008); for Western Sahara, see Tarkki (1995); for Morocco, see Sayahi (2004, 2005, 2006).

30.3 Portuguese: Main Sources

30.3.1 General

Major corpora of written Portuguese include the *Corpus do português* (www.corpusdoportugues.org); the Reference Corpus of Contemporary Portuguese (CRPC) (www.clul.ul.pt); the Tycho Brahe project, a searchable database of historical Portuguese texts (www.tycho.iel.unicamp.br); and the Colonia Corpus of Historical Portuguese (corporavm.uni-koeln.de/colonia/).

30.3.2 European Portuguese

In Portugal, the *Atlas Linguístico-Etnográfico de Portugal e da Galiza* (ALEPG) began collecting materials in the 1970s, and representative data are found in Ferreira *et al.* (2008). The Azores are represented by the *Atlas Linguístico-Etnográfico dos Açores* (www.culturacores.azores.gov.pt/alea/).

30.3.3 Brazilian Portuguese

Brazilian Portuguese is represented in the Lácio-web corpus (www.nilc.icmc.usp.br/lacioweb). A DVD+book with samples of spoken Brazilian Portuguese can be obtained at www.c-oral-brasil.org. A corpus of written Brazilian Portuguese is the *Banco de Português* (www2.lael.pucsp.br/corpora/bp/). The status of the *Atlas linguístico do Brasil* project can be found at twiki.ufba.br/twiki/bin/view/Alib/. The Brazilian state of Mato Grosso do Sul is covered by Oliveira (2007) and southern Brazil by Koch, Silfredo and Altenhofen (2002). Norma Culta transcriptions for Brazilian cities include Calloou and Lopes (1991) for Rio de Janeiro; Castilho, Preti and Urbano (1986–1990) for São Paulo, Hilgert (1997–2009) for Porto Alegre, Mota and Rollemburg (1994) for Salvador; and Sá (2005) for Recife.

30.3.4 Portuguese in Africa and Asia

Sources for Lusophone Africa include Gonçalves (1996) and Jon-And (2011) for Mozambique; Inverno (2009a, 2009b) for Angola; Jon-And (2011) for Cape Verde; and Figueiredo (2010) for São Tomé. A searchable database of written African Portuguese is available at www.clul.ul.pt. Portuguese in Goa is described by Wherritt (1985) and that in Macau by Baxter (2009). Several sources on East Timor Portuguese can be downloaded from profesdeptemtl.wix.com/lingua-portuguesa-timor-leste#!.

30.4 Spanish Dialect Zones and Characteristics

30.4.1 Spain

The principal division of Iberian Spanish dialects is north-south, with the northern dialects of Castile (including the capital, Madrid), León, Cantabria, the Basque Country, Aragon, and Spanish-speaking areas of Catalunya differentiated as a group from the southern varieties of Extremadura, Andalusia, and Murcia. The Canary Islands constitute a separate dialect cluster, most closely related to western Andalusian Spanish. The Spanish of Gibraltar is essentially that of the neighboring Spanish province of Cádiz, modified by contact with English. Modern standard European Spanish, based mostly on the Castilian dialect and therefore often called *Castellano*, is promulgated by the Real Academia Española.

30.4.1.1 Phonetics and Phonology

Most of Peninsular Spain, except for the southwestern provinces and part of the Valencia/Alicante region, distinguishes the sibilant phonemes /θ/ and /s/, for example, *caza* [‘ka.θa] “hunting” versus *casa* [‘ka.sa] “house.” In regions where this contrast is neutralized, [s] predominates in urban areas and [θ] in surrounding provincial regions. In northern Spain, /s/ receives an apico-alveolar realization [ʃ]. In most of Spain, the opposition /k/ - /χ/ has been neutralized in favor of /χ/, but /k/ is retained in parts of northern Spain and sporadically elsewhere in the Peninsula as well as in much of the Canary Islands. Word-final /n/ is velarized to [ŋ] in the northwest and in Extremadura, Andalusia, and the Canary Islands. In the southern regions, syllable- and word-final /s/ is aspirated to [h] or elided. In eastern Andalusia, loss of word-final /s/ is reflected in laxing of the preceding vowel. The singular-plural distinction is therefore maintained through vowel quality, as in *perro* [‘pe.ro] “dog” versus *perros* [‘pe.ɾɔ] “dogs.” In southern Spain and the Canary Islands, word-final /l/ and /r/ are routinely elided, whereas in preconsonantal contexts the two liquids are frequently neutralized, most often in favor of [ɾ].

30.4.1.2 Morphosyntax

Most of Peninsular Spain employs *le* and *les* as masculine direct object clitics; the etymological *lo* and *los* predominate in the southwest and the Canary Islands. There is an increasing tendency in Peninsular Spain to employ the present perfect tense rather than the simple pretérito in contextual frames that do not include the moment of speaking, for example, *mi amigo ha llegado ayer*, lit. "my friend has come yesterday."

30.4.2 Latin America

The most robust dialect classification of Latin American Spanish includes the following major regions: Mexico and Guatemala; Honduras, El Salvador and Nicaragua; Costa Rica; the Caribbean basin (Cuba, Puerto Rico, Dominican Republic, Venezuela, northern Colombia, and Panama); the interior of Colombia; the Pacific coast of Colombia, Ecuador, and Peru; the highlands of Ecuador, Peru, Bolivia, and northwestern Argentina; Chile; Paraguay, eastern Bolivia and northeastern Argentina; central and southern Argentina and Uruguay. Most varieties of Spanish in the United States are directly related to those of immigrant populations' home countries, principally Mexico, Puerto Rico, Cuba, the Dominican Republic, El Salvador and Guatemala, and Colombia (Lipski 2008a, Otheguy and Zentella 2011), but a small pocket of descendants of Canary Islanders is still found in southeastern Louisiana (Coles 1999, Lipski 1990).

30.4.2.1 Phonetics and Phonology

Unlike most of Spain, all varieties of American Spanish have a merger of /θ/ and /s/, realized as [s]. The principal phonological variable in Latin American Spanish is the opposition /χ/ - /j/ (e.g., *se calló* "he/she stopped talking" versus *se cayó* "he/she fell down." This opposition is maintained in Paraguay, Bolivia, northeastern Argentina, a few areas in central Colombia, and most of highland Peru and Ecuador (in parts of Ecuador /χ/ is realized not as a lateral but as [ʒ]); elsewhere these sounds are merged as /j/.

The major phonetic variables of Latin American Spanish involve the realization of /x/, /j/, and the trill /r/; and syllable- and word-final /s/, /n/, /l/, and /r/. The posterior fricative /x/ is a weakly aspirated [h] in the Caribbean zone and the Pacific coast of Colombia, Ecuador, and Peru, whereas in Chile /x/ is a palatal fricative [ç] before front vowels, sometimes followed by a glide [j], for example, *gente* "people." Intervocalic /j/ is weak and often elided in contact with front vowels in coastal Peru, Central America, northern Mexico, and New Mexico and Colorado. In most of Argentina and Uruguay, /j/ has traditionally been realized as a voiced fricative, [ʒ], but devoicing to [ʃ] has extended from Buenos Aires and Montevideo to much of the remaining territory. The nominally trilled /r/ receives a fricative realization, [ʒ], throughout the Andean highlands and much of northern Argentina, and variably in Chile, Paraguay, Guatemala, and Costa Rica (in Costa Rica alternating with retroflex [ɻ]). Final /s/ is variably aspirated or elided throughout Latin America except in the Andean highlands and most of Mexico, Guatemala, and Costa Rica. Word-final /n/ is velarized to [ŋ] in the Caribbean, Central America, the Pacific coast of Colombia, Ecuador, and Peru, and variably in the Andean highlands. Syllable-final /l/ and /r/ are variably neutralized throughout the Caribbean dialect cluster and in parts of Chile.

30.4.2.2 Morphosyntax

Latin American differs from European Spanish in the absence of the informal second-person plural pronoun, *vosotros*, and the corresponding object clitic, *os*; *ustedes* is the sole second-person plural pronoun. The second-person singular familiar pronoun is *vos* (instead of *tú*) in all of Argentina, Paraguay, Bolivia, most of Uruguay and Chile, certain regions of Colombia,

Ecuador, and Venezuela, and vestigially in Peru. *Vos* predominates in all of Central America, as well as in western Panama, parts of the Mexican state of Chiapas and small pockets of Cuba. The verb forms that accompany *vos* vary widely, ranging from reflexes of Peninsular Spanish *vosotros* conjugations to the inflexions corresponding to *tú* (Páez Urdaneta 1981).

Most Latin American varieties of Spanish employ the masculine direct object clitics *lo* and *los*; use of *le* and *les* for direct objects is characteristic of Paraguay and Ecuador and variably of Mexico. In the Southern Cone, direct object clitics can accompany animate direct object nouns (e.g., *Lo conozco a Juan* “I know John”) and in the Andean highlands clitic doubling occurs freely with all direct objects (e.g., *Lo veo el carro* “I see the car”). In much of South America the past subjunctive in subordinate clauses is replaced by present subjunctive, for example, *El profesor me aconsejó que estudie [not estudiara] mucho* (“The teacher advised me to study a lot”). In the Caribbean dialect cluster and sporadically elsewhere, infinitives with overt subjects are used in preference to subjunctive forms: *La fiesta empezó antes de yo llegar* (“The party began before I arrived”) [not *antes que yo llegara*]. Also found throughout the Caribbean is the non-inversion of subject pronouns and verbs in questions, for example, *Cómo tú te llamas?* (“What is your name?”), or, *¿Dónde nosotros podemos comer?* (“Where can we eat?”). Found in much of Colombia, Venezuela, Ecuador, and the Dominican Republic (and also in Brazilian Portuguese) is the affirmative use of intrusive *ser* (“to be”): *Lo conocimos fue en la fiesta* (“(where) we met him was at the party”); *Tenemos es que trabajar mucho* (“We have to work a lot”).

30.5 Portuguese Dialect Zones and Characteristics

Azevedo (2005) and Cintra (1995) provide general information on Portuguese dialect variation.

30.5.1 Portugal, Africa, and Asia

The principal dialect division in Portugal is north versus south, with an approximate transition to the north of Coimbra. Lisbon is located in the southern zone and is the model for standard European Portuguese. African and Asian dialects generally follow a southern model, but since most are spoken in contact with other languages and often as a second language, there is considerable within-country variability. Audio samples of the principal varieties of Portuguese are available at cvc.instituto-camoes.pt.

30.5.1.1 Phonetics and Phonology

The northern varieties retain /ʃ/ (e.g., *chave* “key”) in opposition to /ʃ/ (e.g., *caixa* “box”), a distinction neutralized as [ʃ] elsewhere. The diphthongs *ei* [ej] and *ou* [ow] are retained in the north but reduced to simple vowels in central and southern Portugal. In greater Lisbon, *ei* is realized as [ej]. Atonic /a/ is realized as [ɐ], as is /a/ before nasals; atonic /o/ is raised to [u], and atonic /i/ and /e/ are generally neutralized to a high unrounded vowel [w] and sometimes devoiced or elided. Northern Portuguese has lost the /b/ - /v/ distinction. The treatment of sibilants also varies by region. In an area that includes much of Tras-os-Montes, Alto Minho, and Beira-Alta, a four-way distinction is maintained, for example, /s/ (*cego* “blind”), /z/ (*fazer* “do, make”), /ʃ/ (*senhor* “sir,” *passo* “step”), and /ʒ/ (*coisa* “thing”). In parts of Minho, Douro, Beira Alta, and Beira Baixa, the two sibilant points of articulation have been neutralized and only apico-alveolar sibilants /ʃ/ and /ʒ/ are found, whereas in the remainder of the country only alveolar-dental sibilants /s/ and /z/ are used (Azevedo 2005: 186; Cintra 1995: 28). In most of Portugal, syllable- and word-final /s/ and /z/ are palatalized to [ʃ] and [ʒ], respectively. The dialects of the Azores and Madeira Islands generally follow southern Portuguese patterns. In Madeira, /l/ is palatalized after [i] and [j] as

in *vila* “village,” stressed /a/ is often realized as [ɔ] (e.g., *casa* “house”) and the vowel /u/ is fronted to [y] (e.g., *tudo* “all”).

30.5.1.2 Morphosyntax

European and African Portuguese have generally substituted the infinitive for the etymological gerund in progressive constructions: *estou a trabalhar* “I am working” instead of *estou trabalhando*.

30.5.2 Brazil

Dietrich and Noll (2004) and Ilari and Basso (2006) describe dialect variation in Brazil. Brazilian linguists generally distinguish the following general dialect zones: the Northeast; Bahia; Mineiro (centered on Minas Gerais state); Fluminense/Carioca (centered on Rio de Janeiro); Paulista (centered on São Paulo); southern or Gaúcho; and the Amazonian region. No one city is considered the model for standard Brazilian Portuguese, but the educated speech of São Paulo and Rio de Janeiro comes the closest, although the palatalization of syllable-final /s/ in Rio is not emulated by other speakers. Brazilian Portuguese in general differs systematically from European varieties in several ways.

30.5.2.1 Phonetics and Phonology

Atonic /e/ and /o/ are raised in Brazil to [i] and [u], respectively. Prevocalic /b/, /d/, and /g/ are always occlusive, not fricative or approximant as in Europe, and /t/ and /d/ become pre-palatal affricates before /i/ (*tio*>[tʃi̯w] “uncle,” *dia*>[dʒia̯] “day”) except in the southernmost states. A velar fricative [χ] replaces the trill [r] both in intervocalic position (*carro* “car”) and syllable-finally (*por favor* “please”), although in word-final position /r/ is frequently elided. Throughout Brazil, syllable-final /l/ is vocalized to [w] thus *Brasil* [bra̯.'ziw] as opposed to the velarized [t] found in European varieties. Most Brazilians insert a post-vocalic glide [j] in stressed syllables ending in /s/, thus *nós* [nɔjs] (“we”), or *faz* [fais] (“do,” 3s), and an epenthetic vowel after borrowed words ending in consonants other than /r/, /l/ or /s/, thus *Nova York(i)*; the same occurs word-internally after syllable-final stop consonants, for example, *ad[i]vogado* (“lawyer”). Intervocalic /ɲ/ is pronounced as a nasal glide [j], for example, *senhor* (“sir”). Among regional dialects, the Carioca (Rio de Janeiro) variety is noted for palatalizing syllable-final /s/ to [ʃ] ([ʒ]) before voiced segments). Palatalization also occurs in some northeastern dialects (e.g., Ceará), but only before dental and alveolar consonants. In rural São Paulo state, the *Caipira* dialect is noted for having a retroflex syllable-final [ɿ], for example, in *porta* (“door”). In casual speech, loss of the final /r/ in verbal infinitives is widespread, as is the delateralization of /ʎ/ to [j] (*mullher*>*muié[r]*, “woman”).

30.5.2.2 Morphosyntax

Spoken Brazilian Portuguese eschews object clitics in favor of full pronouns, for example, *vejo él(e)* (“I see him/it”; cf. European *vejo-o*). Progressive verb forms use *estar* + gerund, for example, *estou falando* (“I am talking”), rather than the European *estar* a + infinitive (*estou a falar*). In spoken Brazilian Portuguese, *a gente* (“the people”) is used pronominally in preference to *nós* “we,” and for the majority of the country (except for the southernmost states) *você* “you” is the familiar second-person pronoun instead of European *tu*; formal address is achieved with *o senhor/a senhora* (“sir, madam”). European Portuguese *vós* is not used; only *vocês* expresses second-person plural. Double negation (repeating *não* before and after the verb) is very common, especially when a negative response is being emphasized, and in casual speech the first instance of *não* often disappears, for example, (*não*) *tenho não* (“I don’t have”). *Sei não* alternates with *sei lá* (“I don’t know”).

In casual, vernacular speech, plural /-s/ is placed only on the first element of plural noun phrases, for example, *aquelas coisa(s) nova(s)* ("those new things"). Also found in much vernacular speech, though condemned by prescriptivists, is the use of the third-person singular verb form for all other person-number combinations except first-person singular, for example, *vocês foi lá* ("you (pl.) went there," cf. standard *foram*), or *nós trabalha na cidade* ("we work in the city," cf. standard *trabalhamos*). In many vernacular varieties, the final /-s/ of first-person plural verbs is not pronounced, and in verbs of the first conjugation (in *-ar*), the ending becomes *-emo*, for example, *nós trabalhemo/trabalhamo na cidade*.

30.6 Current Research Trends

A survey of recent trends in Spanish and Portuguese dialectology is given in Lipski (2008b; also 1989). Earlier taxonomic and rural-oriented studies have given way to approaches that focus on dynamic urban environments with particular emphasis on socio-phonetic variation. The *Proyecto para el Estudio Sociolingüístico del Español de España y de América* (PRESEEA; at presea.linguas.net) is a coordinated effort to produce and consolidate sociolinguistic corpora representative of regional and social variation throughout the Spanish-speaking world.

Dialect variation has also informed linguistic models. Regionally distributed syntactic phenomena such as null subjects, non-inverted questions, and double negation have resulted in expanded theories, for example, by Bosque (1999b), Camacho (2006), Duarte (1995), Kato (2000), Ordóñez and Treviño (1999), and Toribio (2000). Regionalized pragmatic features such as politeness strategies have entered the picture, for example, the studies in Placencia and García (2006). Phonological theory has also been applied to dialect variation (e.g., Hualde 1989, Lipski 1999, Morris 2000).

30.7 Future Research

Vital to both the Iberian Peninsula and Latin America is the updating of now outdated dialect atlas materials to reflect the increasingly urbanized linguistic ecology of the twenty-first century. In addition, vast areas of Latin America are lacking accurate descriptions altogether. At least two other promising directions for future research can be mentioned.

The first is the systematic study of regional and social patterns of intonation, particularly in spontaneous speech. A consistent framework for the study of Spanish and Portuguese intonation is now widely accepted (e.g., Beckman *et al.* 2002, Frota 2013, Sosa 1999, Truckenbrodt *et al.* 2008), and many sentence-types have been classified for both European and American varieties of Spanish and Portuguese. Comparative studies are beginning to move beyond idealized and laboratory-elicited patterns to search for the defining traits of entire dialect regions.

Another facet of Spanish and Portuguese dialect variation is stable bilingual contact as a source of dialect features. There are studies of Ibero-Romance varieties that have arisen from sustained dialect or language contact, for example, Mirandese (Quarteu and Frías Conde 2002) and Barranquenho (Alvar 1996; Clements 2009: Chap. 8) along the Portugal-Spain border and "Fronterizo" in northern Uruguay (Elizaincín *et al.* 1987, Elizaincín 1992). In many other cases, analyses of regional and social dialects are handled separately from accounts of "interference" or code-switching in bilingual contact environments, even in cases where hundreds of thousands of speakers are involved. Bilingualism is a fact of life for millions of speakers of Spanish and Portuguese, making the integration of the dialectal traits of bilingual speech communities into general accounts of Spanish and Portuguese dialect variation a crucial step.

REFERENCES

- Alvar, Manuel. 1978. *Atlas lingüístico y etnográfico de las Islas Canarias*. Las Palmas: Excmo. Cabildo Insular de Gran Canaria. 3 vol.
- Alvar, Manuel. 1995. *Atlas lingüístico y etnográfico de Cantabria*. Madrid: Arco Libros.
- Alvar, Manuel. 1996. Barranqueño. In Alvar 1996a (ed.), 259–262.
- Alvar, Manuel (ed.). 1996a. *Manual de dialectología hispánica: el español de España*. Barcelona: Ariel.
- Alvar, Manuel (ed.). 1996b. *Manual de dialectología hispánica: el español de América*. Barcelona: Ariel.
- Alvar, Manuel, Antonio Llorente, Tomás Buesa, and Elena Alvar. 1979–1983. *Atlas lingüístico y etnográfico de Aragón, Navarra y Rioja*. Zaragoza: Diputación Provincial de Zaragoza.
- Alvar, Manuel, Antonio Llorente, and Gregorio Salvador. 1973. *Atlas lingüístico y etnográfico de Andalucía*. Granada: Universidad de Granada and Consejo Superior de Investigación Científica.
- Azevedo, Milton. 2005. *Portuguese: a linguistic introduction*. Cambridge: Cambridge University Press.
- Baxter, Alan. 2009. O português em Macau: contato e assimilação. In Carvalho (ed.), 277–312.
- Beckman, Mary, Manuel Díaz-Campos, Julia Tevis McGory, and Terrell Morgan. 2002. Intonation across Spanish, in the Tones and Break indices framework. *Probus* 14.9–36.
- Bills, Garland and Neddy Vigil. 2008. *The Spanish language of New Mexico and southern Colorado: a linguistic atlas*. Albuquerque: University of New Mexico Press.
- Bosque, Ignacio. 1999a. *Gramática descriptiva de la lengua española*. Madrid: Espasa.
- Bosque, Ignacio. 1999b. On focus vs. WH-movement: the case of Caribbean Spanish. *Sophia Linguistica* 44–45.1–32.
- Butt, John and Carmen Benjamin. 2011. *A new reference grammar of modern Spanish*. Oxford: Oxford University Press. 5th ed.
- Callou, Dinah Maria Isensee and Célia Regina Lopes. 1991. *A linguagem falada culta na Cidade do Rio de Janeiro: matérias para seu estudo*. Rio de Janeiro: UFRJ/Faculdade de Letras.
- Camacho, José. 2006. In situ focus in Caribbean Spanish: towards a unified account of focus. *Selected proceedings of the 9th Hispanic Linguistics Symposium*, ed. Nuria Sagarra, Almeida Jacqueline Toribio, 1–23. Somerville, MA: Cascadilla.
- Canfield, D. Lincoln. 1962. *La pronunciación del español en América*. Bogotá: Instituto Caro y Cuervo.
- Canfield, D. Lincoln. 1981. *Spanish pronunciation in the Americas*. Chicago: University of Chicago Press.
- Caravedo, Rocío. 1989. *El español de Lima: materiales para el estudio del habla culta*. Lima: Pontificia Universidad Católica del Perú, Fondo Editorial.
- Carvalho, Ana (ed.). 2009. *Português em contato*. Frankfurt and Madrid: Vervuert/Iberoamericana.
- Castilho, Ataliba Teixeira de, Dino Preti, and Hudinilson Urbano. 1986–1990. *A linguagem falada culta na cidade de São Paulo: matérias para seu estudo*. São Paulo: T. A. Queiroz.
- Chavarría Ubeda, Carmen. 2010. *Atlas lingüístico etnográfico de Nicaragua*. Bergen, Norway: University of Bergen.
- Cintra, Luis Felipe Lindley. 1995. *Estudos de dialectologia portuguesa*. Lisbon: Livraria Sá da Costa Editora.
- Clements, J. Clancy. 2009. *The linguistic legacy of Spanish and Portuguese: colonial expansion and language change*. Cambridge: Cambridge University Press.
- Coles, Felice. 1999. *Isleño Spanish*. Munich: LINCOM Europa.
- Consejo Superior de Investigación Científica. 1962. *Atlas lingüístico de la Península Ibérica*. Madrid: Consejo Superior de Investigación Científica.
- Cotton, Eleanor and John Sharp. 1988. *Spanish in the Americas*. Washington: Georgetown University Press.
- Cunha, Celso Ferreira da and Cintra, Luis Lindley. 1984. *Nova gramática do português contemporâneo*. Lisbon: Edições J. Sá da Costa.
- Dietrich, Wolf and Volker Noll. 2004. *O português do Brasil: perspectivas da pesquisa atual*. Frankfurt and Madrid: Vervuert/Iberoamericana.
- Duarte, Maria Eugênia. 1995. A perda do princípio Evite Pronome no português Brasileiro. Doctoral dissertation, Universidade Estadual de Campinas.
- Elizaincín, Adolfo. 1992. *Dialectos en contacto: español y portugués en España y América*. Montevideo: Arca.
- Elizaincín, Adolfo, Luis Behares, and Graciela Barrios. 1987. *Nos falemo brasílero*. Montevideo: Editorial Amesur.

- Esqueva, Manuel and Margarita Cantarero. 1981. *El habla de la ciudad de Madrid: materiales para su estudio*. Madrid: Consejo Superior de Investigaciones Científicas Instituto "Miguel de Cervantes."
- Fernández Juncal, Carmen. 2005. *Corpus de habla culta de Salamanca*. Burgos: Instituto Castellano y Leonés de la Lengua.
- Ferreira, Manuela Barros, José Saramago, Luisa Segura, Gabriela Vitorino, Ernestina Carrilho, and Maria Lobo. 2008. *Atlas linguístico-ethnográfico de Portugal e da Galiza: a criação de gado-1*. Lisbon: Centro de Linguística da Universidade de Lisboa/Imprensa Nacional-Casa da Moeda.
- Figueiredo, Carlos Filipe Guimarães. 2010. A concordância plural variável no sintagma nominal do português reestruturado da comunidade de Almoxarife, São Tomé. Doctoral dissertation, University of Macau.
- Frota, Sonia. 2013. *Prosody and focus in European Portuguese: phonological phrasing and intonation*. New York: Routledge.
- García Marcos, Francisco Joaquín. 2000. *Corpus sociolinguístico del español en Almería*. Almería: Servicio de Publicaciones, Universidad de Almería.
- García Moutón, Pilar, and Francisco Moreno Fernández. Atlas lingüístico (y etnográfico) de Castilla-La Mancha. www2.uah.es/alecman/
- Gómez Molina, Ramón, and Marta Albelda. 2001. *El español hablado de Valencia: materiales para su estudio*. Valencia: Universitat de València, Facultad de Filología, Departamento de Filología Española.
- Gonçalves, Perpétua. 1996. *Português de Moçambique: uma variedade em formação*. Maputo: Livraria Universitária.
- Gutiérrez Marrone, Nila. 1992. *El habla de la ciudad de La Paz: materiales para su estudio*. La Paz: Ediciones Signo.
- Heap, David. 2008. The linguistic Atlas of the Iberian Peninsula (ALPI): a geolinguistic treasure 'lost' and found. *Toronto Working Papers in Linguistics* 27.87–96.
- Heras Poncela, María del Rosario. 1999. *El habla culta de la zona metropolitana de Guadalajara*. Guadalajara: Universidad de Guadalajara, Centro Universitario de Ciencias Sociales y Humanidades.
- Hilgert, José Gaston. 1997-2009. *A linguagem falada culta na cidade de Porto Alegre: materiais para o seu estudo*. Florianópolis: Editora Insular/Porto Alegre: Universidad Federal do Rio Grande do Sul.
- Hualde, José Ignacio. 1989. Delinking processes in Romance. *Studies in Romance linguistics*, ed. Carl Kirschner and Janet DeCesaris, 177–193. Amsterdam: John Benjamins.
- Ilari, Rodolfo and Basso, Renato. *O português da gente: a língua que estudamos : a língua que falamos*. São Paulo: Editora Contexto.
- Instituto Caro y Cuervo. 1981. *Atlas lingüístico-ethnográfico de Colombia*. Bogotá: Instituto Caro y Cuervo, 6 vol.
- Inverno, Liliana Cristina Coragem. 2009a. Contact-induced restructuring of Portuguese morphosyntax in interior Angola. Doctoral dissertation, University of Coimbra.
- Inverno, Liliana Cristina Coragem. 2009b. A transição de Angola para o português vernáculo: estudo morfossintático do sintagma nominal. In Carvalho (ed.), 87–106.
- Jon-And, Anna. 2011. Variação, contato e mudança linguística em Moçambique e Cabo Verde: a concordância variável de número em sintagmas nominais do português. Doctoral dissertation, Stockholm University.
- Kato, Mary. 2000. The partial pro-drop nature and the restricted VS order in Brazilian Portuguese. *Brazilian Portuguese and the null subject parameter*, ed. Mary Kato and Esmeralda Negrão, 207–240. Frankfurt and Madrid: Vervuert-Iberoamericana.
- Koch, Walter, Mário Silfredo, and Cléo Vilson Altenhofen (eds.). 2002. *Atlas linguístico-ethnográfico da região sul do Brasil (Alers)*. Porto Alegre: Editorial da UFRGS, 2 vols.
- Kramer, Johannes. 1986. *English and Spanish in Gibraltar*. Hamburg: H. Buske.
- Lasarte Cervantes, María de la Cruz. 2008. *El español hablado en Málaga III: corpus oral para su estudio sociolinguístico: nivel de estudios superiores*. Málaga: Editorial Sarriá.
- Levey, David. 2008. *Language change and variation in Gibraltar*. Amsterdam and Philadelphia: John Benjamins.
- Lipski, John. 1985. *The Spanish of Equatorial Guinea*. Tübingen: Max Niemeyer.
- Lipski, John. 1986. Sobre el bilingüismo anglo-hispánico en Gibraltar. *Neuphilologische Mitteilungen* 87.414–427.
- Lipski, John. 1987. Contemporary Philippine Spanish: comments on vestigial usage. *Philippine Journal of Linguistics* 18.37–48.
- Lipski, John. 1989. Beyond the isogloss: trends in Hispanic dialectology. *Hispania* 72.801–809.
- Lipski, John. 1990. *The language of the isleños: vestigial Spanish in Louisiana*. Baton Rouge: Louisiana State University Press.

- Lipski, John. 1994. *Latin American Spanish*. London: Longman.
- Lipski, John. 1996. *El español de América*. Madrid: Cátedra.
- Lipski, John. 1999. The many faces of Spanish /s/-weakening: (re)alignment and ambisyllabicity. *Advances in Hispanic linguistics*, ed. Javier Gutiérrez-Rexach and Fernando Martínez-Gil, 198–213. Somerville, MA: Cascadilla Press.
- Lipski, John. 2000. The Spanish of Equatorial Guinea: research on la hispanidad's best-kept secret. *Afro-Hispanic Review* 19.11–38.
- Lipski, John. 2008a. *Varieties of Spanish in the United States*. Washington, DC: Georgetown University Press.
- Lipski, John. 2008b. Homeless in post-modern linguistics? (re/dis)placing Hispanic dialectology. *Studies in Hispanic and Luso-Brazilian Linguistics* 1.211–221.
- Lloyd, Paul. 1987. *From Latin to Spanish*. Philadelphia: American Philosophical Society.
- Lope Blanch, Juan. 1971. *El habla de la Ciudad de México*. Mexico City: Universidad Nacional Autónoma de México.
- Lope Blanch, Juan. 1979. *El habla popular de la Ciudad de México*. Mexico City: Universidad Nacional Autónoma de México.
- Lope Blanch, Juan. 1990. *Atlas lingüístico de México*. Mexico City: Colegio de México/Fondo de Cultura Económica. 4 vol.
- Martorell de Laconi, Susana, María Hondrogiannis, and Edda Beatriz Soto. 2000. *Habla culta de la ciudad de Salta: materiales para su estudio*. Salta: Instituto Salteño de Investigaciones Dialectológicas "Berta Vidal de Battini."
- Morales, Amparo and María Vaquero de Ramírez. 1990. *El habla culta de San Juan*. Rio Piedras: Editorial de la Universidad de Puerto Rico.
- Moreno de Alba, José. 1988. *El español en América*. Mexico City: Fondo de Cultura Económica.
- Moreno Fernández, Francisco. 2005. *Corpora of spoken Spanish language – the representativeness issue. Linguistic informatics – state of the art and the future*, ed. Yuji Kawaguchi, Susumu Zaima, Toshihiro Takagaki, Kohji Shibano, and Mayumi Usami, 120–144. Amsterdam and Philadelphia: John Benjamins.
- Moreno Fernández, Francisco. 2009. *La lengua española en su geografía*. Madrid: Arcos.
- Morris, Richard. 2000. Constraint interaction in Spanish /s/-aspiration: three Peninsular varieties. *Hispanic linguistics at the turn of the millennium: papers from the 3rd Hispanic Linguistics Symposium*, ed. Héctor Campos, Elena Herburger, Alfonso Morales-Front, and Thomas J. Walsh, 14–30. Somerville, MA: Cascadilla.
- Mota, Jacyra and Vera Rollemburg. 1994. *A linguagem falada culta na cidade de Salvador: materiais para seu estudo*. Salvador: Universidade Federal da Bahia, Instituto de Letras.
- Noll, Volker. 2009. *Das amerikanische Spanisch: ein regionaler und historischer Überblick*. Tübingen: Niemeyer.
- Oliveira, Dercir Pedro de (ed.). 2007. *Atlas linguístico de Mato Grosso do Sul (ALMS)*. Campo Grande, MS: Editora UFMS.
- Ordóñez, Francisco and Esthela Treviño. 1999. Left-dislocated subjects and the *pro-drop* parameter: a case study of Spanish. *Lingua* 107.39–68.
- Orthegui, Ricardo and Ana Celia Zentella. 2011. *Spanish in New York: language contact, dialectal leveling, and structural continuity*. Oxford and New York: Oxford University Press.
- Páez Urdaneta, Iraset. 1981. *Historia y geografía hispanoamericana del voseo*. Caracas: Casa de Bello.
- Penny, Ralph. 2000. *Variation and change in Spanish*. Cambridge: Cambridge University Press.
- Penny, Ralph. 2002. *A history of the Spanish language*. Cambridge: Cambridge University Press.
- Perini, Mário. 2002. *Modern Portuguese: a reference grammar*. New Haven, CT: Yale University Press.
- Pharies, David. 2007. *A brief history of the Spanish language*. Chicago: University of Chicago Press.
- Pineda, Miguel Angel (ed.). 1985. *Sociolingüística andaluza 2: Material de encuestas del habla urbana culta de Sevilla*. Seville: Universidad de Sevilla.
- Placencia, María Elena and Carmen García (eds.). 2006. *Research on politeness in the Spanish-speaking world*. Mahwah, NJ: Routledge.
- Programa Interamericano de Lingüística y Enseñanza de Idiomas (PILEI). 1971–73. *Cuestionario para el estudio coordinado de la norma lingüística culta de las principales ciudades de Iberoamérica y de la Península Ibérica*. Madrid: Departamento de Geografía Lingüística, Universidad Complutense.
- Quarteu, Reis and Xavier Frías Conde. 2002. L'mirandés: ūa lhéngua minoritaria en Pertual. *IANUA* 2.89–105.
- Quesada Pacheco, Miguel. 2010. *Atlas lingüístico-etnográfico de Costa Rica*. San José: Editorial UCR.
- Quilis, Antonio y Celia Casado-Fresnillo. 1995. *La lengua española en Guinea Ecuatorial*. Madrid: Universidad Nacional de Educación a Distancia.

- Quilis, Antonio y Celia Casado-Fresnillo. 2008. *La lengua española en Filipinas*. Madrid: Consejo Superior de Investigaciones Científicas.
- Rabanales, Ambrosio and Lidia Contreras. 1979. *El habla culta de Santiago de Chile: materiales para su estudio*. Santiago de Chile: Universidad de Chile, Facultad de Filosofía y Letras, Departamento de Lingüística y Filología.
- Real Academia Española. 2009–2011. *Nueva gramática de la lengua española*. Madrid: Real Academia Española. www.rae.es
- Rodríguez Cadena, Yolanda. 2009. *El habla de Barranquilla: materiales para su estudio*. Barranquilla: Universidad del Atlántico.
- Rosenblat, Ángel and Paola Bentivoglio. 1979. *El habla culta de Caracas: materiales para su estudio*. Caracas: Ediciones de la Facultad de Humanidades y Educación Instituto de Filología "Andrés Bello," Universidad Central de Venezuela.
- Sá, Maria da Piedade Moreira de. 2005. *A linguagem falada culta na cidade do Recife: materiais para seu estudo*. Recife: Universidade Federal de Pernambuco, Programa de Pos-Graduação em Letras.
- Salvador Salvador, Francisco and Gonzalo Águila Escobar (eds.). 2006. *El habla culta de Granada: materiales para su estudio*. Granada: Editorial de la Universidad de Granada.
- Samper Padilla, José Antonio, Clara Eugenia Hernández Cabrera, and Magnolia Troya Déniz. 1998. *Macrocorpus de la norma lingüística culta de las principales ciudades del mundo hispánico* (CD-ROM). Las Palmas de Gran Canaria: Universidad de Las Palmas.
- Sayahi, Lotfi. 2004. The Spanish language presence in Tangier, Morocco: a sociolinguistic perspective. *Afro-Hispanic Review* 23(2).54–61.
- Sayahi, Lotfi. 2005. El español en el norte de Marruecos: historia y análisis. *Hispanic Research Journal* 6.3: 195–207.
- Sayahi, Lotfi. 2006. Phonetic features of northern Moroccan Spanish. *Revista Internacional de Lingüística Iberoamericana* 4(2).167–180
- Sosa, Juan. 1999. *La entonación del español*. Madrid: Cátedra.
- Tarkki, Pekka. 1995. *El español en los campamentos de refugiados de la República Árabe Saharaui Democrática*. Helsinki: University of Helsinki, Ibero-American Institute.
- Thomas, Earl. 1974. *A grammar of spoken Brazilian Portuguese*. Nashville: Vanderbilt University Press.
- Toribio, Almeida Jacqueline. 2000. Setting parametric limits on dialectal variation in Spanish. *Lingua* 110.315–341.
- Truckenbrodt, Hubert, Filomena Sandalo, and María Bernadete Abaurre. 2008. Elements of Brazilian Portuguese intonation. *Journal of Portuguese Linguistics* 8.77–115.
- Universidad Nacional de Buenos Aires. 1987. *El habla culta de la ciudad de Buenos Aires: materiales para su estudio*. Buenos Aires: Universidad Nacional de Buenos Aires.
- Vida Castro, Matilde. 2007. *El español hablado en Málaga: corpus oral para su estudio sociolingüístico: nivel de estudios bajo 1*. Málaga: Sarriá.
- Wherritt, Irene. 1985. Portuguese language use in Goa, India. *Anthropological Linguistics* 27.437–451.
- Zamora, Juan and Jorge Guitart. 1988. *Dialectología hispanoamericana*. Salamanca: Ediciones Almar, 2nd ed.
- Zamora Vicente, Alonso. 1967. *Dialectología española*. Madrid: Gredos.

31 Dialects of the Slavic Languages

VLADIMIR ZHOBOV AND
RONELLE ALEXANDER

31.1 Introduction

The Slavic (or Slavonic) languages, of which the best known and most widely spoken is Russian, are a branch of the Indo-European language family, closely related to the Baltic branch. As a group, they are marked by considerable structural unity, yet each of them displays rich dialectal diversity and dialects of one language frequently shade into those of neighboring languages. Slavic scholars have produced numerous regional dialect atlases, dialect dictionaries, local dialect descriptions, and scholarly investigations of dialectal phenomena. Indeed, Slavic dialectology has contributed to general linguistics in several ways. In its famous “*thèses*” of 1929, the Prague Linguistic Circle not only articulated the modern concept of a standard language (and its relation to dialects) but also introduced the term “linguistic geography,” and first proposed the idea of a pan-Slavic linguistic atlas. The concept of Sprachbund or linguistic alliance, coined by Nikolai Trubetzkoy in 1928, is also associated with the Prague circle. The importance of dialects in the well-known Balkan Sprachbund (Albanian, Greek, Balkan Romance, and Balkan Slavic dialects) has been highlighted by Balkan Slavists (Alexander 2012).

Slavic dialectology introduced structuralist concepts early (Stankiewicz 1957 and Ivić 1958), and Slavic dialectal material (from Bulgarian) was used to demonstrate the potential of the innovative approach of dialectometry (Osenova, Heeringa, and Nerbonne 2009; Prokič 2010). Slavic is one of the best-studied branches of Indo-European in terms of dialectal variation. The close relationship among Slavic languages has allowed for their representation together in atlas format, on the model of dialectology, in the All-Slavic Linguistic Atlas (Institut russkogo jazyka 2010–2012). Sociolinguistic aspects of dialects are also very relevant within Slavic, especially with respect to Czech diglossia, where both *spisovná čestina* (written Czech) and *obecná čestina* (“common” or spoken Czech), function as standards (Sgall *et al.* 1992), and in South Slavic where urban dialects which differ markedly from the literary standard have achieved considerable prestige, such as that of Split in Croatia (Jutronić 2010) or Veliko Tărnovo in Bulgaria (Videnov and Bajčev 1999).

The following sections identify the major dialect groups of all Slavic languages within the three major sub-branches (East, West, and South) and identify major isoglosses within the best-known languages. The focus is restricted to dialects within each country’s borders, despite the scholarly significance of a number of diaspora dialects. The data in each case are presented according to the dialectological traditions of each of the individual regions: phonological isoglosses are given more weight both because of their comparative salience within

linguistic structure and because they have been best studied. Russian, as the language with by far the most speakers, is treated in some detail; the other languages are of necessity treated in a more cursory manner.

31.2 East Slavic

The East Slavic languages—Russian, Ukrainian, and Belarusian—constitute a dialect continuum, with a high degree of mutual understanding. The first dialect map of Russian (Durnovo, Sokolov, and Ušakov 1915) regarded Belarusian and Ukrainian (called White Russian and Little Russian, respectively) as dialects of Russian (called Great Russian). Each was termed a *narečije*, called here in English “macro-dialect.”

31.2.1 Russian

Russian, the major Slavic language, is spoken natively (or quasi-natively) by 110 million speakers in the Russian Federation and 40 million more in virtually all former Soviet republics. It is the official language throughout the Russian Federation and co-official in Belarus, Kazakhstan, Kyrgyzstan, and Tajikistan (along with the native language of each of these units). Basic resources for dialectology are Avanesov and Bromlej (1986–1996), numerous regional dialect dictionaries, the largest of which (Filin 1965–) has now reached the letter T, and major survey handbooks (Avanesov and Orlova 1965, Kasatkin 2005).

The systematic study of Russian dialects, however, is normally restricted to the so-called territories of old formation. Later expansions—both east of the Urals and to the south and east in European Russia—are not taken into account in dialect classifications. Russian dialects proper are divided into two large macro-dialects (*narečija*)—Northern (NR) and Southern (SR); the area transitional between them, called Central Russian (CR), is not accorded the same status. The primary isoglosses are phonetic (the development of unstressed vowels and elements of the consonant inventory), but the division is also supported by morphological and lexical features. Standard Russian combines the vocalism of SR and the consonantism of NR.

The major characteristic of NR is “okanje,” or the consistent distinction between /o/ and /a/ in unstressed syllables; south of this isogloss, unstressed /o/ and /a/ merge according to various rules. The major characteristic of SR is the lenition of the voiced velar stop /g/ to a fricative ([ɣ] or more rarely [ɦ]); north of this isogloss /g/ remains a stop. Neither feature is found in the central group, which is characterized by crisscrossing isoglosses of other basic traits of NR and SR.

Another major set of differences concerns prosody, specifically the ratio of stressed to unstressed vowels (in different positions) with respect to their relative length. Using a convention where 3 indicates stress and 1 and 2 indicate, respectively, greater and lesser quantitative reduction of unstressed vowels, one may describe these differences numerically, using in each instance a tetrasyllabic word with penultimate stress. The standard pronunciation, and that of many dialects, shows a gradation of unstressed syllables (1-2-3-1). In NR dialects all unstressed vowels are equal, either very close in length to the stressed one (2-2-3-2) or quite different from it (1-1-3-1), whereas in CR dialects pretonic vowels are frequently equal in length to, or even longer than, the tonic one (1-3-3-1). Finally, in some SR dialects the rhythmic structure depends on the quality of the stressed vowel: the pretonic vowel is weak if the tonic vowel is low (1-1-3-1) but quite long if it is high (1-3-3-1). These rhythmic patterns (instrumentally confirmed by Požarickaja 2005: 28–33) appear to be correlated with the specific development of vowels in the first pretonic position. The average rate of speech is also different: NR is the fastest, SR the slowest, with CR in between.

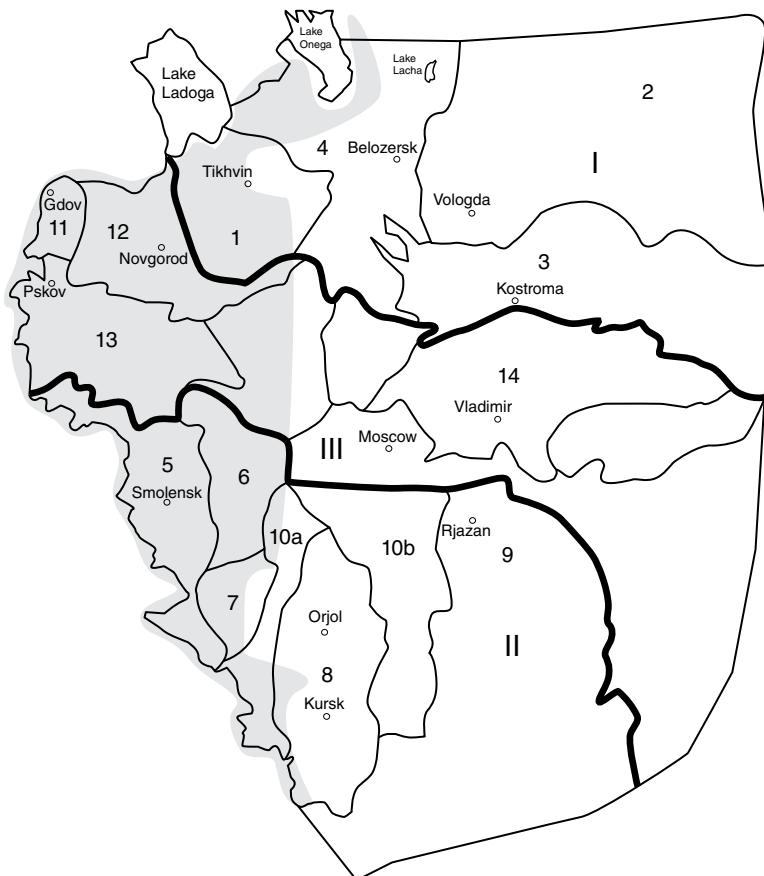


Figure 31.1 Map of Russian dialect regions discussed in the text.

Legend: darker lines indicate borders between major regions, labeled with Roman numerals; lighter lines indicate borders between sub-regional dialects areas, labeled with Arabic numerals. I North Russian: 1. Ladoga-Tikhvin; 2. Vologda; 3. Kostroma; 4. Interzonal. II South Russian: 5. Western; 6. Upper Dnepr; 7. Upper Desna; 8. Kursk-Orjol; 9. Rjazan; 10a Western interzonal; 10b Eastern interzonal. III Central Russian: 11. Gdov; 12. Novgorod; 13. Pskov; 14. Vladimir-Volga. Gray area: Western dialect zone.

Each of these three regions is subdivided into a number of regional dialects (*govory*). The main regions and their sub-dialects are shown in Figure 31.1. Only dialects discussed in the text are indicated on the map. The following text gives first the additional traits characterizing each of the two major *narečija*, and then the most significant traits of the regional dialects within it.

31.2.1.1 North Russian

As noted above, the major trait of NR is (a) *okanje*, or the non-merger of unstressed /o/ with other vowels, for example, *domá* “at home” (versus *daváj* “give [imper]”), *górod* “city” (versus *ókna* “window [gen.sg.]”). Other basic traits of NR are (b) the loss of intervocalic /j/, sometimes with ensuing assimilation and contraction of vowels, for example, *délajet* > *délaet*, *délaat*,

or *délat* "s/he does"; *nóvaja* > *nóvaa* or *nóva* "new [f.]"; (c) assimilation of the cluster /bm/, for example, *om:án* "swindle"; simplification of word-final /st/, for example, *mós* "bridge"; (d) the change of /e/ to /jo/ in post-tonic syllables before a hard consonant, for example, *výn-íosu* "bring out," *óz'oro* "lake"; (e) the syncretism of dative and instrumental plural, for example, *k pustým vedrám* "to empty pails," *s pustým vedrám* "with empty pails"; (f) the presence of a post-posed declinable particle used for emphasis, for example, *dóm-ot* "home," *žená-ta* "woman," *ízby-te* "cellars," *sérdce-to* "heart."

There are four major regional dialects within NR. Moving from St. Petersburg eastwards they are (i) Ladoga-Tikhvin, (ii) the Onega-Lacha-Belozersk complex (sometimes called "inter-zonal"), (iii) Vologda to the northeast of this complex, and (iv) Kostroma to the southeast of it.

The **Ladoga-Tikhvin** dialect is characterized by (a) /i/ as the reflex of jat' in certain stems and consistently in loc.sg. for example, *xlib* "bread," *zvir* "beast," *na stolí* "on the table [loc.]"; and (b) an inserted vowel in liquid + obstruent clusters, for example, *stolób* "pillar," *sérjep* "sickle," *górob* "back."

The **Vologda** dialect is characterized by (a) distinct reflexes of jat versus etymological /e/ and /e/ from front jer (e.g., *siéno* "hay" but *sémi* "seven," *otéc* "father"), and also of /o/ under Common Slavic rising pitch versus /o/ from back jer and /o/ under falling pitch (e.g., *vuólja* "will" but *son* "dream," *zóloto* "gold"), giving a seven-vowel system as opposed to a five-vowel system in most other dialects; (b) the raising of /e/ and /a/ between soft consonants, for example, *v slíne* "in the hay" [versus *sieno* "hay"], *gríesj* "filth" [versus *grázny* "dirty"]; (c) the shift of /v/ to /w/ in other than pre-vocalic position, for example, *traw* "grass [gen.pl.]," *tráwka* "small grass"; (d) the merger of /c/ and /č/, for example, *cáj* "tea," *jajcó* "egg"; (e) the ending /é/ in dat-loc of third declension nouns, for example, *grézié* "filth" [dat-loc]; (f) the affixes /sé/ and /s'o/ in reflexive verbs, for example, *umýls'e*, *umýls'o* "washed [oneself]."

The **Kostroma** group is characterized by (a) progressive assimilation after paired soft consonants, for example, *réd'kia* "radish"; (b) partial assimilation of the cluster /vn/, for example, *damnó* "long ago," *mnúk* "grandson"; (c) the ending -tō in 2pl. imperative, for example, *ídítō* "go!", and present tense *nesetō* "you carry."

The "interzonal" groups are characterized primarily by (a) the shift of /hv/ to [f], for example, *fos(t)* "tail"; (b) the 2sg. verbal forms *dasí* "give" and *jesí* "eat"; (c) the masc-neut gen.sg. endings as in *nóvoyo* or *nóvoo* "new."

31.2.1.2 South Russian

As noted above, the major trait of SR is (a) the lenition of /g/, for example, *noyá* "leg"—*nox* "leg [gen.pl.]". Other basic traits of SR are (b) the merger of unstressed /o/ and /a/ after hard consonants ("akanje"), for example, *damá* "at home"; (c) varying shifts of non-high unstressed vowels after soft consonants, all subsumed under the general term "jakanje"; (d) the insertion of a vowel in initial consonant clusters, for example, *p[la]šenica* "wheat", *s[a]moródina* "currant"; (e) the ending -e in gen.sg. of feminine nouns with hard consonant stem, for example, *žené* "woman," *rabóté* "work"; (f) the ending -y in nom.pl. of neuter nouns, for example, *pátny* "stains," *ókny* "windows"; (g) the ending -e in both gen-acc and dat-loc of personal and reflexive pronouns, for example, *mené*, *tebé*, *sebé* (also dat-loc *mné*); (g) the ending -t̄ in both singular and plural verbal endings, for example, *nesét̄* "s/he carries," *nesút̄* "they carry"; (i) the ending -ut̄ in 3pl. of first and second conjugation verbs, for example, *píšut̄* "write," *kól'ut̄* "slaughter," *dyjsút̄* "breathe," *nósut̄* "carry."

Moving from west to east, the regional dialects of SR are (i) the western group centered on Smolensk, (ii) Upper Dnepr, (iii) Upper Desna, (iv) Kursk-Orjol, (v) southern interzonal, and (vi) Rjazan.

The **western** dialect is characterized by (a) "Žizdrina dissimilative jakanje," where first pretonic non-high vowels following soft consonants are lowered before a tonic non-low

vowel but raised before a tonic low one, for example, *siastrý* "sister [gen.sg.]," *ríaké* "river [loc.sg.]," *p'atnó* "stain," but *sístrá* "sister [nom.sg.]," *l'isá* "forest [gen.sg.]."; (b) the adjectival endings -yj and -ej, without expected consonant softening, in masc.nom.sg., for example, *molodýj* or *molodéj* "young"; (c) plural endings after the paucal numerals, for example, *dvá mužíkí* "two men"; (d) second singular present *dasí* "give" and *jesí* "eat."

The **Upper Dnepr** dialect is characterized primarily by "dissimilative moderate *jakanje*," where first pretonic non-high vowels are raised before all soft consonants and before hard consonants followed by tonic /a/, but lowered in all other positions, for example, *típiér* "now," *strił'át* "shoot," *l'isá* "forest [gen.sg.]," *n'islá* "carried," but *p'atnó* "stain," *p'aklí* "bake [past.part.pl.]".

The **Upper Desna** dialect is characterized by (a) "*Žizdrina* dissimilative *jakanje*"; (b) the replacement (by /i/) or loss altogether of unstressed initial /o/, for example, *itéc*, *téc* "father," *ip'át* "again"; (c) 2sg.pres. *dadíš* "give," *jediš* "eat"; (d) levelling of consonant alternation in 1sg.pres. of second conjugation, for example, *víd'u* "see," *prós'u* "ask."

The **Kursk-Orjol** dialect is characterized by (a) "*Sudža* dissimilative *jakanje*," where first pre-tonic non-high vowels following soft consonants are lowered before tonic /i/, /u/, and original /o/, but raised elsewhere, for example, *n'asú* "I carry," *l'asú* "forest [loc.sg.]," *n'así* "carry [imper.]," *s'amú* "family [acc.sg.]," *p'lacó* "shoulder," but *tip'éri* "now," *glijidéjt* "to look," *p'it'órká* "number 5"; (b) the shift of /č/ to /š/, for example, *šlaj* "tea," *dóš'ka* "daughter"; (c) progressive assimilation after both paired soft consonants and /j/, for example, *bán'ká* "bath-house," *kopéjká* "kopek," but *dóš'ka*; (d) partial assimilation of the cluster /vn/, for example, *damnó* "long ago."

The **Rjazan** dialect is characterized by (a) different types of assimilative-dissimilative *jakanje*; (b) strong reduction, or loss, of second pre-tonic vowels, for example, *yvarjú* "I speak," *pšlá* "she went"; (c) the retention of stressed /e/ before hard consonants, for example, *séstry* "sisters," *nesét* "s/he carries"; (d) the seven-vowel system in some dialects, and (e) progressive assimilation after all soft consonants and /j/, for example, *bán'ká* "bath-house," *dóč'ká* "daughter," *kopéjká* "kopek."

The **southern interzonal group** is divided into two types. The western half shares certain traits of Kursk-Orjol, frequently shifts neuter nouns to masculine, for example, *takój molokó* "such milk," *bol'sój oknó* "big window," and has -oju for instr.sg. of feminine adjectives: *nóvoju*. The eastern half shares certain traits of Rjazan, and also loses -t in 3sg.pres., for example, *nesjó* "s/he carries," *délajo* "s/he does."

31.2.1.3 Central Russian

The regional dialects of CR, moving again from west to east, are Gdov, Pskov, and Vladimir-Volga.

The **Gdov** dialect is characterized by (a) a reduction pattern whereby first pretonic /o/ is retained before mid vowels but merges with /a/ elsewhere, for example, *vodé* 'water [loc.sg.], *malokó* "milk," but *vadá* "water [nom.sg.]," *vadý* "water [gen.sg.]," *górad* "city"; (b) loss of -t in 3.pres., for example, *d'élaje* "s/he does," *nesú* "they carry"; (c) 2.pres. forms *dástiš* "give," *jéstíš* "eat." In the **Novgorod** dialect, sometimes grouped here, unstressed /o/ and /a/ remain distinct.

The **Pskov** dialect is characterized by (a) "strong *jakanje*," where first pretonic non-high vowels merge consistently with /a/, for example, *n'aslá* "carried," *n'así* "carry!", *ríaká* "river [nom.sg.]," *ríaké* "river [loc.sg.]," *p'aták* "five kopek coin"; (b), the lowering of /u/ in post-tonic and second pretonic syllables, for example, *ókan* "perch," *røkavá* "sleeves"; (c) the affricativization of soft alveolar stops, for example, *mátsj*, "mother," *džéni* "day"; (d) the merger (in the Pskov region) of /c/ and /č/, for example, *túca* "cloud," *úlica* "street"; (e) the lack of agreement in passive participles used predicatively, for example, *jágody nabráń* "the strawberries have been picked."

The **Vladimir-Volga** dialect is characterized by (a) “incomplete *okanje*,” where /o/ and /a/ remain distinct in first pretonic syllables, but are both reduced to /ə/ elsewhere, for example, *gəlová* “head,” *górad* “city,” and where initial /o/ is raised in second pretonic syllables, for example, *ugurcý* “cucumbers”; (b) the contraction of /aje/ to /a/, for example, *délat* “s/he does,” *znát* “s/he knows.”

The development of linguistic geography led to a new, more complex stratification of linguo-territorial formations within Russian dialects. The analysis of newly collected data (Avanesov and Bromlej 1986–1996) led to the concept of dialect “zones.” Like the three major “regions,” they are larger areas defined by bundles of isoglosses. Unlike the regions, distinctions among which are correlative, the dialect zones are defined by independent features. Furthermore, the boundaries of dialect zones are less definite, such that a single dialect “group” can be included in several dialect zones. These eight zones are defined geographically: northern, northwestern, northeastern, southern, southwestern, southeastern, western, and central. Of these, the western zone is the most clearly defined, by (a) hard pronunciation of the affricate /č/; (b) a prothetic /j/ in forms of third-person personal pronouns, for example, *joná, jonó, jony*; (c) an extended stem of demonstrative pronouns, for example, *tája túju, tóje*; and (d) predicative usage of participles, for example, *pójezd ušovšy* “the train left.”

31.2.2 Ukrainian

Ukrainian is the official language of Ukraine. In the 2001 census, out of the total population of 48.5 million, 37.5 million declared themselves ethnic Ukrainians and 32.5 million considered Ukrainian their native language, a 2.8 % increase compared to 1989. (Of course in this context “native language” does not always mean “most frequently used language”; most Ukrainian speakers are equally fluent in Russian.) There are three main dialects of Ukrainian: Northern (NU, also called Polissian), South-Western (SWU), and South-Eastern (SEU). The standard language is based on SEU. Basic resources include Zakrev’ska and Matvijas (1984–2001) and Bevzenko (1980).

The major traits of NU are (a) diphthongization or raising of mid-vowels in new closed syllables under stress, for example, *kuoni*, *kueni*, *kuiní*, *kuní* and *kiní* “horse,” *prin'uós, prin'ués, prin'uís* “brought,” *p'ieč, pič* “oven”; (b) lack of merger of unstressed /o/ and /a/ except in the far northwest, for example, *galavá* “head”; (c) the ending -je in neuter nouns following long soft consonants, for example, *žit'lé* “existence”; (d) the ending -y on nom. pl. adjectives, for example, *molodý* “young.”

The major traits of SWU are (a) raising of unstressed mid vowels, for example, *se'ló* “village,” *kuróva* “cow”; (b) devoicing of final obstruents (unlike in Standard Ukrainian and most other dialects), for example, *snik* “snow,” *zup* “tooth”; (c) the ending -je on neuter nouns, with no long consonants, for example, *žit'é* “existence”; (d) hard -t in verbal endings, for example, *xódít* “s/he walks,” *xódítat* “they walk”; (e) archaic past tense with affixed auxiliary, for example, *hodívjem* “I was going”; (f) use of enclitic prounoun objects.

The major traits of SEU are (a) softening of hard /r/, for example, *r'áma* “frame,” *komór'a* “pantry”; (b) shift of /f/ to /xv/, for example, *xvábrika* “factory”; (c) lack of alternation in first singular present, for example, *xód'u* “I walk,” *sid'ú* “I sit”; (d) contracted third singular present, for example, *zná* “s/he knows,” *dúma* “s/he thinks.”

31.2.3 Belarusian

Belarusian, together with Russian, is the official language of Belarus, although the dominant language in most public spheres is Russian. Out of 9.5 million people in the 2009 census, about 5 million listed Belarusian as their native language and another 1.2 million as a second

language, yet only 2 million state they use Belarusian at home. There are two main dialects of Belarusian, North-Eastern (NEB), and South-Western (SWB), with a transitional zone between them located around the capital, Minsk. Basic resources include Avanesov, Krapiva & Mackevič 1963 and Blinava and Mjaceljskaja 1980.

31.3 West Slavic

West Slavic consists of a Lekhitic group, represented by Polish (40 million speakers), a Czech-Slovak group (10 million speakers of Czech and 5 million of Slovak—the two languages are largely mutually intelligible), and the now isolated Sorbian (also known as Lusatian or Wendish), with a rapidly dwindling number of speakers.

31.3.1 Polish

The major dialect areas of Polish correspond to historical-geographic regions: Małopolska ("Little Poland," in the southeast, centered on Kraków), Wielkopolska ("Great Poland," a north central region centered on Poznań), Mazowsze (Mazovia, in the northeast, centered on Warsaw), and Śląsk (Silesia, in the southwest, centered around Katowice). Scholars disagree as to which of the first two is the base of the standard language. A fifth region, Kashubia (in the far north, centered on Gdańsk), is regarded by some as a Polish dialect and by others as a separate language. Basic resources include Dejna (1993) and a great number of dialect atlases, especially Dejna (1998–2002) and Stieber *et al.* (1964–1978).

There are two basic isoglosses over the larger territory: (a) voicing of word-final obstruents in Little Polish, Great Polish, and Silesia, for example, [brad muj] for /brat muj/ "my brother" versus devoicing in Mazovia and Kashubian, for example, [vus muj] for /vuz muj/ "my wagon"; (b) the maintenance of a three-way distinction of sibilants in Great Polish and southern Silesia, for example, alveolo-palatal [śano] "hay," [żarno] "grain," [ćasto] "dough," [dżadek] "grandfather" ~ post-alveolar [duša] "soul," [žaba] "frog," [čapka] "hat," [džuma] "plague" ~ alveolar sam "same," zamek "castle," car "tsar," sadza "soot," versus the merger of the latter two groups in Little Polish, northern Silesia and Mazovia, for example, *dusa, zaba, capka*, a phenomenon called "*mazurzenie*".

Great Polish dialects are characterized by (a) diphthongization of /a/, /o/ and /y/, for example, *ptaćk, ptɔćk, ptoćk* "bird," *bioso, boęso* "barefoot," *leżyi* "s/he lies," *żyjeć* "life," *myjysi* "mice"; (b) asynchronic and raised pronunciation of nasalized vowels before stops and affricates, for example, *fšyndże* "everywhere," *pwožundek* "order," and the loss of nasality before fricatives in the south, for example, *ćyški* "heavy," *kšuška* "book"; (c) prothetic /v/ before initial /o/, for example, *vovies* "oats."

Little Polish dialects are characterized by (a) the shift of final /x/ to /k/, for example, *byuek* "I was," *dak* "roof"; (b) asynchronic and raised pronunciation of the front nasal in the mountains, for example, *sſynty* "holy," *prynžy* "faster" versus synchronic and lowered pronunciation in the foothills, for example, *sſfaty, prązy*.

Mazovian dialects are characterized by (a) the pronunciation of soft labials (phonetically palatalized in other dialects) as clusters, for example, *vjara* "faith," *bjala* "white," *fiłut* "joker," *mjasto* "place," with variants such as *jara* or *żara*, *bżala*, *ńastu*; (b) the merger of /y/ and /i/, for example, *lodi* "ice [pl.]," *ribi* "fish [pl.]".

Silesian dialects are characterized by (a) mixture of post-alveolar and alveopalatal consonants into an intermediate "palatalized post-alveolar" sound, for example, *śidło* "awl," *żelezny* "iron [adj.]"; (b) diphthongization of /a/ and /o/, for example, *ptaćk, ptoćk* "bird," *koęvouł* "smith," *guya* "wood"; (c) fricative /r/, for example, *přibić* "to hammer"; (d) anticipation of softness *ńejiše* "s/he carries," *zaiś* "but"; (e) initial stress in the south.

Kashubian is characterized by (a) the diphthongization of /o/, for example, *goura* or *guyra* "wood," *kuoza* or *kyeza* "goat"; (b) frequent fronting of ó, for example, *toę* "this," *uýškoe* "bed"; (c) the pronunciation of soft labials as clusters, for example, *cerpiec* "endure," *bjic*, *bçic* "beat," *vžara* "faith," *traſqa* "hit, score"; (d) schwa for short high vowels, for example, *dëmu* "smoke [gen.]," *rëbë* "fish [pl.]," *cëgní* "drag," *dësha* "soul"; (e) replacement of syllabic /l/ by /ɔł/ or /ɔł/, for example, *vøuk* "wolf," *mþučec* "keep silent"; (f) soft consonants before syllabic /r/, for example, *mjártwi* "dead"; (g) initial stress in the north and free stress in the south.

31.3.2 Czech

The largest dialect of Czech is Bohemian, spoken in the western and central regions, followed by Central Moravian (or Hanák) to the southeast, centered on Brno, and Lach (or Silesian) further to the southeast; some classifications also include Eastern Moravian (or Moravian-Slovak), centered on Ostrava. The basis of standard Czech is the Bohemian dialect. The primary differentiating features are the pronunciation of long vowels (marked with acute accents), and "metaphony," or the shift of /a/ and /u/ to /e/ and /i/ after soft consonants, for example, *jev* "phenomenon," *jih* "south," *klíče* "key [gen.sg.]." Basic resources include a dialect atlas (Balhar and Jančák 1992–2005) and authoritative overviews (Koudela 1964, Belič 1988).

Bohemian dialects are characterized by (a) the shift of /ý/ to /ei/ and of /é/ to /í/, for example, *streic* "uncle," *beik* "bull," *mlíko* "milk"; (b) the addition of /v/ before initial /o/, for example, *vohěn* "fire"; (c) the pl.instr. ending -ma, for example, *s kravama* "with cows." In eastern Bohemian the softness of all labials before /e/ was lost and restored later as a separate /j/ by a more general rule, but the soft /m/ is still pronounced differently: *mničet'* versus *mjet'* "copper." In southwestern Bohemian (around České Budějovice) the softness, again as /j/, is preserved even before /i/: *bjič* "whip."

Central Moravian dialects are characterized by (a) frequent loss of vowel length, for example, *prah* "threshold," *blato* "mud," *žaba* "frog," *březa* "birch," *vietr* "wind"; (b) the shift of /ý/, /í/ and /ei/ to /é/, of /ou/ to /ó/ and of /y/ to /e/, for example, *dobré* "good," *stréc* "uncle," *nělepší* "best," *móka* "flour," *rebe* "fish [pl.]"; (c) the occurrence of metaphony only within roots, for example, *ležet* "s/he lies," *cizí* "foreign [pl.]," but *duša* "soul," *píju/píjo* "I drink"; (d) sandhi voicing before vowels and sonorants, for example, *g mostu* "towards the bridge."

Eastern Moravian dialects are characterized by (a) the consistent shift of /ou/ to /ú/, for example, *múka* "flour," *vedú* "they know"; (b) the frequent shift of /í/ to /é/, for example, *kamének* "pebble"; (c) the complete absence of metaphony, for example, *ležat* "s/he lies," *cuzí* "foreign [pl.]"; (d) the retention of syllabic /r/ but the shift of syllabic /l/ to /u/, for example, *puný* "full."

Lach dialects are characterized by (a) the lack of vowel length, for example, *muka* "flour," *dobry* "good," *nejlepši* "best"; (b) the limitation of metaphony to the shift /a/ > /e/ and only within the root, for example, *ležet* "s/he lies" but *cuzí* "foreign [pl.]" *duša* "soul," *piju* "I drink"; (c) the replacement of syllabic liquids by a sequence of vowel+liquid, for example, *pylny* "full," *kryk* or *kyrk* "throat"; (d) sandhi voicing as in Central Moravian; (e) palatalization of /s/, /z/, /t/, /d/ before front vowels and /j/, for example, *šeňo* "hay," *žima* "winter," *něbudžeće* "be not!" [imper. pl.]; (f) fixed stress on the penultimate syllable.

31.3.3 Slovak

Slovak, the official language of the Slovak Republic, is spoken natively by 80% of the total population of 5.4 million and also by minorities in the Czech Republic, Hungary, and the Serbian province of Vojvodina, where it has official status. There are three main

dialects: central, western, and eastern. The eastern dialect group, with fixed penultimate stress and the loss of vowel quantity, constitutes a transition to Polish. The central group shares several important features with the northernmost South Slavic language, Slovene. Krajčovič (1988) and Štolc (1968–1984) are basic resources.

31.3.4 Sorbian

The two varieties of Sorbian, both located in the far eastern regions of Germany, are considered separate languages. Upper Sorbian, with fewer than 20,000 speakers, is centered around Budišyn (German Bautzen) and shares certain features with Czech, for example, the lenition of /g/. Lower Sorbian, with fewer than 10,000 speakers, is centered around Chošebuz (German Cottbus) and shares certain features with Polish, for example, the three-way distinction of sibilants.

31.4 South Slavic

South Slavic represents a gradual speech continuum whose dialects are treated under the rubric of a number of different national languages. In the far west is Slovenian (2 million speakers) and in the south and east are Macedonian (2 million speakers) and Bulgarian (7 million speakers). The intervening region contains a single language system, formerly Serbo-Croatian, now viewed as separate national-ethnic languages: Croatian (5.5 million speakers), Bosnian (3 million speakers), Serbian (8.7 million speakers), and Montenegrin (number of speakers disputed). Nevertheless, these still constitute a single communicative system, most frequently called BCS by outsiders and often called “central South Slavic” by natives. There are no linguistic atlases available, but Ivić (1981) gives phonological summaries for all of South Slavic except Bulgarian.

31.4.1 Slovenian

The small mountainous Slovenian region has the densest dialectal diversity of all of Slavic. Its seven major zones, called by Slovenianists “dialectal bases” (Logar and Rigler 1983; Logar 1993), are differentiated primarily by vocalic and prosodic features. The central dialects, Upper and Lower Carniolan, form the basis of the standard language. They are closer to the Styrian zone on the east and the Rovte zone on the west than to the more peripheral zones: Pannonian in the far northeast (bordering with Hungary), Carinthian in the north, and Littoral in the west; each of the latter borders with, and extends into, Hungary, Austria, and Italy, respectively. Contact phenomena in each of these three zones are evident in a number of dialectal features.

31.4.2 Bosnian/Croatian/Serbian

Until 1991, the language known as Serbo-Croatian was divided into three major groups that are quite distinct from one another: Kajkavian, Čakavian and Štokavian (Brozović and Ivić 1988). Kajkavian is spoken in northwestern Croatia, adjacent (and transitional) to Slovenian; it forms the base of the urban dialect of Zagreb. Čakavian is spoken in Istria, along the northern and central Dalmatian coast, and on most islands; it also forms the base of the urban dialect of Split. Because these regions are now completely within Croatia, Kajkavian and Čakavian are now considered dialects of Croatian (Lončarić 1996, Lisac 2009). Štokavian is spoken throughout the rest of Croatia and in all of Bosnia-Herzegovina, Serbia, and Montenegro.

Kajkavian dialects are characterized by (a) complex vowel systems, opposing broad /e/ [ɛ] (from front nasal and original /e/, for example, *pet* “five,” *selo* “village”) to /e/ (deriving from *jat* and the *fers*, for example, *leto* “summer,” *pes* “dog”) and /o/ (deriving from original /o/, for example, *most* “bridge”) to raised /o/ [ɔ] (deriving from back nasal and vocalic /l/, for example, [zɔp] “tooth,” [vɔk] “wolf”); (b) the periphrastic future with *be-auxiliary* and the active participle, for example, *bum došel* “I will come”; (c) accentual systems with tonal oppositions in long vowels and preservation of original accent position; (d) distinct plural case forms, for example, *k ženam*, *o ženah*, *s ženami* “to /about/ with women”; (e) retention of final /-l/, for example, *došel* “came.”

Čakavian dialects are characterized by (a) the preservation of Common Slavic pitch distinctions and original accent placement, combined with a complex set of vowel lengthenings (Langston 2006); (b) mixed reflexes of *jat*, where the southeast has /i/, for example, *misto* “place,” whereas central and northwestern groups have /e/ after dental consonants before non-front vowels but /i/ elsewhere, for example, *leto* “summer,” *ded* “grandfather,” *mera* “measure” but *dite* “child,” *rika* “river”; (c) distinct plural case forms, for example, *k ženan*, *o žena s ženami*; (d) retention of final /-l/, for example, *bil* “was.”

Štokavian dialects are divided into “New Štokavian” and “Old Štokavian.” The former group is characterized by (a) the syncretism of dat.-loc.-instr. pl., for example, *k ženama*, *o ženama*, *sa ženama*; (b) shift of final /-l/ to /-o/, for example, *bio* “was”; (c) the “neoštakavian retraction”—the shift of accent one syllable toward the beginning of the word resulting in a new rising tone. Subgroups of Štokavian are defined according to the consistency of the “neoštakavian” changes (Ivić 1958), and the reflex of *jat*. In the first instance, dialects are either (i) fully New Štokavian, (ii) limited New Štokavian or (iii) Old Štokavian. In the second instance, they are “ekavian,” for example, *mleko* “milk,” “ikavian,” for example, *mliko* “milk,” or “(i)jekavian,” with /je/ in short syllables and /ije/ in long syllables, for example, *koljeno* “knee” but *mlijeko* “milk.” Croat dialectologists now divide Štokavian into western [W] and eastern [E] zones (Lisac 2003).

There are three major “fully New Štokavian” dialects: (a) the **East Herzegovina** dialect (ijekavian), which has spread broadly throughout the region due to migrations, and which now forms the base for all the current standard languages [W]; (b) the **Western** dialect, sometimes called “younger ikavian,” located in western Herzegovina, along the Croatian coast, in parts of Bosnia and a pocket of northern Vojvodina [W]; (c), the **Šumadija-Vojvodina** dialect (ekavian), covering northern and north-central Serbia [E].

There are four major “limited New Štokavian” dialects: (d) **Eastern Bosnian** dialects (ijekavian) with only partly retracted accents [W]; (e) **Archaic Slavonian** dialects, located in northeastern Croatia, (mixed ekavian and ikavian), with mostly unretracted accents, and with both archaic Čakavian tonal contours and the new rising neoštakavian tones [W]; (f) the **Zeta-Lovćen** dialect (ijekavian), in Montenegro and southwestern Serbia, with mostly unretracted accents but preservation of length, and syncretism only of dat. and loc. pl. [E]; (g) the **Kosovo-Resava** dialect (ekavian), covering south-central and eastern Serbia, with shortening of post-tonic length and incomplete retractions [E].

Southeastern Serbian dialects (ekavian) are sometimes classed within Štokavian [E] as **Prizren-Timok** and sometimes as a separate major group called **Torlak**. Accents are unretracted and vowel length is completely lost. In addition, the reflex of both *fers* is /ə/, for example, *dən* “day,” *sən* “sleep.” This group is also characterized by traits seen in Bulgarian and Macedonian, all of which have been termed “Balkan,” since they characterize languages of the Balkan Sprachbund (including also Albanian, Greek, and Balkan Romance). The major traits are (a) drastic reduction (or loss altogether) of case distinctions; (b) replacement of infinitive by a complementizer clause; (c) post-posed definite articles; (d) expression of

future tense by an unchanging particle derived from the verb “want” plus the present tense; and (e) analytic formation of adjectival comparison. The region comprising Torlak, Macedonian, Bulgarian, and Slavic dialects spoken in Greece and Albania is frequently called “Balkan Slavic.” Many Bulgarian linguists, however, consider (and call) this entire area Bulgarian (BAN 2001).

31.4.3 Bulgarian

The basic dialect division of Bulgarian into Eastern (EB) and Western (WB), is determined by reflexes of the *jat* vowel. In the west, *jat* is consistently replaced by /e/, and although *jat* reflexes in the east are less uniform, an open vowel is usually involved. The most frequent reflex is the alternation /ja/ ~ /e/. The former reflex, /ja/, occurs only under stress when none of following segments is a post-alveolar or soft consonant (including /j/), or a front vowel: WB *bél* “white” *béli* “white [pl.]” *gréh* “sin” *grexové* “sins,” *gréška* “blunder” versus EB: *bíál* but *béli*, *gríáh* but *grexové*, *gréška*. This feature is correlated with the opposition of hard and soft consonants in that it expands the frequency and the positional characteristics of soft consonants: the opposition hard/soft embraces all consonants in EB but is limited to /k/, /g/, /l/, and /n/ in WB. The development of the so-called etymological *ja* is also to a great extent parallel to that of *jat*: WB *tojágá* “twig” *tojágí* “twigs”; EB: *tojágá*, *to(j)égi*.

WB is further subdivided into northern, southern, and “transitional” (to Torlak Serbian), the main isogloss being the reflex of the back nasal: NWB *pøt*, SWB *pat*, TWB *put* “road.” EB is subdivided into Moesian and Balkan in the northeast and Rupic in the south, the best-known subtype of which is the Rhodope dialect. Basic general resources are BAN (1966–1981) and Stojkov (1993).

31.4.4 Macedonian

Macedonian is divided into three major zones. The northern zone shares many traits with southernmost Torlak, whereas the eastern zone shares many traits with southwestern Bulgarian. The standard language, established in 1945, is based on the central-western zone. Prior to 1945, Serbian linguists viewed Macedonian as “old Serbian,” and many Bulgarian linguists continue to view it as a dialect of Bulgarian. A complex of crisscrossing isoglosses has contributed to this “competition” of views. Still, Macedonian has a number of unique traits, including fixed antepenultimate accent and the “Romance” perfect, composed of the auxiliary verb “have” and a neuter passive participle, for example, *imam bideno* “I have been.” The basic resource is Vidoeski (1998–1999).

31.5 Future Tasks

Future tasks include the continuation of instrumental phonetic research on Slavic dialects and of work on dialect atlases and dictionaries, such as the ideographic dictionary of Bulgarian dialects. Other tasks include continuation of work on, and development of further methodologies for, the study of urban dialects and of dialects in the broader culturological sphere. On another level, the breakup of Yugoslavia into smaller states has introduced a significant national-ethnic component into the region’s dialectology, the future direction of which remains to be seen.

REFERENCES

- Alexander, Ronelle. 2012. Convergence and causation in Balkan Slavic. In *Balkanismen Heute*, edited by Thede Kahl et al., 39–45. Vienna: Lit Verlag.
- Avanesov, R.I., and S.V. Bromlej, eds., 1986–1996. *Dialektologičeskij atlas russkogo jazyka (DARJ I–III)*. Moscow: Nauka.
- Avanesov, R.I., K.K. Krapiva, and Ju.F. Mackevič 1963. *Dyjalektalagičnyj atlas belaruskaj movy*. Minsk: Vydateľstva Akademii navuk BSSR.
- Avanesov, R.I., and V.G. Orlova. 1965. *Russkaja dialektologija*. Moscow: Nauka.
- Balhar, Jan, and Jančák, Pavel. 1992–2005 český jazykový atlas 1–5. Prague: Academia. Online version: <http://cja.ujc.cas.cz/cja.html>. Accessed 15 August 2014.
- Bělič, Jaromír. 1988. *Přehled nářečí českého jazyka*. Prague: Státní pedagogické nakl.
- Bevzenko, S.P. 1980. *Ukraïnska dialektologija*. Kiiv, Višta škola.
- Blinava, E.D. and E.S. Mjacel'skaja 1980. *Belaruskaja dyjalektologija*. Minsk: Vyšejšaja škola.
- Brozović, Dalibor, and Pavle Ivić. 1988. *Jezik, srpskohrvatski/hrvatskosrpski, hrvatski ili srpski. Izvadak iz II. izdanja Enciklopedije Jugoslavije*. Zagreb: Jugoslovenski leksikografski zavod.
- BAN (Bulgarska akademija na naukite). 1966–1981. *Bulgarski dialekten atlas I–IV*. Sofia: Institut za būlgarski ezik.
- BAN (Bulgarska akademija na naukite). 2001. *Bulgarski dialekten atlas, obobštavašt tom*. Sofia: Trud.
- Dejna, Karol. 1993. *Dialekty polskie*, 2d ed. Wrocław: Zakład narodowy im. Ossolińskich.
- Dejna, Karol. 1998–2002. *Atlas gwar polskich I–IV*. Warsaw: Upowszechnianie nauki.
- Durnovo, N.N., N.N. Sokolov, and D.N. Ušakov. 1915. *Opyt dialektologičeskoj karty russkogo jazyka*. Moscow: Trudy moskovskoj dialektologičeskoj komissii.
- Filin, F.P., ed. *Slavar' russkih narodnyx govorov 1–43, 1965–2010*.
- Institut russkogo jazyka. 2010–2012. "Obščeslavjanskij lingvističeskij atlas. Publikacii". <http://www.slavatlas.org/publications.html> Accessed 20 August 2014.
- Ivić, Pavle. 1958. *Die serbokroatischen Dialekte, ihre Struktur und Entwicklung I*. The Hague: Mouton.
- Ivic, Pavle, ed. 1981. *Fonološki opisi srpskohrvatskih/hrvatskosrpskih, slovenačkih i makedonskih govora obuhvaćenih opšteslovenskim lingvističkim atlasom*. Sarajevo: Akademija nauka i umjetnosti Bosne i Hercegovine
- Jutronić, Dunja. 2010. *Od vapora do trajekta, po čemu će nas pripoznati*. Split: Naklada Bošković.
- Kasatkina, L.L., ed. 2005. *Russkaja dialektologija*. Moscow: Akademija.
- Koudela, Břetislav, 1964. *Vývoj českého jazyka a dialektologie*. Prague: československé státní pedagogické nakladatelství.
- Krajčovič, Rudolf. 1988. *Vývin slovenského jazyka a dialektológia*. Bratislava: Slovenské pedagogické nakladatelstvo.
- Langston, Keith. 2006. *Čakavian prosody, the accentual patterns of the čakavian dialects of Croatian*. Bloomington: Slavica.
- Lisac, Josip. 2003. *Hrvatska dijalektologija I. Hrvatski dijalekti i govorovi štokavskog narječja i hrvatski govorovi torlačkog narječja*. Zagreb: Golden marketing.
- Lisac, Josip. 2009. *Hrvatska dijalektologija II. čakavsko narječe*. Zagreb: Golden marketing.
- Logar, Tine. 1993. *Slovenska narečja*. Ljubljana: Mladinska knjiga.
- Logar, Tine, and Josip Rigler. 1983. *Karta slovenskih narečij*. Ljubljana: Cankarjeva založba.
- Lončarić, Mijo. 1996. *Kajkavsko narječe*. Zagreb: Školska knjiga.
- Osenova, Petja, Wilbert Heeringa, and John Nerbonne. 2009. A quantitative analysis of Bulgarian dialect pronunciation. *Zeitschrift für slavische Philologie* 66: 425–458.
- Požarickaja, S.K. 2005. *Russkaja dialektologija*. Moskva, Paradigma.
- Prokić, Jelena. 2010. *Families and resemblances*. Groningen: Groningen series in dissertations 88.
- Sgall, Petr, et al. 1992. *Variation in language: Code switching in Czech as a challenge for sociolinguistics*. Amsterdam: John Benjamins.
- Stankiewicz, Edward. 1957. On discreteness and continuity in structural dialectology. *Word* 13: 44–59.
- Stieber, Zdisław, et al. 1964–1978. *Atlas językowy kaszubszczyzny i dialektów sąsiednich*, Wrocław: Zakład narodowy im. Ossolińskich,

- Stojkov, Stojko. 1993. *Bulgarska dialektologija* 3rd ed. Sofia: Bulgarska akademija na naukite.
- Štolc, Jozef, ed. 1968–1984. *Atlas slovenského jazyka I–IV*. Bratislava: SAV.
- Videnov, Mixail, and Bojan Bajčev. 1999. *Veliko-Túrnovskijat ezik: Sociolinguisticko proučvane na velikotúrnovskata gradska reč*. Veliko Túrnovo: Abagar.
- Vidoeski, Božidar. 1998–1999. *Dijalektite na makedonskiot jazik I–III*. Skopje: Macedonian Academy of Sciences and Arts.
- Zakrev’ska, Ja. V., and I.H. Matvijas. 1984–2001. *Atlas ukraïns’koї movi I–III*. Kiev: Naukova dumka.

32 Dialects of Arabic

ENAM AL-WER AND RUDOLF DE JONG

32.1 Introduction

Arabic is a member of the Semitic language family, which in addition to Arabic includes ancient languages such as Akkadian, Ugaritic, Phoenician (all extinct), Hebrew, and Aramaic.

The Semitic language family is itself a branch of a larger family called “Afro-Asiatic” (formerly known as “Hamito-Semitic”), whose other branches include Ancient Egyptian (the language of the pharaohs and its descendant, Coptic, now extinct), Berber (languages of North Africa), the Chadic languages (e.g., Hausa) and the Cushitic languages of northeast Africa (e.g., Somali).

Arabic is spoken by approximately 300 million people distributed over 22 different countries in southwest Asia and North Africa, collectively known as “the Arab world,” which stretches from the Arabian (or Persian) Gulf in the east to the Atlantic Ocean in the west, and from the Mediterranean Sea in the north to the Horn of Africa in the south (see map in Figure 32.1).

The region is ethnically and linguistically diverse. Among the non-Arabic indigenous languages that continue to be spoken in various parts are Kurdish (Iraq and Syria), varieties of Berber (North Africa), Aramaic (small communities, primarily in Syria and Iraq) and Armenian, Chechen, and Adyghe (Circassian) in the Levant. Outside the Arab world, varieties of Arabic are spoken by minority groups in Iran, Afghanistan, Turkey, Uzbekistan, Cyprus, and Nigeria, among others. As a liturgical language, Classical Arabic, the language of the Quran, the holy book of Islam, is used for prayer by Muslims the world over.

The homeland of Arabic is the Arabian Peninsula and the Syrian steppe. Everywhere else, Arabic arrived mainly through the Arab Islamic incursions, which began during the second half of the seventh century. It spread north to the eastern Mediterranean region, east to ancient Mesopotamia (Iraq), west to the Nile valley (Egypt and the Sudan), and farther to Roman Africa (today’s Libya, Tunisia, Algeria, and Morocco) and the Iberian Peninsula (Andalusia). By the thirteenth century, it had been transformed from a localized language used in a confined territory into a language fit for literature, science, and the administration of a transcontinental empire.

Arabic began to recede geographically and to decline in importance with the Mongol sacking of Baghdad in 1258, the withdrawal of the Arabs from Andalusia after the fall of Granada in 1492, and later as the leadership of the Islamic empire was taken over by the Ottoman Turks during the fourteenth century. By the beginning of the twentieth century the use of standard Arabic had become restricted to a small minority of elite intellectuals, mainly in the eastern Mediterranean region (the Levant), Iraq, and Egypt. Nonetheless, despite nearly five centuries

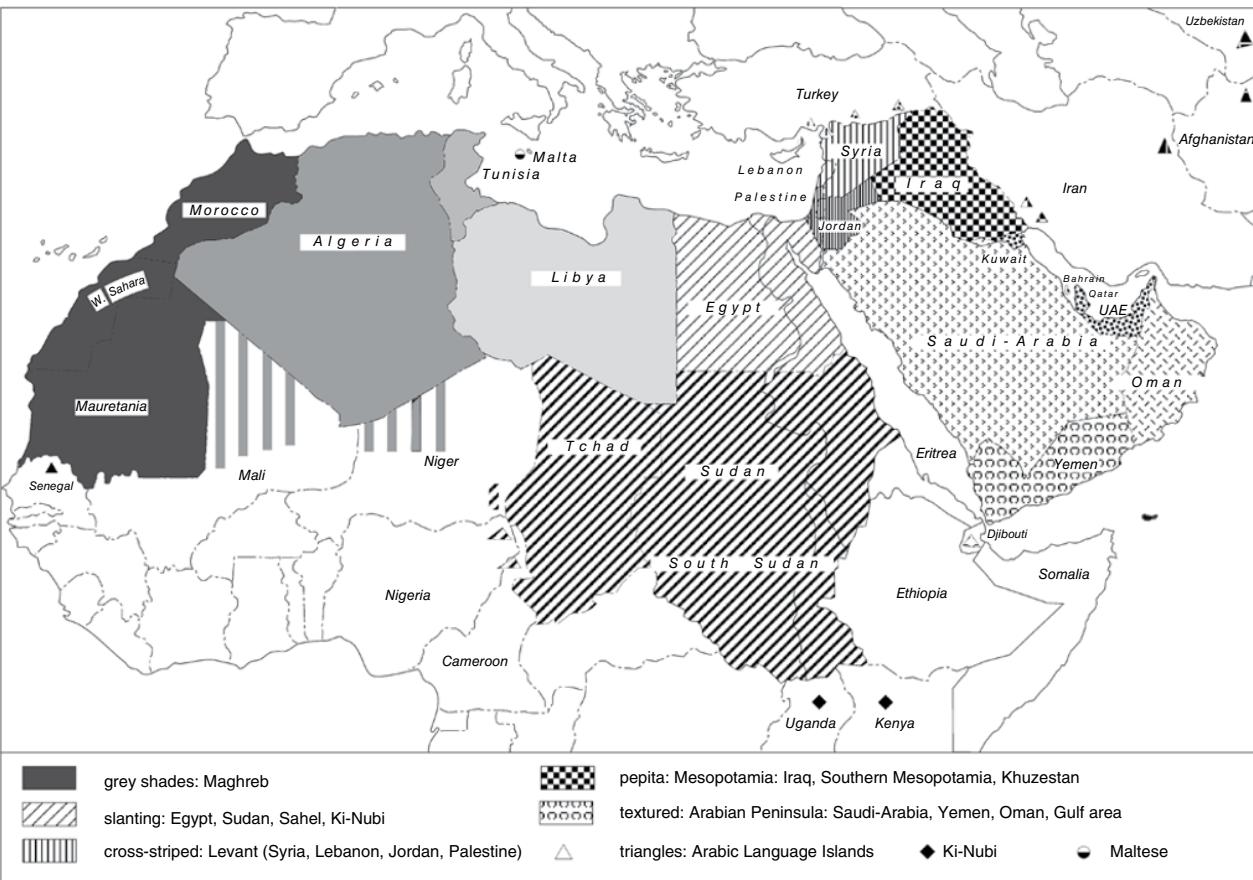


Figure 32.1 Map showing the Arabic-speaking regions and the status of Arabic in the various countries (courtesy of Peter Behnstedt).

of Ottoman rule and the decline of standard Arabic in general, Arabic vernaculars (the spoken dialects, see below) continued to be the mother tongues and the normal medium of communication for the majority of the population, especially in the eastern part (Egypt, the Levant, Iraq, and Arabia). In North Africa, French colonization has had a long-lasting linguistic effect to this day. In these countries (collectively called the *Maghreb*, lit. “the west”), French has encroached on all domains, including those of the local vernaculars (Arabic and Berber). Although standard Arabic (and also Berber in Morocco) is the official language in North African countries, French continues to play a functional role and enjoys considerable social prestige.

32.2 Varieties of Arabic in Contemporary Times

In modern Arabic-speaking societies, Arabic exists in two norms: the standard variety and the vernaculars. In this section, we provide broad definitions, and explain the functions of each norm and pertinent sociolinguistic issues.

32.2.1 Standard Arabic

The term “standard Arabic” is used in this chapter to refer to two forms of the standardized norm: (i) Classical Arabic as it was codified by the ancient philologists in the eighth century; (ii) Modern Standard Arabic (common acronym MSA), which is structurally the same as Classical Arabic, but contains many borrowings from modern European languages, for example, /film/ “film,” and modern coinages, for example, /ha:tif/ “telephone.” Subject-initial sentences are certainly more common in MSA than in Classical Arabic (see Holes 2004).

Written standard Arabic is used in an almost invariant form throughout the Arab world, although regional differences in lexical choices and various idiomatic coinages do occur. Spoken standard Arabic, by contrast, is influenced by the speakers’ mother tongues (the vernacular) in pronunciation and supra-segmental features (stress, intonation).

The linguistic situation in countries where Arabic is the official language is usually described as diglossic (Ferguson 1959). In this situation, the standard variety is nobody’s first acquired language. It is not transmitted naturally from parent to child, but learnt later on through formal instruction of some sort.

Standard Arabic is the language of literature, official documents and the written formal media in general (newspapers, instruction leaflets, school books, etc.). In spoken form it is most consistently used when reading from a scripted text (e.g., news bulletins). It is also the variety used for prayer and sermons in the mosque or church. It is the only variety that is taught in schools as *the* Arabic language. In institutions of higher education, it is used as a medium of instruction in the social sciences and humanities. In most universities, English or French are used in the applied and medical sciences. Notwithstanding the relatively limited domains of usage of the standard variety in everyday life and for ordinary interactions, it has a strong psychological claim on Arabs. As a transnational norm, it is perceived as an important element of unity and solidarity among Arab nations. As the language in which the ancient treatises are written it symbolises a glorious past and as the language of the Quran it is considered divine by many.

32.2.2 Vernacular Arabic

Arabic dialects are the mother tongues of Arabic speakers. They are the first acquired languages in the home, except among ethnic groups who maintain their heritage languages. They are the normal medium of communication in all domains except those described above, which require the standard.

Although there are no conventionalised writing systems for Arabic vernaculars, they feature in many literary works (novels, poetry) and in the arts (lyrics, films, soap operas). They are also widely used in a written form in the electronic social media, cartoons, and advertisements.¹

Departments of Arabic in Arab universities do not generally cater for research on Arabic dialects, which reflects widely held beliefs in the Arab world that the dialects are but a debased offshoot of standard Arabic and studying them is a sign of corruption.

Arabic dialects in various regions of the Arab world have undergone natural processes of standardization, culminating in the emergence of local and regional (koineized) standards. In many cases, the dialect of the capital city emerged as a *de facto* local standard, as with Cairo Arabic, which represents standard Egyptian Arabic. The same is true of the dialects of Baghdad, Damascus, Beirut, Amman, and Tunis. In the Levant, natural standardisation includes city dialects in the region as a whole, thus transcending political borders (Syria, Lebanon, Jordan, and Palestine) (see Al-Wer 2007). In Arabian dialects, a Gulf-wide linguistic norm exists, covering the major cities in the region (Manama, Kuwait, Doha, Abu-Dhabi, etc., see Holes 1990). Ongoing research indicates similar developments in Saudi Arabia, where two major standards are emerging, one based on the Riyadh dialect in the central region and another based on the dialect of Jeddah in the western region.²

Linguistic variation and change in spoken Arabic is structured by an interaction between linguistic and social variables, with constraints dictated by the respective native Arabic dialects of each community, not by the structures or features found in the standard variety. A common finding of variationist sociolinguistic research on Arabic dialects is that the local standards, rather than standard Arabic (the superposed high variety), play the pivotal role in the structure of variation in spoken dialects, and that the trajectory of change in vernacular Arabic is in the direction of features associated with the local standards. Therefore, whereas standard Arabic clearly has a function in Arabic-speaking societies, it does not have a normative effect on variation in the core domains of phonology, morphology and syntax (Al-Wer 2013).

32.3 Arabic Dialect Geography

The usual way to classify Arabic dialects is by region, and within regions reference is made to two linguistic norms: “Bedouin” and “Sedentary.” In this section we begin by defining these norms briefly, and move on to present a more detailed classification according to geography. The frame of reference in the discussion of dialects will be a comparison with Classical Arabic, henceforth abbreviated as CA.

32.3.1 Bedouin Versus Sedentary

Throughout the Arabic speaking world we can distinguish between “Bedouin” and “sedentary” dialects. Bedouin dialects are akin to a norm believed to have been characteristic of the dialects spoken by nomadic tribes; sedentary dialects are akin to dialects that were spoken by settled communities (rural and urban). It is important to stress that the use of these terms in Arabic dialectology does not refer to contemporary lifestyle; most of the Bedouin dialects nowadays are spoken by settled and often highly urbanized populations.

Among the prominent characteristics of Bedouin dialects are a voiced “reflex”³ of CA *q, often /g/; interdental reflexes /θ/ and /ð/ for CA *θ and *ð; and /ðˤ/ as the reflex for both CA *ðˤ and *dˤ (where sedentary dialects have /t/, /d/ and /dˤ/, respectively). In morphology,

Bedouin dialects often maintain a grammatical gender distinction (fem./masc.) in second and third person plural forms of verbs and of pronominals.⁴

In the remainder of this section, we deal with the geographical classification of Arabic dialects, focusing on the most salient features of each group. This classification is based on sets of isoglosses, which may vary from region to region.

32.3.2 West Versus East

Dialectologists divide Arabic dialects into two geographical super groups: eastern and western. These are further subdivided into five main regions:

1. North Africa, or dialects of the Maghreb (including Mauritania, Morocco/Western Sahara, Algeria, Tunisia, and Libya);
2. Egypt and the Sudan, usually referred to as Egyptian dialects;
3. The Arabian Peninsula (including Saudi Arabia, Yemen, Oman, and the Gulf states);
4. Mesopotamia (Iraq and the Iranian province of Khuzestan);
5. The Levant, or Syro-Lebanese dialects (Syria, Lebanon, Palestine and Jordan).

The main criteria distinguishing western dialects (spoken from Egypt to North Africa) from eastern dialects (spoken from Egypt to western Asia) involve morphology. Egypt is a transitional area between the western and eastern regions.

32.3.2.1 Western Dialects

The most important morphological isogloss concerns the first person singular and plural in the verbal imperfect. To the west of this isogloss the first person singular is formed with initial *n* (+ vowel) - + verb stem. To the east the singular is formed without this initial *n*-, but with an initial vowel (often *a*-). In the plural, the first person is *n* (+ vowel) - + verb stem + -*u* in the west, whereas to the east it is formed without the final -*u*.⁵ There is a narrow area of transition between these forms in the western Nile Delta. These differences are illustrated below, using the verb *katab* "to write."

	<i>west</i>	<i>transitional area</i>	<i>east</i>
First p. sg.	<i>ni-ktib</i>	<i>a-ktib</i>	<i>a-ktib</i> "I write"
First p. pl.	<i>ni-ktib-u</i>	<i>ni-ktib-u</i>	<i>ni-ktib</i> "We write"

Another distinctive characteristic of western dialects, especially Algerian and Moroccan dialects, is the (partial) lack of phonemic distinction between short vowels. Some dialectologists recognize only one short vowel phoneme, often written as /ə/. Additionally, in many dialects of the western Maghreb, even the phonemic status of vowel length is not recognized (for a discussion, see Behnstedt and Benabbou 2002).

In North Africa, we can still recognize differences between dialects that date back to the different waves of Arabisation in history: pre-Hilali⁶ dialects (from the seventh century), which were spoken by settled tribes originally from the Peninsula (therefore called "sedentary" dialects); and Hilali dialects (tenth and eleventh centuries), which were spoken by nomadic tribes (thus known as "Bedouin" dialects). Pre-Hilali (sedentary) dialects tend to have a voiceless reflex of CA *q, for example, /q/ or /ʔ/, whereas Hilali (Bedouin) dialects have a voiced /g/ for *q. A special case is Hassaniyya, the dialect spoken in Mauritania, which is also a Bedouin dialect but shows a number of innovations not found in other dialects. Over the centuries, all of these dialects have to a greater or lesser extent undergone influences from Berber languages (see Versteegh 1997).

32.3.2.2 Eastern Dialects

From the eastern Nile Delta of Egypt, a transition zone comprising the Bedouin dialects of the northern Sinai Desert forms a bridge to the dialects of the east. These dialects link up to those of the Hejaz, the western province on the Red Sea coast of present-day Saudi Arabia⁷. The major criteria that distinguish Sinai and Hejaz dialects from those of the central Nile Delta are phonological. The salient distinctions are:

<i>CA</i>	<i>Sinai</i>	<i>Central Delta</i>	<i>Example</i>
*q	/g/	/ʔ/	<i>galb/ʔalb</i> “heart”
*ʒ	/dʒ/~/ʒ/	/g/	<i>dʒamal~ʒamal/gamal</i> “camel”
*θ, *ð	/θ/, /ð/	/t/, /d/	<i>θala:θa/tala:ta</i> “three,” <i>ðe:l/de:l</i> “tail”
*ðˤ, *dˤ	/ðˤ/	/dˤ/	<i>ðˤarab/dˤarab</i> “he hit”

In addition to phonological differences, these dialects can be distinguished by maintenance versus neutralization of gender distinctions in the plural (see Bedouin above): for example, *ðˤarabaw* “they (masc.) hit”/*ðˤaraban* “they (fem.) hit”; whereas the (sedentary) central Delta dialects have a *communis* form for both genders, thus, *dˤarabu* “they (masc. or fem.) hit.”

The Northwestern Bedouin group is hypothesized to continue south through the Hejaz province and along the Red Sea coast toward Tihama province (see De Jong 2000). It is distinguished from the Northeastern Bedouin Arabian group, commonly known as Najdi dialects, which are spoken in the central peninsula and in the Gulf countries; several Bedouin dialects spoken in Iraq and the Levant are related to this large group. In morphology, Najdi dialects, similarly to CA, preserve the verbal imperfect ending *-n*, for example, *tifmali:n* “you (second fem sg) do”; *tgu:lu:n* “you (second masc pl) say.” In other dialects these forms are without final *-n*, thus *tifmali*, *tgu:lu*. Najdi dialects are generally more conservative than other Bedouin dialects, but this should not be understood to mean that they lack innovations. A good example of an innovative Najdi feature is *Najdi resyllabification*: a phonotactic constraint prohibiting the appearance of strings CaCaCv (C=consonant, a = /a/, v=any long or short vowel). Such strings are resyllabified to become CCICv (I=short vowel /i/, /u/ or /a/ depending on surrounding sounds): for example, (*katab + -aw* →) *katabaw* → (Najdi) *ktibaw* “they (masc.) wrote” (see Ingham 1982).

Najdi and the Northwestern group share an innovation called the *gahawah-syndrome* (lit. “coffee-syndrome”), a phonotactic constraint prohibiting a string CaXC(v)(C) (X=any back spirant h, ħ, ʕ, x or y); the string will be resyllabified to become CaXaC(v)(C). In Najdi, the output of the *gahawa-syndrome* is then subjected to *Najdi resyllabification* (see above; in generative terms, this is a “feeding” relationship). Thus, *gahwah* → *gahawah* (Northwestern surface form; intermediate form in Najdi) → (Najdi) *ghawah* “coffee.”⁸

Another distinctive Najdi innovation, not found in Northwestern dialects, is palatalization or affrication of stops: /g/ → [g̊] or [d̊z], and /k/ → [t̊], for example, *d̊ili:l* (<*gili:l*) “little, few” and *be:tit̊s* (<*be:tik*) “your (fem. sg.) house.” Related Bedouin dialects farther north (Iraq, Syria) affricate /g/ > [dʒ] and /k/ > [tʃ]. This difference in the output of palatalization is one illustration of the different subgroups in Najdi, namely [g̊]/[d̊z] and [t̊] in the Arabian Peninsula, versus [dʒ] and [tʃ] in Syrian and Mesopotamian Najdi varieties.

In some Gulf countries, sedentary and Bedouin dialects can be distinguished along the voiced-voiceless /q/ criterion (see above). For instance, some sedentary dialects in villages in the mountainous interior of Oman have /q/ or /k/, but Bedouin dialects in Oman have (voiced) /g/. However, all dialects of Oman, both sedentary and Bedouin, have interdental reflexes [θ] and [ð] for CA *θ and *ð, and [ðˤ] for CA *ðˤ and *dˤ.⁹

A further subdivision in the Arabian Peninsula is the group of Southwest Arabian dialects (Yemen, the Hadramawt, and Aden). The dialect of the Shi'ite population of Bahrain is

related to this group. The dialect divisions of Yemen are highly complicated and extremely diverse, with at least 16 different zones being identified (Vanhove 2009, "Yemen," EALL). One of the major features in verbal morphology found mainly in the mountain range in the west of the country is the "k-perfect" (where other dialects have a "t-perfect"). For example, "I wrote," which in other dialects is *katabt* (or in nearby dialects *katabtu*), may in this western mountainous region be *katabku*, *katabk^w*, *katubk*, *katabuk*, *katbuk* or *katabk*. Another feature is a rather unique gender distinction in the first person singular pronoun, *ana* "I (masc.)" versus *ani* "I (fem.)," which is found in western Yemeni dialects and in continuations of these in the central dialects. In dialects of the Hadramawt, a [j] reflex for Classical Arabic *j (/dʒ/) is found (as in the major dialects of the Gulf). Mutual intelligibility, even between direct neighbors, is often a problem in this region (see Vanhove 2009, "Yemen," EALL).

32.3.2.3 Mesopotamia

The dialects of Mesopotamia can be subdivided into two major groups, which are named after their variant forms of "I said": the *gilit* (or *gələt*) dialects and the *qəltu* dialects. The *gilit* dialects are Bedouin (or of Bedouin origin) and are continuations of the Bedouin dialects of the Arabian Peninsula. These can be further subdivided into two groups: (i) *gilit* spoken in rural areas of northern and central Iraq, in the Sunni area around Baghdad and by Muslims in Baghdad; and (ii) *gilit* spoken in rural areas of southern Iraq and by Muslims in urban areas (see Jastrow 2007, "Iraq," EALL). The *qəltu* dialects are akin to the sedentary norm, and can be further subdivided into three groups: the Tigris group, the Euphrates group, and the Kurdistan group. The Tigris group is spoken in Baghdad (by Christians and in earlier times by Jews) and to the north of Baghdad (in and around Tikrit by Muslims). To the south of Baghdad, it is spoken by Christian communities (and in former times also Jewish communities).¹⁰ The Euphrates group is spoken in Hi:t and ſA:na by Muslims and (formerly also) by Jews. The Kurdistan group comprises Jewish communities in the otherwise Kurdish-speaking northeast of Iraq.

Most Iraqi dialects have an extra phoneme /p/ (not often present in Arabic dialects), which is attributed to influences from Turkish, Persian and Kurdish, and English. Most Iraqi dialects (of both the *gilit* and the *qəltu* type) have typically Bedouin interdental reflexes for CA interdentals *θ and *ð and *ðˤ/*dˤ (see above), although in Christian Baghdadi (and some Jewish dialects of Kurdistan), reflexes for these are plosives /t/, /d/ and /dˤ/, respectively. In the dialect of Basra, the common reflex for CA *j is /y/ ([j]), similarly to several Gulf dialects (see Jastrow 1978).

The first thorough sociolinguistic investigation of Arabic dialects, *Communal Dialects in Baghdad*, written by Haim Blanc (1964), is a study of the dialects of Baghdad, as spoken by the three different religious communities of Muslims (of the *gilit* type), Christians, and Jews¹¹ (both of the *qəltu* type). Dialectal differences along sectarian/religious lines are found elsewhere in the Arab world (e.g., Bahrain and the island of Djerba, off the coast of Tunisia, see Behnstedt and Woidich 2005), but it is not the rule that speakers of different religious backgrounds speak different dialects. In most cases in Egypt and the Levant, for instance, dialect differences between Muslims and Christians are minimal, involving some vocabulary items, mainly in the religious realm, or non-existent.

32.3.2.4 Egyptian Dialects

In terms of numbers of Arabic speakers, Egypt is by far the largest country in the Arab world. The vast majority of Egyptians (currently nearing 90 million) live in the Nile Delta and Valley. The dialects of Egypt can be subdivided into rural, urban, and Bedouin dialects. As stated earlier, the Nile Delta, in a sense, forms a transitional area between the larger western and eastern groups of Arabic dialects. Bedouin dialects are spoken in the western Delta and farther west (for example by the Awla:d ſAli tribe along the Mediterranean coast) and in the eastern Delta and farther east (for example, the tribal dialects of the Sinai, which form the

transition to the dialects of the Arabian Peninsula). Other Bedouin dialects are found throughout Egypt, mostly in desert areas and on the fringes of cultivated land.

Egypt's rural dialects can be divided into three major groups: central Delta dialects, like the dialect of Cairo; Middle Egypt (with a northern and a southern group); and Upper Egypt (with four subdivisions roughly from north to south; see Behnstedt and Woidich 1985). The dialect of Cairo is perhaps the best-known (and most widely understood) Arabic vernacular. This is largely due to Egypt's film and music industries, which have flourished for decades, serving audiences throughout the Arab world.

Some of the major phonological variables include the reflexes of CA *j ([dʒ]) and *q: in Cairo, the central Delta, to the south of Cairo (to approximately 20 km south of Bani Swayf) and in the Fayyoun oasis, their reflexes are /g/ (for *j), which is in fact older than the *j of CA itself,¹² and (voiceless) /?/ (for *q) respectively. In the western and eastern parts of the Delta (and farther west and east) and to the south of Bani Swayf, we find the more typically Bedouin reflexes /j/ (i.e., [dʒ]) (for CA *j) and /g/ (for CA *q). Farther south, the reflex for *j can be /d/ (from around Asyout down to north of Aswan), for example, *dardal* for "bucket" (<*jardal*), or /dʒ/, with minimal friction, in Upper Egypt.

In many locations south of Asyout we find the *gahawah-syndrome* (see above), which is another indication of Bedouin influence, and in the far south a gender distinction in plural forms of the second and third person also bears witness to Bedouin influences.

From the bend in the Nile at Qena and farther south, as well as in the Bahariyya and Farafra oases in the Western Desert of Egypt, there are several locations with the western paradigm for the first person singular and plural in the imperfect: *niktib* "I write" and *nikt(i)bu* "we write," or the "intermediate" paradigm *aktib* (sg.) and *nikt(i)bu* (pl.) (Behnstedt and Woidich 1985). Dialectologists disagree on whether these paradigms were "imported" into Egypt by speakers of Maghrebi dialects who arrived from the west, or actually originated in Egypt (Woidich 1993; Behnstedt 1998; Owens 2003). The possibility of their being of pre-Islamic origin cannot be ruled out either (Manfred Woidich, p.c.).

The Arabic dialect of the ſAba:bda tribe, who are originally speakers of Beja (a Cushitic language) and who live in the southeastern desert of Egypt and northeastern Sudan, is quite noticeably of a Sudanese Arabic type (De Jong 2002). Arabic dialects of Sudan arrived there by way of the Nile from Egypt from the ninth century onward. From Sudan, Arabic was spread farther west by cattle-raising nomadic tribes into what are now Chad, Cameroon, and Nigeria in western Africa (the so-called *bagga:ra* belt, lit. "the cow herders belt"). Part of the population who brought Arabic to Chad are reported to have come from the north, following caravan routes through the Sahara from the Mediterranean (speakers of Hassa:niyya Arabic from Libya). As a result of contact with the languages spoken locally, radical changes have taken place in the inventories of consonants. In Chadian Arabic, for instance, no less than 10 CA consonantal phonemes are lacking, while four new consonantal phonemes have been added to the inventory. The sedentary dialects of Sudan have no interdental sounds, and the nomadic dialects have a reflex /g/ for CA *q.¹³

In the central Sudan and larger towns Sudanese colloquial Arabic is spoken, which is also known as Khartoum Arabic. This variety has incorporated many vocabulary items from Beja (Cushitic) and Nubian (Nilo-Saharan). The consonantal phoneme inventory lacks interdentals, and CA *q may have a reflex /g/, /k/, and in some cases the voiced velar fricative /ɣ/. A feature of nominal morphology is the final stress on the long vowel of the suffixes -i: and -ni: (respectively, the 1sg. possessive and object suffixes; see Abu Manga 2009, "Sudan," EALL).

32.3.2.5 Syro-Lebanese Dialects of the Levant

According to a recent study, the Syrian steppe is actually the first region where Arabic has been attested (in Safaitic inscriptions, see Al-Jallad 2012). With the advent of Islam, the entire

Levant became the new home of Arabic speakers originating from the Arabian Peninsula as well, so that Aramaic (also a Semitic language), which had been widely spoken until then, gradually declined and all but disappeared, nevertheless leaving substrate influences on local Arabic dialects (see Kusters 2009, "Substrate," EALL).

In the Naqab Desert (Negev) and southern Jordan, Bedouin dialects of the Northwestern group are spoken, which link up with the dialects of Sinai (De Jong 2000; 2011). In Syria, the sedentary dialects are spoken in the west, whereas the Bedouin dialects (of the Najdi and Sha:wi types) are found in the eastern desert. In the far northeastern corner of Syria, we find dialects of the *qaltru* type (related to those of northern Iraq and Anatolia) (see Jastrow 1978, 1981; Behnstedt 2009, "Syria," EALL.). Nevertheless, most of the dialects spoken in the Levant are of the sedentary type: CA interdentals have plosive reflexes and reflexes of CA *q are voiceless (/?/ in Beirut, Damascus, Jerusalem). In Amman we find voiced /g/ for CA *q, but this /g/ tends to characterize the speech of men, whereas women use /?/ instead (Al-Wer and Herin 2011). In Druze dialects, /q/ tends to be the reflex for CA *q, and in rural Palestinian its reflex is predominantly /k/ or more back (uvular/velar) /kⁱ/ . Also in rural Palestinian, CA *k has a palatalized reflex [tʃ]. City dialects in the Levant lack the gender distinction in plural verb forms and pronominals.

At a general level, the Levant can be subdivided into three main dialect groups (Versteegh 1997):

- Lebanon and central Syria (including those of Beirut and Damascus);
- Northern Syria (including that of Aleppo);
- Palestine and Jordan (including the Horan region).

Northern coastal dialects stretch from Lebanon through Syria into the Turkish Hatay province; more are found along the Mediterranean Turkish coast as language islands in Cilicia and Antiochia (Antakya).

In phonology, a distinctive feature of northern Syrian dialects is extreme raising of the CA long vowel *a: to /e:/ or /i:/. Another striking feature of vowel phonology is the presence of /o:/ (for *a:), whereas /a:/ is a reflex of *ay (in the dialect of Arwad, an island off the coast of Tartous, Syria) (see Procházka 2013).

In Lebanese dialects, the presence of the diphthongs /ay/ and /aw/ is distinctive. Not only can these reflect CA *ay and *aw, but they also occur as a result of the diphthongization of *i: and *u: (especially near Zahle in the Beqa:f Valley, Lebanon): for example, *m̥mayħ* (<*mniħ*) "good, well" and *mabsawt* (<*mabsu:f*) "happy" (see Wardini 2014, "Lebanon," EALL online).

Palestinian dialects can be subdivided into urban, rural, and Bedouin types, the latter two having interdentals, whereas urban Palestinian has plosives for CA interdentals (Shahin 2008, "Palestinian Arabic," EALL).

32.3.3 Language Islands

Arabic "Language islands" are found in Uzbekistan, Afghanistan, Iran, and Turkey (also Cyprus, of Levantine origin). As a result of their isolation, they often show archaic features, but also unique innovations, which may be due to isolation and to contact with other languages (see, e.g., Jastrow 2011). Maltese is originally an Arabic dialect of North African origin, which came to Malta via Sicily during the four centuries around 1000 CE and in which many influences (notably in vocabulary) from other Mediterranean languages (Sicilian, Italian) as well as of English are traceable. Its codified form has become the official language of Malta, which has its own orthography in the Latin alphabet (Mifsud 2008, "Maltese," EALL).

32.3.4 Creoles

Of special interest for Creole linguistics are varieties of Arabic that emerged as a result of contact with African languages (Lur, Lemdu, Bari, Moro, Bangala, Swahili, among others), such as Ki-Nubi and Juba Arabic spoken in central and eastern Africa (Owens 2006, “Creole Arabic,” EALL; Tosco and Manfredi 2013), and other varieties in the western Sudan and farther west into the *bagga:ra* belt.¹⁴

Additional Resources

In addition to the works cited in the main text, the reader is advised to consult the bibliographies of the lemmas in the *Encyclopedia of Arabic Language and Linguistics* (EALL, also available online), as well as Owens (2006), Retsö (2003, 2006, 2013) and Huehnergard, and Rubin (2011).:

The most important linguistic atlases and volumes that include linguistic maps of various regions include Arnold (1998), Arnold and Behnstedt (1993), Behnstedt (1985a, 1985b, 1997), Behnstedt and Woidich (1983, 1985-99, 2005, 2010, 2012, 2014), Bergsträßer (1915) and Cantineau (1940).

NOTES

- 1 The Arabic writing system lacks graphemes for some of the sounds that occur in the dialects exclusively, for example, [g], [ʃ]. In the social media especially, the Latin script is often used to write the dialects; in this case numerals are used to compensate for absent graphemes, for example, 7 is used to represent [ħ], 9 for [ʃ], 2 for [?] . The choice is based on similarity of the number and the Arabic grapheme that stands for that sound, for example, 7 stands for the Arabic grapheme *ڇ*.
- 2 Sociolinguistic research in five different locations in Saudi Arabia is currently in progress at the University of Essex, UK.
- 3 We speak of a “reflex,” since we cannot prove that the new phoneme is a direct further development of the phoneme (here marked with *) found in Classical Arabic; we only claim that the new phoneme “reflects” the CA phoneme. CA is thus only used as a frame of reference and the asterisk does *not* indicate a proto-form here.
- 4 For information about further features that separate these norms, see Versteegh 1997.
- 5 There are, however, several Upper Egyptian dialects (considered “eastern”) that have the same North African (i.e., western) imperfect paradigm (Behnstedt and Woidich 1985).
- 6 The name refers to an Arab tribe, the Banu Hilal, who invaded Egypt and North Africa.
- 7 This group of Bedouin dialects has been named Northwest Arabian Arabic, see Palva 2008, “Northwest Arabian Arabic,” EALL.
- 8 See De Jong 2007, “Gahawa-syndrome,” EALL.
- 9 See Holes 2008, “Omani Arabic,” EALL. N.B. The author of this lemma is erroneously printed as “Lutz Edzard.”
- 10 As a result of the wars and the social and political unrest that has plagued the country over the past decades, many religious minorities (as well as large numbers of the Sunni and Shi’i majorities) have emigrated from Iraq.
- 11 Jewish Arabic dialects are often interesting since they show deviations from surrounding dialects everywhere and often represent very old and conservative types of Arabic. This is especially so in the Maghreb, but also in Egypt and Iraq.
- 12 See Woidich and Zack 2009; Zeroual, 2008, “Palatalization,” EALL.
- 13 Versteegh 1997; Jullien de Pommerol 2006, “Chad Arabic,” EALL; Abu Manga 2009, “Sudan,” EALL.
- 14 Owens 2006, “Creole Arabic,” EALL; Miller 2007, “Juba Arabic,” EALL; Wellens 2007, “Ki-Nubi,” EALL.

REFERENCES

*Note: because of space limitation, references to lemmas published in the *Encyclopedia of Arabic Language and Linguistics* (EALL) are cited in the main text under author, year and title but not listed below.

- Al-Jallad, Ahmad. 2012. *Ancient Levantine Arabic: A Reconstruction Based on the Earliest Sources and the Modern Dialects*. Ph.D. dissertation, Harvard University.
- Al-Wer, Enam. 2007. The formation of the dialect of Amman. In *Arabic in the City*. Catherine Miller, Enam Al-Wer, Dominique Caubet, Janet Watson (eds). New York: Routledge. 55–76.
- Al-Wer, Enam. 2013. Sociolinguistics. In Jonathan Owens (ed.) *Handbook of Arabic Linguistics*. 241–263.
- Al-Wer, Enam and Herin, Bruno. 2011. The lifecycle of *Qaf* in Jordan. *Langage et Société*, 138, 59–76.
- Arnold, Werner. 1998. *Die arabischen Dialekte Antiochiens*. Wiesbaden: Harrassowitz.
- Arnold, Werner and Behnstedt, Peter. 1993. *Arabisch-Aramäische Sprachbeziehungen im Qalamun (Syrien)*. Wiesbaden: Harrassowitz.
- Behnstedt, Peter. 1985. *Die nordjemenitischen Dialekte. Teil 1: Atlas*. Wiesbaden: Ludwig Reichert.
- Behnstedt, Peter. 1997. *Sprachatlas von Syrien. Semitica viva* 17. Wiesbaden: Harrassowitz.
- Behnstedt, Peter. 1998. La frontière orientale des parlers maghrébins en Egypte. In Jordi Aguadé (ed.) *Peuplement et Arabisation au Maghreb Occidental*. 85–96.
- Behnstedt, Peter and Benabbou, Mostafa. 2002. "Zu den arabischen Dialekten der Gegend von Táza (Nordmarokko)". In Werner Arnold and Hartmut Bobzin (eds) *Sprich doch mit deinen Knechten aramäisch, wir verstehen es!* 60 Beiträge zur Semitistik, Festschrift für Otto Jastrow zum 60. Geburtstag. Wiesbaden: Harrassowitz. 62–65.
- Behnstedt, Peter and Woidich, Manfred. 1983. A VIII 12 Ägypten. *Arabische Dialekte: 1:1 Mio. (Tübinger Atlas des Vorderen Orients (Tavo))*. Wiesbaden: Ludwig Reichert (map edition).
- Behnstedt, Peter and Woidich, Manfred. 1985–99. *Die ägyptisch-arabischen Dialekte* (5 vols.). Wiesbaden: Reichert.
- Behnstedt, Peter and Woidich, Manfred. 2005. *Arabische Dialektgeographie, eine Einführung*. Leiden: Brill.
- Behnstedt, Peter and Woidich, Manfred. Wortatlas der arabischen Dialekte, 4 vols: 2010 (I), 2012 (II), 2014 (III), due to appear (IV). Leiden-Boston: Brill.
- Bergsträßer, Gotthelf. 1915. *Sprachatlas von Syrien und Palästina. Zeitschrift des Deutschen Palästina-Vereins* 38 Leipzig.
- Blanc, Haim. 1964. *Communal Dialects in Baghdad*. Cambridge (Mass.): Harvard University Press.
- Cantineau, Jean. 1940. *Les parlers arabes du Hôrân* (Collection linguistique 49). Paris: Klincksieck.
- De Jong, Rudolf. 2000. *A Grammar of the Bedouin Dialects of the Northern Sinai Littoral, Bridging the Linguistic Gap between the Eastern and Western Arab World*. Leiden–Boston–Köln: Brill.
- De Jong, Rudolf. 2002. "Notes on the Dialect of the Ababda". In W. Arnold and H. Bobzin (eds) *Sprich doch mit deinem Knechten aramäisch, wir verstehen es: Festschrift für Otto Jastrow*. Wiesbaden: Harrassowitz. 337–359.
- De Jong, Rudolf. 2011. *A Grammar of the Bedouin Dialects of Central and Southern Sinai*. Leiden–Boston: Brill.
- EALL online. *Encyclopedia of Arabic Language and Linguistics*. Lutz Edzard and Rudolf de Jong (gen. eds), Ramzi Baalbaki, James Dickins, Mushira Eid, Pierre Larcher, Janet Watson (ass. eds). Leiden: Brill. (Online continuation of EALL, includes all lemmas of (printed) EALL).
- EALL. 2006 (A-Ed); 2007 (Eg-Lan); 2008 (Lat-Pu); 2009 (Q-Z) (four volumes + index vol.).
- Encyclopedia of Arabic Language and Linguistics*. Kees Versteegh (gen. ed.), Mushira Eid, Alaa Elgibali, Manfred Woidich, Andrzej Zaborski (ass. eds). Leiden–Boston: Brill.
- Ferguson, Charles. 1959. The Arabic Koine. *Language* 35/4: 616–630.
- Holes, Clive. 1990. *Gulf Arabic*. London: Routledge.
- Holes, Clive. 2004. *Modern Arabic: Structures, Functions and Varieties*. Revised Edition. Georgetown Classics in Arabic Language and Linguistics Series. Washington D.C.: Georgetown University Press.
- Huehnergard, John and Rubin, Aaron. 2011. Phyla and Waves: Models of Classification of the Semitic Languages. In Stefan Weninger, Geoffrey Khan, Michael P. Streck, and Janet C.E. Watson (eds) *The Semitic Languages: An International Handbook Handbücher zur Sprach- und Kommunikationswissenschaft (HSK)* 36. Boston – Berlin: De Gruyter Mouton. 259–278.

- Ingham, Bruce. 1982. *North east Arabian dialects*. London: Kegan Paul International.
- Jastrow, Otto. 1978, 1981. *Die Mesopotamisch-Arabischen Qəltu-Dialekte* (2 vols). Wiesbaden: Steiner.
- Jastrow, Otto. 2011. Dialect Differences in Uzbekistan Arabic and their Implications for the History of Arabic. 9th Meeting of the Association Internationale de Dialectologie Arabe (AIDA) in Pescara (Italy), March, 28–31.
- Owens, Jonathan. 2003. Arabic dialect history and historical linguistic mythology. *Journal of the American Oriental Society*. 123: 715–740.
- Owens, Jonathan. 2006. *A linguistic History of Arabic*. New York: Oxford University Press.
- Procházka, Stephan. 2013. Traditional Boatbuilding – Two texts in the Arabic dialect of the island of Arwād (Syria). In Renaud Kuty, Ulrich Seeger and Shabo Talay (eds) *Nicht nur mit Engelszungen, Beiträge zur semitischen Dialektologie, Festschrift für Werner Arnold zum 60. Geburtstag*. Wiesbaden: Harrassowitz. 275–288.
- Retsö, Jan. 2003. *The Arabs in Antiquity: Their history from the Assyrians to the Umayyads*. New York: RoutledgeCurzon.
- Retsö, Jan. 2006. Thoughts about the Diversity of Arabic. In Lutz Edzard and Jan Retsö (eds) *Current Issues in the Analysis of Semitic Grammar and Lexicon II: Oslo–Göteborg Cooperation 4th–5th November 2005*, Wiesbaden: Harrassowitz. 23–33.
- Retsö, Jan. 2013. What is Arabic? In Jonathan Owens (ed.) *The Oxford Handbook of Arabic Linguistics*. Oxford: Oxford University Press. 433–450.
- Tosco, Mauro and Manfredi, Stefano. 2013. Pidgins and Creoles. In Jonathan Owens (ed.) *The Oxford Handbook of Arabic Linguistics*. Oxford: Oxford University Press. 495–519.
- Versteegh, Kees. 1997. *The Arabic Language*. Edinburgh: Edinburgh University Press.
- Woidich, Manfred. 1993. Die Dialekte der ägyptischen Oasen: westliches oder östliches Arabisch? *Zeitschrift für Arabische Linguistik* 18:340–356.

33 Dialects in the Indo-Aryan Landscape

ASHWINI DEO

33.1 Introduction

The Indo-Aryan branch of the Indo-European language family currently occupies a significant region of the Indian subcontinent, its member languages being spoken in the bulk of North India, as well as in Pakistan, Bangladesh, Nepal, Sri Lanka, and the Maldives. Its closest relative is the neighboring Iranian branch of Indo-European. The historical depth of the textual record and the geographical breadth of the Indo-Aryan linguistic area, the diversity of its languages (226 in all), and its many speakers (about 1.5 billion in number) all serve to make Indo-Aryan a complex object of linguistic investigation. This chapter offers an overview of the broad structure of the Indo-Aryan branch and a classification of its major member languages, tracing briefly the historical record that leads to its synchronic distribution. It is crucial to note that Indo-Aryan is not one language, and a comparative study of the “dialects” of Indo-Aryan necessarily involves a comparison between several mutually unintelligible languages—many of them with millions of speakers, deep literary records, and complex dialectal differences within. The level of resolution at which variation within Indo-Aryan can be considered in this brief survey is thus different from the intra-linguistic level at which dialectal variation is usually studied.

The presence of Indo-Aryan in the Indian subcontinent can be dated back to approximately the early second millennium B.C.E. The influx of Indo-Aryan speakers first occurred in the mountainous areas of Afghanistan and Pakistan and the river plains of Punjab, with gradual migrations eastward and southward over the following millennia. Within the early phase of Indo-Aryan expansion, we can identify the initial geographical center as the Upper Indus valley (now in Pakistan) and the later center (toward the end of the Vedic period) as the Gangetic plains of North India.¹ By the time of the Buddha (sixth century BCE), most of North India (i.e., north of the Vindhya mountain ranges and the Narmada river) was Indo-Aryan speaking, these groups having displaced the original languages of the region, which included Dravidian and Austro-Asiatic languages as well as languages of unknown stock (see Witzel 1999 for a detailed review of substrate evidence). Gradually, over the next millennium and a half, Indo-Aryan spread towards the South, occupying areas south of the Narmada river, a region corresponding to the modern territory of Marathi and Oriya. It is this contiguous geographical territory over which the modern dialectological landscape of Indo-Aryan is to be found. The non-contiguous Indo-Aryan languages (which include, for instance, Sinhala (Sri Lanka), Divehi (Maldives), Parya (Tadzhikistan), and Romani (mainly Eastern Europe)) are the result of pre-modern migrations of Indo-Aryan speakers into non-Indo-Aryan territory (Masica 1993: 22).

Fifteen of the 22 official languages recognized by the Eighth Schedule of the Indian Constitution are Indo-Aryan. This status allows the use of these languages for both educational and regional administrative purposes. Pakistan and Bangladesh recognize only Urdu and Bangla respectively as their official languages; despite sizeable speaker populations, Punjabi, Siraiki, and Sindhi do not have official status in Pakistan. Nepali is the official language in Nepal, Sinhala in Sri Lanka, and Dhivehi in the Maldives. The scheduled Indo-Aryan languages of India include Assamese, Bangla, Dogri, Gujarati, Hindi, Kashmiri, Konkani, Maithili, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Sindhi, and Urdu. Of these, Hindi (written in Devanagari script) has special status as the official language of the Union of India. Speaker populations range from 422 million (Hindi) to 2 million (Dogri).² All other Indo-Aryan varieties are classified as “non-scheduled” or unofficial languages, and accorded the status of “mother-tongue” varieties or dialects of the regional standard variety. The map in Figure 33.1 shows the geographic location of the major Indo-Aryan languages.

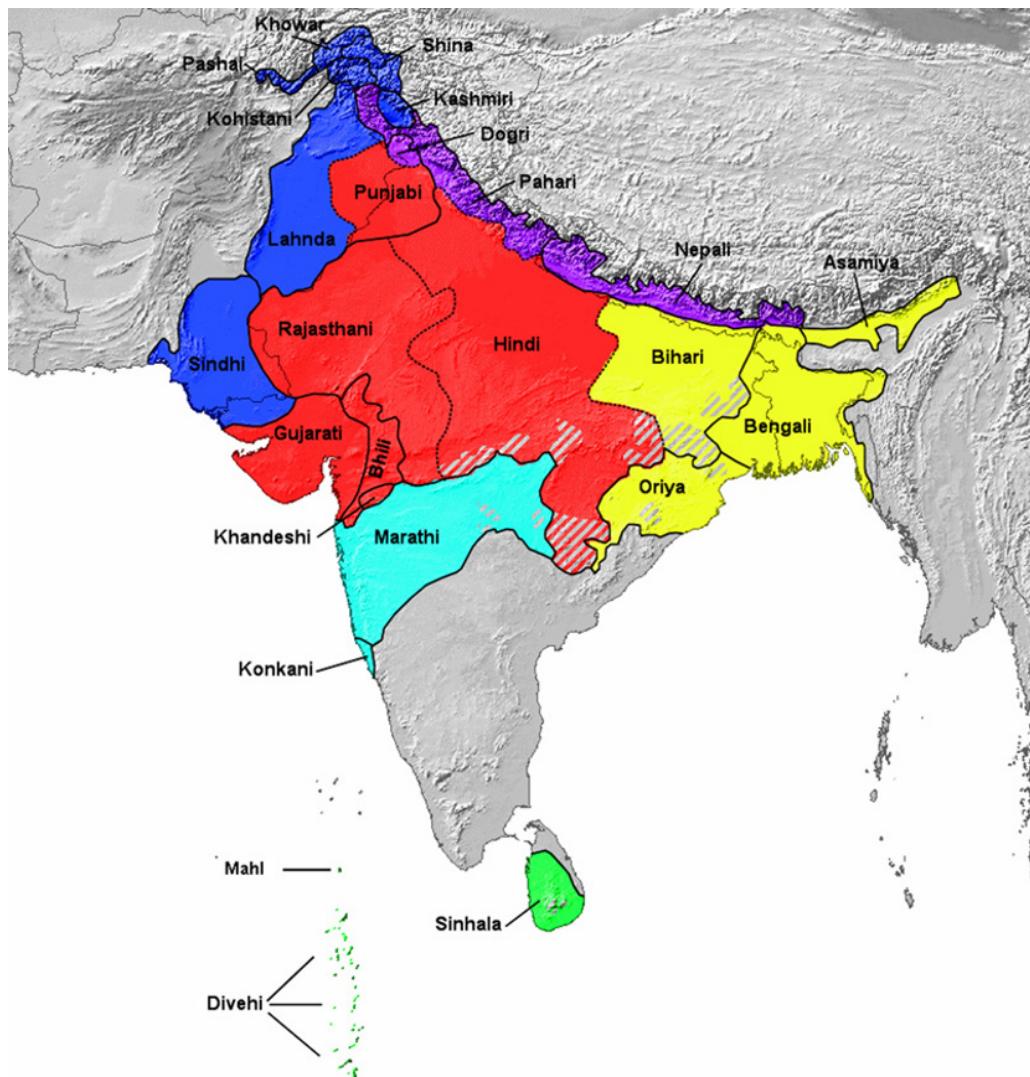


Figure 33.1 The distribution of the Indo-Aryan languages.

33.2 The Historical Record of Dialect Variation

Chronologically, the evolution of Indo-Aryan can be divided into three main stages: the Old Indo-Aryan (OIA) stage, which includes early and late Vedic texts as well as the later Epic and Classical Sanskrit literature; the Middle Indo-Aryan (MIA) Prakrit stage, with the textual record for Prakrit starting around the third century BCE and continuing until the ninth century CE; and the New Indo-Aryan (NIA) stage, whose literatures are attested from the twelfth century CE and whose modern forms constitute the current Indo-Aryan landscape.

The earliest clear evidence for regional dialect variation within Indo-Aryan languages comes from the Ashokan inscriptions, which are dated to the third century BCE. The Ashokan inscriptions contain 33 edicts that are found in scattered locations in modern-day India, Pakistan, Nepal, and Bangladesh. The language of these inscriptions exemplifies the Early MIA stage and is commonly known as inscriptional Prakrit. Based on the distribution of certain linguistic innovations observed there, Bloch (1950) identified three geographical dialect areas—Eastern (E), the Northwestern (NW), and Southwestern (SW). Southworth (2005) resolves this three-way distinction in the Ashokan inscriptions to an earlier binary division between NW on the one hand and E and SW on the other. The shared innovations that have been presented to motivate this classification include:

- (1) a. The three-way distinction between OIA sibilants /s/ ([s]), /ś/ ([ʃ]), and /ʂ/ ([ʂ]) is neutralized to [s] in the E and SW, but retained in the NW.
- b. OIA syllabic /r/ (phonetic [r̩]) is vocalized to /a/ in the E and SW but to /i/ in the NW (e.g., OIA *kṛta* → E MIA *kāta*, SW MIA *kata*, but NW MIA *kit̩(r)a*).
- c. OIA consonant clusters of the form /C_ir/ undergo assimilation in the E and SW dialects and change to /C_iC_i/ (e.g., OIA *agra* “foremost” → E MIA *agga*, SW MIA *agga*, but NW MIA *agra*).
- d. OIA consonant clusters of the form /rC_i/ undergo assimilation in the E and SW and change to /C_iC_i/; these undergo metathesis in the NW (e.g., OIA *garbha* “womb” → E MIA *gabbha*, SW MIA *gabba*, but NW MIA *grahba*).

Innovations that are exclusively attested for the Eastern dialect include the following:

- (2) a. The merger of OIA /n/ ([n]), /ñ/ ([ñ]), and /ṇ/ ([ɳ]) to /n/ ([n]) (e.g., OIA *jñātika* “kinsman,” *dāna* “gift,” *prāṇa* “breath” → *nātika*, *dāna*, *pāna*). These are retained in the SW and NW.
- b. The full merger of OIA /r/ and /l/ to /l/ (e.g., OIA *rājā* “king,” *rūpāṇī* “spectacles” → *lājā*, *lūpāṇī*).³
- c. The masc./neut. nom. sg ending -ah changes to -e, whereas it is retained (with some exceptions) in its allomorphic form -o in the SW and NW.

Later evidence for dialect variation in MIA comes from traditional grammars and commentaries on the dramatic literature of Classical Sanskrit and Prakrit. These contain valuable evidence of regional MIA dialects in the first millennium CE (better known as the literary Prakrits), their linguistic features, and their internal sociolinguistic stratification. The more important literary Prakrits include Śauraseni, associated with the Northwestern region, Mahārāṣṭri (the Prakrit par excellence), associated with the Southern region, and Māgadhi, associated with the East.

The distinct dialectal divisions identified for Early and Middle MIA have implications for the subgrouping of modern NIA languages, which is controversial. The Stammbaum model fails to work satisfactorily for this region. As Masica (1993: 446) describes it, this region is characterized by few internal natural barriers, unstable political units that do not coincide

with linguistic units, and significant internal migrations. The result has been dialect continua without sharp boundaries separating mutually unintelligible languages. Nevertheless, broad geographical divisions can be identified within the current linguistic landscape, which correspond to the general territories of the modern NIA languages, including (Western) Hindi, with its multiple dialectal variants, and Nepali, which occupy the North-Central region of the subcontinent; Punjabi, Sindhi, and Lahnda to the Northwest; Rajasthani and Gujarati to the West; Marathi and Konkani to the South; and Bengali, Oriya, Assamese, and Eastern Hindi to the East.⁴

Hoernle (1880) proposed a classification that assumed two broad branches in remote times that gave rise to the contemporary languages—a Southern-Eastern branch (which grouped Marathi together with Bengali, Oriya, and Eastern Hindi) and a Northwestern branch (grouping Western Hindi and Nepali with Punjabi, Sindhi, and Gujarati). This hypothesis was further refined by Grierson, whose Inner-Outer hypothesis closely builds on the affinities between the Southern and the Eastern languages observed by Hoernle. Grierson proposed a model involving two distinct waves of immigration into the subcontinent, one of which led to the settlement of Northern India and gave rise to Western Hindi and its dialects—the “inner group”—and another encircling wave that corresponds to languages surrounding this inner group in a semi-circle—the “outer” group. This hypothesis, like Hoernle’s, groups together the contemporary Eastern and Southwestern languages (a grouping supported by evidence from Middle Indo-Aryan dialect classification) and further posits that Northwestern languages such as Sindhi and Lahnda are closer to these than to Western Hindi. Grierson first posited this model before setting out on the ambitious Linguistic Survey of India, described in §3 (LSI 1.1: 116–118), then further revised it after the survey (Grierson 1931–33). The revised model, in addition to the Inner and Outer branches, includes an Intermediate branch, whose members are considered to be “inner” languages superimposed on an “outer” substratum. These include Punjabi, Gujarati, Nepali, Eastern Hindi, and Rajasthani. Grierson bases his view on a number of grammatical and phonological criteria:

- (3) a. The Outer languages form their perfective aspect forms with an /-l/ affix (first attested in late MIA), whereas the Inner languages retain the perfective paradigm derived from OIA.
- b. The Inner languages tend to be more analytic and retain fewer inflectional features than the Outer languages.
- c. The Inner languages preserve the distinction between OIA sibilants /s/ ([s]), /ʃ/ ([ʃ]), and /ʂ/ ([ʂ]), which is reduced in different ways in the Outer languages.

Grierson’s criteria, and indeed the entire Inner-Outer hypothesis, were severely criticized by Chatterjee (1926), according to whom Grierson had in some cases inaccurately represented the geographical distribution of features and in others was describing changes of relatively recent origin or cases of independent development (Masica 1993: 449–450). Later proposals, including Chatterjee’s own, assume different divisions, which emerge out of varying criteria being privileged over others. Masica describes in some detail the incompatible classifications implicitly or explicitly provided by Turner (1975), Katre (1968), Cardona (1974), and Nigam (1972) and comes to the conclusion that:

Perhaps a wiser course would be to recognize a number of overlapping genetic zones, each defined by specific criteria.

... We might therefore be well-advised to give up as vain the quest for a final and “correct” NIA historical taxonomy, which no amount of tinkering can achieve, and concentrate instead on working out the history of various features. (Masica 1993: 460)

While this is a sensible caution against striving for sharp, Stammbaum-like classifications, Southworth's revival of Grierson's original construal of Indo-Aryan regional divisions deserves mention here. Southworth (2005: 135 ff) introduces linguistic evidence overlooked by Grierson and later scholars to adduce support for a Griersonian picture of dialect divisions.

- (4) a. The usage of the OIA gerundive in *-tavya* to mark future meaning is observed the Eastern and Southwestern languages in contrast to the North-Central languages.
- b. The North-Central languages have preserved some instances of phonemic contrast between long and short high vowels (/i/ ≠ /i:/ and /u/ ≠ /u:/), whereas length in the Eastern and Southwestern languages is allophonic.
- c. Although there are exceptions, the Eastern and the Southwestern languages typically exhibit initial word accent. By contrast the North-Central languages typically stress the rightmost heavy syllable (*modulo* final extrametricality). The North-Central pattern is taken to be the conservative one inherited from Late OIA and MIA. (Turner 1916: 47, Chatterjee 1926: 280).
- d. The lateral /l/ changes to /n/ in several lexical items in the Eastern and the Southwestern languages, which is far less common in the North-Central languages.

Southworth synthesizes his linguistic arguments with textual and archaeological evidence to build a convincing picture of the routes of expansion of Indo-Aryan speaking populations into the sub-continent and the resulting dialect space that has generated the patterns of distribution observed in the NIA languages.

33.3 Studies and Sources

Linguistic study of the NIA languages was first undertaken by the scholar-administrators of the East India Company/British Government, initially in the context of training officers who came to the colony and later as part of the larger Orientalist pursuit. In addition to grammars of individual NIA languages written in the 19th century, Beames (1872–1879 [1966]) undertook a comparative investigation of seven major languages—Hindi, Punjabi, Sindhi, Gujarati, Marathi, Oriya, and Bangla. Hoernle (1880) represents the other main nineteenth-century contribution to the regional taxonomy of Indo-Aryan (and its historical implications).

The first systematic study of contemporary spoken Indo-Aryan languages and dialects was undertaken by Grierson, through the monumental *Linguistic Survey of India* (LSI), a comprehensive survey of the languages and dialects of British India. The survey was planned as a collection of specimens in which a standard passage was “to be selected for comparison and this was to be translated into every known dialect and subdialect spoken in the area covered by the operations” (LSI 1.1: 17). In order to access idiomatic language, this translation was to be complemented by a piece of folklore or some other passage in narrative prose or verse. There was further a list of words and sentences whose equivalents in each language were sought. Conducted by the (colonial) Indian Government between 1894 and 1928, the project published 11 volumes of results, of which 7 are exclusively dedicated to the Indo-Aryan languages. Grierson's work depended crucially on the cooperation of local government officers in identifying dialect communities and collecting/correcting samples obtained from a large part of the subcontinent. For each regional variety, the LSI provides a translation of the Prodigal Son story, and for most varieties, also a brief (often single page) grammatical sketch that identifies their phonological and morphological properties. For many of the varieties, we also find the narrative passage and word/sentence lists. The survey distinguishes between standard and non-standard varieties of individual languages and also identifies

dialect continua that cannot be subsumed under a distinct regional language. The classification in the LSI builds on Grierson's ideas of prehistoric dialect regions, but can also be taken to be a neutral synchronic description of the linguistic space. Bengali, Assamese, Eastern Hindi (Magahi), and Oriyā are assigned to the Eastern group, Marāthī and its dialects to the Southern group, Sindhi, Lahnda, and the Dardic languages (including Kashmiri) to the North-Western group, whereas Western Hindi, Punjabi, Rajasthani, Gujarati are assigned to the Central group (which corresponds to the Outer group of the original proposal). This is also the grouping assumed in the main by the SIL Ethnologue.

An important outcome of the LSI was the identification and basic description of languages and dialect continua that could not be easily assimilated within any of the major regional languages. The Bhil and Khāndeśī languages in the Southwest and the Pahārī and Gujurī languages in the Northwest are notable examples, to which are dedicated entire sub-volumes of the LSI (LSI 9.3 and LSI 9.4 respectively). Both of these language groups are assigned to the Central Group of Indo-Aryan and are represented by dozens of distinct, sometimes mutually unintelligible varieties.⁵ The LSI data have generated an invaluable high-resolution picture of the Indo-Aryan language and dialect space in the first decade of the twentieth century, which has never been fully replicated in scope or linguistic rigor. The entire contents of the 11 volumes are now available in searchable form at <http://dsal.uchicago.edu/books/lsi/>.

Another invaluable resource is Sir Ralph Lilley Turner's *A Comparative Dictionary of the Indo-Aryan Languages* (4 vols., 1962–1966). Turner, an army officer turned comparative philologist, assembled this dictionary based on his own and other scholars' work over a period of 50 years. The main dictionary contains 14,189 lexical entries, either attested in OIA or MIA or reconstructed, and provides their reflexes (as applicable) in a number of standard and non-standard NIA languages. The dictionary is available in searchable form at <http://dsal.uchicago.edu/dictionaries/soas/>.

For individual languages, there were some smaller-scale dialect surveys undertaken in the post-independence period. To mention a few: Ghatage (1962–1968) documents several dialects of Marathi (including those spoken by diasporic communities) while Gill (1973) is a *Linguistic Atlas of the Punjab*. For Bengali, there are descriptions of isolated dialects (Chaudhuri 1939, 1940; Sen 1972; Goswami (1939)) but no unified survey apparatus that identifies sub-regional dialect patterns and relevant linguistic variables. Gusain (2000–2008) contains grammars of Rajasthani dialects compiled with a set template, enabling comparability and identification of points of variation. Dialectological resources for Hindi are quite scattered, especially because "Hindi" is an umbrella term covering the Western and Eastern varieties, which have distinct grammars and contain several sub-varieties that are arguably distinct languages as well. The best resource compiling work on standard and non-standard Indo-Aryan languages and their dialects is still Masica (1993); its bibliography has been organized by language and puts together available material for each language and its dialects in one location. Shapiro (2007) also provides some references for work on distinct varieties of Hindi.

Two ongoing efforts at documenting the languages of India should be mentioned in this context. The first is the *People's Linguistic Survey of India* (<http://peopleslinguisticsurvey.org>), an ambitious survey of the entire country's languages initiated in 2010 by the Bhasha Research and Publication Center. Spearheaded by Ganesh Devy, the project "envision[s] the creation of a Linguistic Survey rooted in people's perception of language."⁶ The survey, carried out by native speakers of the distinct language communities, is articulated as a political movement for the identification and self-assertion of the country's distinct linguistic varieties. It consists of a basic grammatical template and also asks for information about the geographical and linguistic contexts of the speech community and samples of oral literature. The multi-faceted goals of this survey project and its reliance on native-speaker researchers make it a politically empowering project. Its results are in the process of being published; for

Indo-Aryan, data on the languages of four states—Maharashtra, Rajasthan, Jammu & Kashmir, and Uttarakhand—are available, with the volume on Gujarat about to be published at the time of this writing.

The second effort is the web-based Language Information Service (LIS)-India (<http://www.lisindia.net>), initiated by the Central Institute of Indian Languages (CIIL), which provides census data and information on intra-linguistic dialect variation and historical evolution for both scheduled and non-scheduled Indian languages.

33.4 Language and Dialect in the Indo-Aryan Context

In 1956, following regional demands and widespread agitations for linguistically based states, the Indian Government implemented a major reform of the boundaries of India's administrative states along linguistic lines by passing The States Re-organisation Act of 1956.⁷ This drawing of linguistic-political boundaries, which were further articulated over the next several decades through the creation of more linguistic states, has been central to defining how linguistic identity is “officially” established in the Indian context. Movements for regional autonomy (and “statehood” status) have gone hand-in-hand with movements for official linguistic recognition, as in the formation of the states of Maharashtra (Marathi) and Gujarat (Gujarati) in 1960, Punjab (Punjabi) in 1966, and Goa (Konkani) in 1987.⁸

The distinction between “language” and “dialect” is especially fraught in a situation where the political and administrative recognition of regional languages has obliterated sub-regional linguistic identities, creating new hierarchies and threatening the historically stable differentiation between varieties (Khushchandani 1991, Annamalai 2001, Dasgupta and Sardesai 2010). Historically, India has been tolerant of multilingualism with distinct varieties co-existing side by side, as well as long-term retention of native languages by diasporic communities in non-native geographical contexts. The multilingual nature of most social interactions, widespread diglossia, and non-discrete boundaries between speaker communities create fluid zones of linguistic interaction, where linguistic identity can often be a shifting notion determined by sociopolitical goals. In modern India, the fluid boundaries between language and dialect have been used systematically to both delineate new and suppress old identities.

Shapiro and Schiffman (1983: 5) provide a clear example of the former with Punjabi, a dialect that turned into a language within a span of a few decades. Punjabi was widely considered a dialectal form of Hindi before India's independence. Punjabi was used at home but Hindi or Urdu prevailed in education. After independence, strong demand for a Punjabi-speaking state led to “changing popular and official attitudes towards the code.” The formation of Punjab in 1966 in turn led to the language becoming a medium for education and broadcasting and to a corresponding increase in publication, thus crystallizing its linguistic identity and distinctness from standard Hindi.

Suppression of old identities is more easily observed. The pressure for linguistic homogeneity through public media and education within administrative-linguistic states, which contain significant proportions of minority linguistic populations, has led to unequal relations between dominant and minority communities (Emeneau 1962, Khushchandani 1991). Thus, minority languages like Bhili (9.5 million in 2001), Khāndesi (2 million in 2001), and Halbi (0.59 million in 2001), while assigned “mother tongue” status in census counts, are being subsumed under the regional standard determined by political boundaries.

A historical example of the suppression of previously dominant linguistic varieties comes from the Hindi dialect continuum and the emergence of Khari Boli as the prestige dialect in the eighteenth and nineteenth centuries (King 1994: ch. 2–5). The Hindi belt, a contiguous region encompassing several dialect varieties and often distinct languages as well, occupies

a large central region of the subcontinent.⁹ The boundaries of this broad dialect continuum, following Masica (1993), often include Bihari, Magahi, and Maithili, which are the Eastern-most varieties, the last of which was, in fact, declared an official language in 2003. Rajasthani (with its dialects) is at the western boundary of the continuum. It is recognized as a language for literary purposes, but does not have official status yet.

Gumperz (1957, 1958) distinguishes between three hierarchical forms of speech within this continuum: local level village dialects; regional dialects found in small market centers that are relatively uniform over a large area; and Standard Hindi, whose primary base of native speakers is found only in large cities like Delhi, Agra, and Lucknow. Within the Hindi continuum, there are some regions where the three forms are mutually unintelligible, whereas in others, the three are relatively close to each other.¹⁰

Some of these varieties, such as Braj and Awadhi, the dominant languages of the North Indian pre-modern literary traditions, were recognized as independent languages for much of the New Indo-Aryan linguistic period. However, since independence, these languages, together with Kanauji, Bundeli, Bagheli, Bangaru, and several others, have been subsumed under Hindi as regional varieties, with their literatures co-opted within the great Hindi literary tradition. The reason for this language-to-dialect transition over a century has to do with the rise of Khari Boli as a literary dialect and lingua franca in the nineteenth century. It acquired this status in part due to its role in the training of British officers in colleges for administrators (e.g., Fort William College in Calcutta). The rise of this dialect was accompanied by a growing social movement that sought to differentiate between Hindi and Urdu (King 1994). This polarization, which was underpinned by socioeconomic as well as political motivations, further consolidated the status of Khari Boli as the authoritative dialect across North India and ensured that this variety of Hindi emerged after independence as the official language of the Union of India, with a blurring of former regional and cultural differences.

33.5 The Hindi Belt: Dialect Differences and Emergence of the Standard

Identifying dialects of Hindi is complicated for several reasons: variation among scholars regarding the boundaries of the Hindi belt and varieties that can or should be subsumed under the Hindi umbrella; inconsistency in language planning policies and census practices over time; and diglossia, with varying degrees of control over and movement between the varieties involved. The one point of consensus is that there are two sets of dialects—Western and Eastern—subsumed under the Hindi umbrella (Masica 1993, Shapiro 1989: 3–5). Western dialects include Braj, Bundeli, Harianvi/Bangaru, Kanauji, and vernacular Hindustani (also called Khari Boli). Eastern Hindi has three major dialects: Awadhi, Bagheli, and Chattisgarhi (Shapiro 2007: 251). Some salient linguistic differences between the sets include:

- (5) a. Eastern varieties form the future with the suffix /-ba/, inherited from the Old Sanskrit gerundive in *-tavya*, while Western dialects developed a new future paradigm based on the OIA present stem.
- b. The perfective paradigm in Eastern dialects has an added /-l/ affix, absent in Western dialects.
- c. The ergative clitic /-ne/, characteristic of western Hindi, is absent in Eastern dialects.
- d. Lexical accent tends to be word-initial in Eastern dialects, whereas Western dialects retain older stress patterns.

Many of these differences correspond to the broader, historically rooted Eastern/North-Central divide for Indo-Aryan languages in the Grierson/Southworth model discussed above.

The Hindi language region has for much of its history been a loose network of only partially mutually intelligible dialects. The regional literary traditions that developed in some of these varieties, notably Braj, Avadhi, and Maithili, did not remain confined to their region of origin but spread broadly across Northern India, reinforcing speaker-competence in and intelligibility between multiple dialectal varieties. Until the mid-nineteenth century, there was very little assimilation of dialects in favor of any standard in the Hindi-speaking region. The current situation is markedly different, with clear recognition of a prestigious Hindi standard among speakers of all varieties and varying degrees of competence in this standard relative to more local varieties. This trend of assimilation, effected over the past one and half centuries, is best understood as the consequence of another process of conscious differentiation, the polarization of Hindi and Urdu.

33.5.1 The Evolution of Modern Standard Hindi

Modern Standard Hindi, the official national language of India and the language used across the country in both education and administration, takes as its dialectal base Khari Boli, the dialect spoken in and around Delhi. This language is virtually identical in grammatical structure to Urdu, the official language of Pakistan, and one of the scheduled languages of India. This differentiation of a single dialect into two socioculturally distinct entities emerged through a conscious fashioning of the two linguistic identities and their harnessing for religious and nationalistic expression in the nineteenth and the twentieth centuries (King 1994, Trivedi 2003). The polarization between Hindi and Urdu and the crystallization of each have had far-reaching effects on the social dialectology of the Hindi belt in India and on language policy and education in the multilingual South Asian context. In the early nineteenth century, Urdu or Hindustani (as it was called by the colonial British administrators) was the Khari Boli dialect with Perso-Arabic influence, spoken predominantly by Muslims, but also used by literate, elite Hindus. Another version of Khari Boli, the so-called Hindu version, relied on Sanskrit for its learned vocabulary and was written in the Devanagari script, shared by several Indo-Aryan languages. In 1837, the British colonial government passed Act 29, replacing Persian with Urdu as the official language for judicial and administrative purposes in the North West Provinces and parts of the Central Provinces of India. This sparked a demand in several regions of North India that Devanagari, the script in which Hindi is written, be used as an alternative script alongside the modified Persian script (in which Urdu is written), for administrative and judicial business in Northern India (King 1994). This demand became an organized literary and sociopolitical movement in the late nineteenth century, called the Hindi-Nagari movement. While Urdu already had a well-established poetic tradition, Hindi prose and poetry only began to emerge and develop in the mid-to-late 1800s. As both Urdu and Hindi cultivated distinct literary styles of the same Khari Boli dialect base, the reliance on distinct classical languages (Perso-Arabic and Sanskrit, respectively) increased and the languages became markers of religious and political identity. King (1994), describing this evolution over the second half of the nineteenth century, calls it a process of “multi-symbol congruence,” by which the Hindi language, the Devanagari script, and the Hindu religion became more closely identified with one another and were jointly opposed to Urdu, its Persian-based script, and Islam. The Hindi-Nagari movement, driven by a desire to define Hindi as distinct from (and superior to) the culturally, literarily, and politically dominant Urdu, was underpinned by this multi-symbol congruence (Brass 1974, King 1994, Dalmia 1997). The Indian Independence movement, in particular political bodies like the Indian National Congress, adopted Hindi/Hindustani as the movement’s common language, citing

its religiously neutral Hindi-Urdu grammatical core and its intelligibility to large swathes of the North Indian population. Despite political efforts to amalgamate the two sociocultural entities and dissociate Hindustani from issues of script, Perso-Arabic versus Sanskritic influence, and religion, the Hindi language movement and the literary and linguistic norms it established in opposition to Urdu prevailed (King 1994, Dasgupta and Sardesai 2010). After independence and the concomitant separation of colonial India into India and Pakistan, the Indian constitution of 1950 provided that Hindi, written in Devanagari script, was to be the national language of the Indian republic. Over the following decades, Hindi was increasingly “cast in a heavily Sanskritized mould,” in education as well as print media and television (Shapiro 1997: 256).

This top-down, state-imposed norm of Modern Standard Hindi (MSH) has had far-reaching effects on the dialectal situation in the Hindi belt. Until constitutional sanction for Hindi, MSH had relatively few native speakers in either its deliberately crafted literary version, or its dialectal base, Khari Boli. The majority of the so-called Hindi/Hindustani speaking population spoke one of the regional Eastern or Western dialects, and depending on access to education, had some familiarity with the standard. The current situation is markedly different. Khubchandani (1997), describing the code fluidity of the Hindi belt, divides the population into five categories:

- (6) a. Bilinguals of the north-central region who view their own primary speech in terms of substandard variations of the prestigious MSH standard.
- b. Bilinguals who use primary varieties in their intimate rural milieu and MSH in modern settings.
- c. Bilinguals who use primary varieties for all oral communication and relegate MSH to written use.
- d. Illiterate monolinguals who only speak primary varieties but say they belong to the Hindi fold.
- e. “Real” users who speak and write MSH and nothing but MSH.

There are two points to note here. First, distinct varieties that are controlled by bilingual speakers are often perceptually assimilated into the larger Hindi fold. That is, bilingual speakers see themselves as monolingual speakers, where MSH and the variety they control (e.g., Braj or Awadhi) are taken to be stylistic variants of a single language. Second, the last category of speakers—monolingual MSH speakers—has increased dramatically in the Hindi belt as a consequence of educational policies and media exposure. Thus, in terms of both perception and actual linguistic behavior, we see assimilatory trends as the Hindi standard is imposed over a complex dialect continuum.

NOTES

- 1 This research was supported by NSF grant BCS-1255547/BCS-1660959. Most of our evidence for the geographical location and onward trajectory of the Indo-Aryan speaking population from the North West comes from the texts of the Vedic period, which mention river names and tribal boundaries.
- 2 http://www.censusindia.gov.in/Census_Data_2001/Census_Data_Online/Language/Statement4.aspx
- 3 This particular change has been problematic for linguistic reconstruction. First, Vedic texts exhibit a partial merger of PIE */r/ and */l/ to /r/. Second, none of the later attested languages (including modern counterparts in the same region) exhibits the full merger. The [l-r] variation, moreover, was consciously perceived as a sociolinguistic marker, whereby [r] distinguished Aryan Brahminical

- (high-class) speech, whereas [l] was stigmatized from the late Vedic period as non-Brahminical (Bloch 1965, Deshpande 1979, Southworth 2005: 161–167).
- 4 The Dardic languages to the North (which include Kashmiri and Shina) form a separate subgroup that is sometimes considered to be an independent branch of Indo-Iranian and sometimes located within Indo-Aryan.
 - 5 In the Bhili and the Khandeshi case, I know from personal fieldwork in the linguistic region that mutual intelligibility often has to do more with long-term contact and multilingualism than with grammatical properties of the languages.
 - 6 <http://peopleslinguisticsurvey.org/aboutus.aspx?page=PLSI>
 - 7 The Sri Lankan Government passed its Official Language Act in 1956, making Sinhala the only official language of a country with a large Tamil-speaking minority.
 - 8 See Brass (1974) and Windmiller (1956) for overviews and perspectives. This theme of linguistic regionalism extends to Pakistan, formed in 1947, which had two major linguistic contingents, Urdu and Bangla. Differences between these first led to the creation of a separate provincial ministry for the Bangla speaking region, and ultimately to the separation of Pakistan into two distinct countries, Pakistan (Urdu majority) and Bangladesh (Bangla majority) in 1972.
 - 9 Hindi speakers are concentrated in the Indian states of Uttar Pradesh, Uttarakhand, Madhya Pradesh, Chhattisgarh, Bihar, Jharkhand, Haryana, Rajasthan, Himachal Pradesh, and Delhi.
 - 10 This hierarchical pattern is also found in the Bhili and Khandeshi dialect continuum where the highest standard is either Gujarati or Marathi depending on the geographical location of the community and usually mutually unintelligible with the members of the dialect continuum.

REFERENCES

- Annamalai, E. 2001. *Multilingualism in India: Political and Linguistic Dimensions*. New Delhi: Sage Publishers.
- Beames, John. (1872–1879). A Comparative Grammar of Modern Indo-Aryan Languages of India, Munshiram Manoharlal, Delhi, Republished 1966.
- Bloch, Jules. 1950. *Les inscriptions d'Asoka*. Paris: Société d'édition les Belles Lettres.
- Brass, Paul R. 1974. Language, Religion and Politics in North, Cambridge University Press, Cambridge, Republished in 2005 by Backinprint.
- Cardona, George and Dhanesh Jain. 2007. *The Indo-Aryan languages*, Routledge, New York.
- Chatterjee, Suniti Kumar. 1926. *The Origin and Development of the Bengali Language*. London: George Allen and Unwin. (Reprinted 1975 by Rupa & Co., Calcutta).
- Chaudhuri, Sambha Chandra. 1939. Notes on the Rangpur dialect. *Indian Linguistics* 7: 297–315.
- Chaudhuri, Sambha Chandra. 1940. North Bengal dialects: Rajshahi. *Indian Linguistics* 8: 418–431.
- Dalmia, Vasudha. 1997. *The nationalization of Hindu traditions: Bharatendu Hariśchandra and nineteenth-century Banaras*. Cambridge, Mass.: Oxford University Press.
- Dasgupta, Probal and Madhavi Sardesai. 2010. 'Sociolinguistics in South Asia', In Ball, Martin J. (ed.), *The Routledge Handbook of Sociolinguistics Around the World*, 81–88, Routledge, Oxford.
- Deshpande, Madhav M. 1979. *Sociolinguistic Attitudes in India: An historical reconstruction*. Ann Arbor, MI: Karoma Publishers.
- Emeneau, Murray B. 1956. 'India as a linguistic area', *Language* 32, 3–16.
- Ghatage, Amrit. 1962–1968. A Survey of Marathi Dialects (vol 1–9), The Maharashtra State Board for Literature and Culture, Bombay.
- Gill, Harjeet Singh. 1973. Linguistic atlas of the Punjab, Dept. of Anthropological Linguistics, Punjabi University, Patiala.
- Goswami, Krishnappa. 1939. Linguistic notes on Chittagong Bengali. *Indian Linguistics* 8: 493–536.
- Grierson, Sir George Abraham Grierson. 1903–1928, Linguistic Survey of India, volume I–XI, Office of the Superintendent of Government Printing, India, Calcutta, Reprinted 1967 by Motilal Banarsi das.
- Gumperz, John T. 1957. 'Language Problems in the Rural Development of North India', *The Journal of Asian Studies* 16.2, 251–259.
- Gumperz, John T. 1958. 'Dialect Differences and Social Stratification in a North Indian Village', *American Anthropologist* 60, 668–682.

- Gusain, Lakhani. (2000–2008). Bagri, Shekhawati, Mewati, Marwari, Mewari, Dhundhari, Harauti, and Wagri grammars, Lincom Europa, Munich.
- Hoernle, A. F. Rudolf. 1880. *A Comparative Grammar of the Gaudian Languages*, Trübner and Co., London, Reprinted 1975.
- Katre, Sumitra M. 1968. *Problems of Reconstruction in Indo-Aryan*. Simla: Indian Institute of Advanced Study.
- Khubchandani, Lachman. 1991. 'India as a sociolinguistic area', *Language Sciences* 13, 265–288. Khubchandani, Lachman. 1997. 'Indian diglossia', In *Revisualizing boundaries: a plurilingual ethos*.
- King, Christopher. 1994. *One Language, Two Scripts: The Hindi Movement in Nineteenth Century North India*, Oxford India, Delhi.
- Masica, Colin P. 1993. *The Indo-Aryan Languages*, Cambridge Language Surveys, Cambridge University Press, Cambridge.
- Nigam R. C. 1972. *Language Handbook on Mother tongue in Census*. New Delhi: Government of India (Census Centenary monograph No. 10).
- Sen, Nilmadhav. 1972. Some dialects of Bangladesh: an outline. *Indian Linguistics* 33: 143–152.
- Shapiro, Michael. 2007. 'Hindi', In Cardona, George and Dhanesh Jain (eds.), *The Indo-Aryan languages*, 250–, Routledge, New York.
- Shapiro, Michael and Harold Schiffman. 1983. *Language and Society in South Asia*, Foris, Dordrecht.
- Southworth, Franklin. 2005. *Linguistic archaeology of South Asia*, Routledge.
- Trivedi, Harish. 2003. 'The progress of Hindi, Part 2: Hindi and the nation', In Pollock, Sheldon (ed.), *Literary Cultures in History*, 958–1022, University of California Press, Berkeley.
- Turner, Ralph L. (1962–1966). *A comparative dictionary of Indo-Aryan languages*, Oxford University Press, London, Republished by Motilal Banarsi Dass Publishers, New Delhi.
- Witzel, Michael. 1999. 'Substrate Languages in Old Indo-Aryan (R̥gvedic, Middle and Late Vedic', *Electronic Journal of Vedic Studies (EJVS)* 5, 1–67.

34 Dialects of Chinese

CHAOJU TANG

34.1 General Introduction

Chinese dialect research is a big topic, especially if we include all levels of grammar and all aspects of variation. First, it has a long history. Second, Chinese dialects entail rich variation and present a wealth of research questions, encompassing historical phonology, family classification, field work methodology, data collection and corpus construction, atlas production, and dialect preservation in a linguistic environment dominated by *Putonghua*, or “Standard (Modern) Chinese.”¹ As all of these subtopics cannot be addressed in the space available, this chapter will focus on four main issues. Section 34.2 briefly reviews the history of Chinese dialect research and the establishment of Chinese Dialectology. Section 34.3 discusses the classification of Chinese dialects. Section 34.4 addresses the major sources of data on Chinese dialects and some of the methods they have employed. Section 34.5 reviews some remaining questions, including dialect preservation and future work on Chinese dialects.

34.2 The Development of Chinese Dialect Research

34.2.1 Seminal Work in Ancient China

Research on Chinese dialects can be traced back to the Zhou-Qin period (Zhou and Qin Dynasties, i.e., B.C. 770–B.C. 206), but reached its heyday in the Xihan (West Han) period (B.C. 206–A.D. 25). It was during this period that Yang Xiong (B.C. 53–A.D. 18), a native of Chengdu (now in Sichuan province) and a famous literary scholar and linguist in the Han dynasty (B.C. 206–A.D. 220), edited the *Fāngyán*, the first Chinese dialect dictionary, which is acknowledged to be a milestone for dialect research in China and, arguably, the world. The *Fāngyán* builds on the tradition of imperial emissaries who made annual surveys of regional dialect vocabulary throughout China during the Zhou Dynasty (B.C. eleventh century–B.C. 256). Its full name is *Yóuxuān shǐzhě juédài yǔshí biéguó fāngyán*, which literally means, “Local speeches of other countries in times immemorial explained by the Light-Carriage Messenger.” In its preface, Yang Xiong explains that he spent 27 years collating and editing the *Fangyan*, which contains about 9,000 dialectal characters in 13 volumes. Yang Xiong is consequently recognized as the first Chinese dialectologist and his *Fangyan* is called the first “Dialect Geography” because of its wide coverage of regional dialects.

34.2.2 The Establishment of Modern Chinese Dialectology

A second turning point in Chinese dialect research occurred in the early 1900s. In 1918, “Geyao Yanjiuhui,” a society for the study of “ballads” or folk songs, was established at Peking University. Because the ballads were sung in dialect, the work of this society contributed greatly to Chinese dialect research. In 1924, the Association of Dialect Research was set up to further dialect investigation. Later, professional institutes, such as the Institute of History and Philology at Academia Sinica, carried out surveys and collected first-hand dialect data. Several classic publications illustrate the remarkable achievements of this period. Representative of these are *Xiandai Wuyu Yanjiu* (“Research on Modern Wu Dialects,” Chao 1928); *Fangyan Diaocha Fangfa* (“Methodology of Dialect Research,” Cen 1956) and *Fangyan Diaocha Fangfa Gailun* (“General Introduction to Research Methodology of Dialects,” Cen 1956). As a result of this work, Chinese Dialectology was well established as a discipline in parallel with the allied disciplines of Chinese Exegetics, Philology, and Phonology.

34.2.3 The Boom Period of Chinese Dialectology

The campaign to popularize *Putonghua* was launched in 1956. Although this national movement reflected the government’s concerns about using local dialects, it coincided with a program of research on those dialects, including a full survey of Chinese dialects initiated during this period. To facilitate the fieldwork, a series of professional manuals were made available to researchers. These include: *Fangyan Diaocha Zibiao* (“Dialect Survey Forms,” actually a word table called “Questionnaire of Characters for Dialect Surveys”),² *Fangyan Diaocha Cihui Shouce* (“A Handbook for Research on Dialectal Lexicon,” Ding 1989); *Hanyu Fangyan Diaocha Jianbiao* (“Worklist for Dialectal Research,” Ding and Li 1956); *Hanyu Fangyan Diaocha Shouce* (“A Handbook for Research on Chinese Dialects”, Li 1957), *Fangyan Diaocha Cihuobia* (“Word Table for Chinese Dialects”) (Beijing University, 1989); *Hanyu Fangyan Gaiyao* (“General Introduction to Chinese Dialects,” Yuan 1960); and, *Hanyu Fangyan Zihui* (“Collections of Chinese Dialectal Characters”). These documents were very instructive at the time and are still helpful for today’s research. They stimulated a boom period of research on Chinese dialects, beginning with the first issue of the *Dialect Journal*, published by the Chinese Academy of Social Sciences, which appeared in 1979. Later publications, such as *The Atlas of Chinese Languages* (Li 1987), *Xiandai Hanyu Fangyan Dacidian* (“Great Dialect Dictionary of Modern Chinese, in Series,” Li 2002), and *Xiandai Hanyu Fangyan Yinku* (“Audio Database for Modern Chinese Dialects,” Hou 1994, 2003) exemplify the more recent achievements of this boom period.

34.2.4 Contemporary Chinese Dialectology and Dialect Preservation Work

Chinese dialect research continues to develop today and now benefits from technological advances such as the development of specialized computer software for dialect research. Research results are now presented in many forms, including both text and audio files. According to Cao Zhiyun, the current state of Chinese dialect research shows the following developments: 1) The number of researchers continues to increase; 2) Academic seminars, conferences, and congresses are frequently held; 3) Methodologies have multiplied; and 4) Technological advances have allowed the application of more scientific techniques of data collection and analysis (Cao 2012).

Since 1955, when *Putonghua* was officially prescribed as the national common language of all ethnic communities in the territory of the People’s Republic of China, Chinese dialects have faced significant challenges. Some dialects are dying out, while others are splitting into smaller groups and still others are leveling with neighboring varieties. Since “language is the

heart of a culture," (Ladefoged n.d.), and dialect is therefore the heart of local or regional culture, conservation work is very important. Dialects are often seen as the most important part of Chinese people's cultural heritage.

34.3 The Classification of Chinese Dialects

Research on Chinese dialects involves studying historical phonology, fieldwork investigation, database construction, atlas production, and classification of dialect groups. Exactly how to classify individual dialects in groups is one of the most pivotal and controversial questions in current Chinese dialectology. Diverse criteria and multiple classificatory dimensions have led to different groupings. Moreover, the taxonomic relations of Chinese dialects are dynamic, as reflected in the results of recent research; for example, the newly defined dialects of Jin, Hui, and Pinghua raised the number of dialectal groups from 7 to 10 (see discussion below).

One basic issue is variability in the definition of "dialect" by different groups of scholars, such as anthropologists, historical linguists, sociolinguists and those concerned with history, politics and other extra-linguistic matters. In China, all of the sub-varieties of Han speech³ were called *Fangyan* (the Wade–Giles system spells it *Fangyen*), or "dialects," in contrast to *Putonghua*, or standard speech (which is based on the Mandarin dialect of China's capital city, Beijing). The word *fangyan* is a compound of *fang* ("direction, locality, side, place, region, area") and *yan* ("speech, talk, language, word, saying"). Thus, *fangyan* means "regional speech" and is usually translated into English as "dialect," which has led to some confusion. Some European and Western dialectologists prefer to consider some Chinese "dialects" to be "languages" (Mair, 1991:6), because some of them are mutually unintelligible. Alternative terms, such as "regionialect" (Defrancis, 1984: 57) and "topolect" (Mair 1991: 7), were suggested to address the mistranslation of *fangyan*. "Regionalects" refer to mutually intelligible sub-varieties of Chinese speech, whereas "topolect" is used when there is no clear-cut difference between "language" and "dialect" (Groves 2008). Although the definition of "dialect" involves criteria of mutual intelligibility, sound pronunciation, phonological parameters, and etymological cognates, Chinese speech varieties are called "dialects" due mainly to their unified writing system and shared (cognate) morphemes. For instance, the Yue dialects of southeastern China, which include Cantonese, the prestige variety of Yue spoken in Guangdong province and in nearby Hong Kong, as well as by many people of Chinese ancestry in North America, are generally not mutually intelligible with the Mandarin dialects spoken in northern and southwestern China because of wide-ranging differences at the phonological, syntactic and lexical levels.

There are many disagreements about classifying Chinese dialects. Chinese dialectologists have devoted a great deal of work to establishing a relatively accurate classification scheme, but the rich diversity of Chinese dialects is very complex. First, the hierarchical structure of dialectal families is difficult to determine. Second, there are multiple criteria for grouping dialects, which result in diverse dialectal groupings. Influenced by western sinologists who suggested a hierarchical model of language varieties (i.e., family-group-branch-language-dialect-subdialect; Mair 1991), Chinese experts proposed several layers of classification in a family tree of Chinese dialects, based on the latest research: supergroup (*dàqū*); group (*qū*); subgroup (*piān*); cluster (*xiǎopiān*); and local dialect (*diǎn*; Li 1987).

34.3.1 Criteria for Classification

The review of previous research that was used to sort out the early classification of Chinese dialects focused mainly on two factors: geographical distribution and phonological evolution. Chinese dialect division along geographic lines can be traced back to Yang Xiong's

time; for example, the dialects of his own state of Chu were classified based on the North-South division of the state at that time. In the early twentieth century, Zhang Taiyan (also Zhang Binglin), Li Jinxi, and Chao Yuen-Ren developed more specific divisions, as shown in Table 34.1. The water systems (river: “jiang”, “he”; lake: “hu”, sea: “hai”) were used as the reference points for dialect description and grouping.

Another approach to Chinese dialect classification examines the phonological and phonetic development of phonemes in each dialect. In the evolution from Middle Chinese to Modern Chinese, variability in initial and coda segments and in tone establishes important dialect divisions. In the 1930s, the major division between Mandarin and non-Mandarin was addressed by Wang Li, who was regarded as the first dialectologist to move beyond impressionistic classification criteria by using phonological features, especially the evolution of Middle Chinese voiced initials (cf. Wang 1936).

The number of major groups of Chinese dialects established has varied depending on the different phonological characteristics used to classify them. Wang Li established five groups in the 1930s by using major phonetic characteristics (cf. Wang S-Y 1996: 249, Yan 2006). Li Fang-Kuei then proposed eight major groups based on phonological features of Middle Chinese (Li 1937: 1–13). Ting Pang-hsin (Ting 1982: 258) adapted seven major groups according to 17 different evolutionary features of Middle Chinese. Ramsey (1987) classified Chinese dialects into Mandarin and non-Mandarin groups, including Wu, Xiang, Gan, Kejia (Hakka), Yue, and Min. Norman (1988: 182) grouped Chinese dialects into three major branches according to 10 phonological, grammatical, and lexical features. The Northern group included Northern Mandarin, Southern Mandarin, Northwest Mandarin, and Southwest Mandarin; the Central Group comprised Xiang, Wu, and Gan; and the Southern Group was made up of Kejia (Hakka), Yue, and Min. Lau (2002) proposed a new dialect classification with four groups: Northern, Wu, Min, and Gan-Yue. *The Language Atlas of China* proposed 10 (super)groups and detailed subgroups (Li 1987). These are shown in the family tree diagram of Figure 34.1).

In summary, several conclusions can be obtained from Table 34.1 and Figure 34.1:

1. All of the groups of dialects evolved from Han Chinese, their common ancestral “parent”, with Northern Mandarin as a backbone, as shown in Figure 34.1.
2. Chinese dialects show a primary split into Mandarin and non-Mandarin groups, both of which, along with their sub-branches, are in flux. The number and names of the divisions vary in different accounts. For example, the Mandarin group has been called *Beifanghua* (“Northern speech”), “Northern,” and so on (Yuan 1960, Huang 1987, Norman 1988, Lau, 2002).
3. An important discrepancy relates to the Min group. Yuan Jiahua (1960) bifurcated Min into South Min and North Min. Zhan Bohui (1981) suggested a tripartite division: North Min, East Min, and South Min. Huang Jinhu (1987) further split Min into East Min, South Min, Puxian, Central Min, and North Min.
4. The Xiang dialect was separated from the Mandarin branch and then split into Old Xiang and New Xiang (Zhan 1981).
5. Kejia (Hakka) and Gan were once merged as a Ke-Gan group and then re-separated.
6. Jin became an independent dialect from the Mandarin branch. Pinghua and Huiyu, shown in parentheses in Figure 34.1, are newly recognized dialects and need to be further studied before they can be confidently classified (see Tang 2009: 32–33).

Another approach to phonological classification focuses on tone features, since Chinese is tonal language. Modern Chinese dialects developed from Middle Chinese, which underwent a tone split from four basic categories: *Ping* (level tone); *Shang* (rising tone); *Qu* (falling tone); and *Ru* (entering or checked tone). Each of these split into two registers, called *Yin* (upper register with voiceless initials) and *Yang* (lower register with voiced initials;

Table 34.1 Chinese dialect classification based on geographical distribution. Numbers in the top row are column labels to aid in comparison of the different schemes. (Refer to Yan 2006: 8–9; Cf. Wang L. 1981: 487; Wang S-Y 1996: 249; He 1995: 414–415.)

Scholar (down)	Number (across)												
	1	2	3	4	5	6	7	8	9	10	11	12	
Zhan (1915)	HeBei-ShanXi	Shaan-Gan	YuDong-Shandong-JiangHuai		Chuan-Yun-Qian-Gui	YuNan-E-Xiang-Gan		GuangDong		FJ	SuH	Hui	
Li (1937)	HeBei	HeXi	HeNan	JiangHuai	JinSha	JiangH	JHu	YueHai		MH	OH	TaiHu	Zh-Y
Chao (1934)	Mandarin Groups					HuaNan (South-East) (including Xiang, Gan)		non-Mandarin Groups					
	HuaBei (North-East) (including Jin)							Kejia	Yue	Hai-Nan	Min	Wu	
Chao (1934)	Mandarin Groups			XiaJiang	Shangjiang (Upper Yangtze)			non-Mandarin Groups					
	Beifang					Xiang	Gan-Kejia	Kejia	Yue	ChSh	Min	Wu	Wan
Chao (1928)	Beifang			XiaJiang	XiNan (South-West)				Yue		Min	Wu	
										S	N		
Chao (1928)	Mandarin Groups			XiaJiang	XiNan (South-West)		non-Mandarin Groups						
	Beifang					Xiang	Gan	Kejia	Yue		Min	Wu	Hui
										S	N		

Key to abbreviations: FJ = Fujian; SuH = SuHang; JiangH = JiangHan; JHu = JiangHu; MH = MinHai; OH = Ou'Hai; Zh-Y = ZheYuan; ShangJ = ShangJiang (Upper Yangtze); ChSh = ChaoShan; XiaJ = Xiajiang (Lower Yangtze); N = North; S = South; Kejia (Hakka); Beifang = Beifanghua (Northern Mandarin speeches), XiNan = South-West.

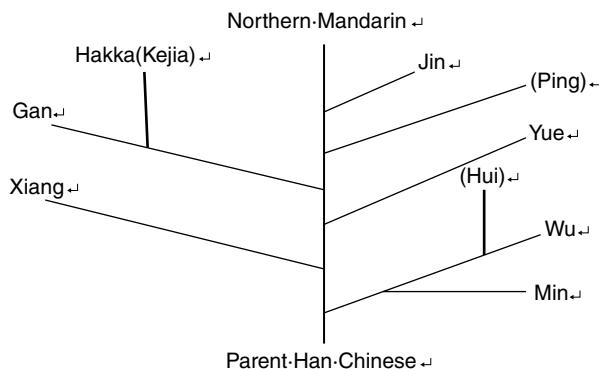


Figure 34.1 Family tree diagram showing the branching of Chinese dialect groups from Han Chinese, their common ancestral “parent.”

Cheng 1973: 95), thereby generating eight tones. Over time, some Yin tones merged with other Yin tones or with Yang tones, whereas others split into more tone units. The merger or split of tones in Yin and Yang registers determines the main division of dialects into Mandarin and non-Mandarin groups (Cheng 1973, 1991). Mandarin dialects reduced the number of tones to five, whereas the non-Mandarin dialects retain more than five tones, along with complex syllable patterns and checked codas (Li 1987).

34.3.2 Measures of the Relationship Between Chinese Dialects

Efforts to measure the degree of relation among Chinese dialects have relied on objective, subjective, and mixed measures. Objective measures have often focused on phonological features shared by pairs of dialects, or on a combination of lexical and phonological correspondence, as in the early lexicostatistical calculation of Wang (1960). Further examples of this lexical-phonological approach are found in the work of Cheng (1973, 1982, 1991). Genealogical relationship is established based on phonological correspondence among cognates' syllable structures, involving initial, onset, final and tone segments (Cheng 1991, 1993, 1997). Later, structural measures of relationship were based on listeners' judgments of production data from dialect pairs. The results of this work showed that the basic split between Mandarin and non-Mandarin branches matched for the most part the structural difference based on syllable structure.

A more subjective approach was taken by Tang and van Heuven (2007a, 2007b, 2008, 2009), who measured (dis)similarities between Chinese dialects based on the listener judgments of recordings of native speakers' pronunciations of Swadesh-like basic vocabulary items. The results generally confirmed previous taxonomies of major groups and sub-groups, but Changsha (Xiang) and Nanchang (Gan) were ambiguously split between Mandarin and non-Mandarin groups (for more details, refer to Tang 2009: Chap. 3).

Tang and van Heuven also tried mixed methods to correlate both the various objective and subjective measures with the taxonomy of dialect groupings. The results showed again that the primary split was basically correct, but several subgroups or clusters were always problematic. Taiyuan (Jin) consistently deviated from Mandarin (details in Tang 2009: Chap. 5).

Structural distance measures can be used to indicate the closeness of the relationship between Chinese dialects, but the results do not always correlate accurately with phonological features. Moreover, neither structural nor phonological proximity may correspond definitively with the genetic relations between dialects. Of course, dialects differ at many linguistic levels (phonetic, phonological, lexical, morphological, syntactic, etc.)

The inconsistent and dynamic discrepancies in the classification of Chinese dialects create a dilemma. Structural methods based on phonological characteristics are too complex, and in most cases the measures used involve many dimensions. Mutual intelligibility is generally used as the most important criterion for distinguishing languages from dialects and is therefore often applied as a primary means to measure the closeness between dialects. Intelligibility between dialects is not always symmetrical: that is, Dialect A can be more intelligible to Dialect B than Dialect B is to Dialect A, which creates a problem for mutual intelligibility as a criterion in dialect classification. Linguists therefore often use the “mean” intelligibility between two dialects to indicate their proximity. Another challenge is that intelligibility itself is a gradient concept. Tang and Van Heuven (2007a, 2007b, 2008, 2009) worked out a method for measuring the mutual intelligibility of 15 Chinese dialects, in which they obtained scaled values of mutual intelligibility via functional testing of listeners’ understanding of words and sentences produced by 15 native dialect speakers. They then correlated the results with several structural measures from previous research and with their own calculation of Levenshtein distance (cf. Yang and Castro 2009) and used the resulting model to predict the relationships among the dialects. Correlation between subjectively judged and objectively measured similarities is also computed and discussed by Tang (2009), who was the first to apply mutual intelligibility measures to Chinese dialects.

34.4 Sources of Data on Chinese Dialects and Their Methods

The achievements of Chinese dialectology include many sources of data on Chinese dialects, employing various methods of data collection and analysis, which are reviewed in this section. They can be divided into general categories.

34.4.1 Missionary Documents

In the 1900s and even earlier, European and American missionaries came to China to share their religious beliefs and to promote what they saw as the interests of the Chinese people. In order to do their work, these missionaries had to learn Chinese dialects from local people, as *Putonghua* was not as widely used then as it is today. Their archives, including handwritten notes, diaries, manuscripts, annotations, reports, correspondence, books, monographs, journals, magazines, and textbooks, contain many descriptions of Chinese dialects and are thus valuable sources of data on the speech of that time. A good example is a bilingual textbook (with the title of *Chinese Lessons for First Year Students in West China*) of the Sichuan dialect (a southwest type of Mandarin, with the Chengdu variety as its representative), compiled by Omar Leslie Kilborn, M.D, the founder of the West China Hospital (Kilborn, 1917). Grootaers (1948, 1958) recorded many local dialects in manuscript form, based on fieldwork in Zhangjiakou, Wanquan, and Xuanhua counties of the Hebei Province.

34.4.2 Dialect Dictionaries or Monographs

Many dialect dictionaries and monographs have been prepared by Chinese dialectologists, for example, *Hanyu Fangyan Cihui* (“Lexicon of Chinese Dialects”) and *Hanyu Fangyin Zihui* (“Phonic Collection of Chinese dialectal Words”). The former is a large lexical database that contains 905 common words in Standard Mandarin and their equivalents (very often but not always expressed by cognate words) in 18 dialect localities grouped in 7 regions: MANDARIN (Beijing, Ji’nan, Shengyang, Xi’an, Chengdu, Kunming, Hefei, and Yangzhou);

WU (Suzhou and Wenzhou); XIANG (Changsha); GAN (Nanchang); HAKKA (Meixian); YUE (Guangzhou and Yangjiang); and MIN (Xiamen, Chaozhou, and Fuzhou). The latter is a collection of phonic data with 2,400 words from 17 dialects in 7 groups in the first edition: MANDARIN (Beijing, Ji'nan, Xi'an, Taiyuan, Hankou, Chengdu, and Yangzhou); WU (Suzhou and Wenzhou); XIANG (Changsha and Shuangfeng); GAN (Nanchang); HAKKA (Meixian); YUE (Guangzhou and Yangjiang); and MIN (Xiamen, Chaozhou, Fuzhou, and Jian'ou). The second edition includes 3,000 words from 20 dialects: in the MANDARIN group, Hefei was added and Hankou changed into Wuhan; Yangjiang was added to the YUE group and Jian'ou to the MIN group. The data are arranged in alphabetical order by their Pinyin form in Modern Chinese, phonologically annotated in Qieyun (a rime dictionary of Classical Chinese Phonology) of Middle Chinese, and transcribed with IPA. This collection is very helpful in making comparisons between the phonetic and phonological features of Standard Chinese and dialects, pairs of dialects, or Modern and Middle Chinese. Both the *Cihui* and *Zihui* collections were produced by the Centre for Chinese Linguistics at the Department of Chinese Language and Literature of Peking University, led by Wang Futang (Chief editor).⁴

34.4.3 Dialect Atlases

The most authoritative atlases of Chinese dialects are the *Language Atlas of China* (Li 1987) and the *Linguistic Atlas of Chinese Dialects* (Cao 2008). The former is focused mainly on the classification of Chinese dialects. It includes 35 color maps, among which is a general map of Chinese dialects (A1) and maps for each of 16 subgroups of dialects (B1-B16). A digital version of this atlas is now available from Wikipedia (http://en.wikipedia.org/wiki/Language_Atlas_of_China). The latter atlas is similar to the former but focuses on the geography of Chinese dialects. It contains 510 maps in three volumes, concerning various dialect features, including phonetics (205 maps), lexicon (203 maps), and grammar (102 maps). An example of a map showing the geographic distribution of Chinese and non-Chinese languages, including the major “dialects”, or regional varieties, of Chinese, is shown in Figure 34.2.

34.4.4 Audio Files

Xiandai Hanyu Fangyan Yinku (literally “Dialect Sound Database of Modern Chinese”) is a remarkable set of auditory data on 40 Chinese dialects. It was completed after the publication of *The Language Atlas of China*. The data were made available either in cassette tape or on CD-ROM, along with a book for each dialect (Hou 2003). Data on each dialect include phoneme inventory, core vocabulary, example sentences, and the reading of the fable, “The North Wind and the Sun” in this dialect (Hou 1994, 2003; Tang, 2009). Another auditory database, the “Audio Database of Chinese Language Resources,” initiated by National Languages Committee (also the Mandarin Promotion Council, or the National Languages Promotion Committee) is still under construction. This will be an enormous resource, including the current status of Chinese and ethnic minority languages and dialects, along with dialect-accented *Putonghua*, in the form of an audio database.

34.5 Remaining Questions and Future Work

Contemporary Chinese dialectology has many interesting directions to focus on in the future. Perhaps the most important is the construction and completion of a comprehensive national corpus and database to support further analysis and research. Also important, given the current popularization of *Putonghua*, is the preservation of dialects, as well as undertaking surveys of changes in the use of local dialects, with a view to rescuing endangered varieties.



Figure 34.2 Map of Chinese and Non-Chinese languages, with locations of major regional varieties of Chinese (reproduced by permission from <http://www.dartmouth.edu/~chinese/maps/map2a.html>).

Another area still open to further research is the methodology for testing the mutual intelligibility of Chinese dialects. Finally, work on oral speech at the text level and correlation of the results of structural comparison of dialects with mutual intelligibility is still in progress.

34.6 Discussion and Conclusions

Strictly speaking, many dialects in China, especially in the non-Mandarin branch, cannot be defined as “dialects of Chinese” in a strict sense, because they are not mutually intelligible with Mandarin varieties. However, as discussed above, there are nevertheless several reasons to retain the term “dialects”: (1) Linguistically, Chinese dialects can be grouped together by

similarities in their phonological evolution from Han Chinese (thus the use of the term “Sinitic”) and by their geographical proximity in situations of language contact; (2) Politically, Chinese dialects are considered to be variants of the same parent Han language, given their common phonetic inventory, writing system and grammatical rules. This remains a difficult issue: should contemporary mutual intelligibility be the main criterion in modern dialect classification? We can recognize these varieties as dialects according to their written form, which is common to all of them; however, in making this classification we may have to accept only partial and often asymmetrical mutual intelligibility between some pairs of dialects. On the other hand, some of the inconsistencies between structural similarity and mutual intelligibility might be explained by non-linguistic factors, such as socio-political matters or even experimental defects in the research on intelligibility.

As important as this issue may be (see Crystal 2000:8 and Hudson 1980:35–37 for further discussion), modern research on Chinese dialects focuses more on developing accurate and reasonable grounds for classification. Generally speaking, today’s accounts of Chinese dialect groups fall into two general patterns: either validating the traditional groupings or showcasing new classifications. In either case, however, the criteria for grouping are always in question. Since the criteria are various and multi-dimensional, no single criterion can authorize a definitive division. The best approach would seem to be a composite index, including traditional phonological features, a scale of mutual intelligibility, native speakers’ judgments, experimental data and political and social functions as well. Such an approach might finally resolve some of the debates about the correct classification of Chinese dialects.

NOTES

- 1 In 1955, *Putonghua* (literally “common speech”) was confirmed as a lingua franca in P. R. China after a long discussion. It is thus officially prescribed as a national common language of all ethnic speech communities of China. Other terms, such as “Standard (Modern) Chinese,” “Standard Mandarin Chinese,” or simply “Mandarin” are also used. Outside the Chinese mainland, other terms are used: *guóyǔ* (literally “national language”) mainly in Taiwan and *huáyǔ* (“Chinese language”) in Singapore and Malaysia. In the United Nations, *Putonghua* is often called simply “Chinese.”
- 2 *Questionnaire of Characters for Dialect Surveys* is literally translated as “*Dialect survey forms*,” which is actually a word table designed for gathering information about Chinese dialect pronunciation. The main editors were Chao Yuan-ren, Ding Shengshu, and Li Rong. It was then modified by the Chinese Academy of Sciences Institute of Linguistics in July 1955 and published by Beijing Science Publishing House. This version is based on “*Dialect survey forms*” as the master copy compiled by the Institute of History and Philology at Academia Sinica in 1930 (Refer to Tom Pearl on <http://baike.soso.com/v8561783.htm>).
- 3 Han speech refers to the Chinese language spoken by the Han people, a native ethnic group, which accounts for approximately 90% of China’s population. *Han* comes from the Han Dynasty, which succeeded the short-lived Qin Dynasty that united China, thus *Hanyu* (literally, “Han language”) is descended from an ancient common Han language.
- 4 In my previous publications (Tang and van Heuven 2006, 2007, 2008, 2009, and Tang 2009), I gave a different Pinyin spelling and English translation for this; I hope this version is more accurate and specific.

REFERENCES

- | | |
|---|---|
| Chao, Yuen Ren. 1928. <i>Studies in the Modern Wu Dialects</i> . Tsinghua College Research Institute Monograph, 4, Beijing. | Cao, Zhiyun. 2008. <i>Linguistic Atlas of Chinese Dialects</i> . Beijing: The Commercial Press. |
|---|---|

- Cao, Zhiyun. 2012. The Future Prosperity of Chinese Dialect Research. *Language Teaching & Learning and Research.* (5): 86–92. (in Chinese).
- Cen, Qixiang. 1956. *Fangyan diaocha fangfa* (*Methodology of Dialect Research*). Wenzigaike chubanshe (language reform publishing house). (in Chinese).
- Cheng, Chin-Chuan. 1973. *A Synchronic Phonology of Mandarin Chinese*. The Hague: Mouton. By Chung, Raung-fu. (tran.) 1994. (Translated into Chinese: 國語的共時音韻) Taipei: Crane Publishing.
- Cheng, Chin-Chuan. 1982. A quantification of Chinese dialect affinity. *Studies in the Linguistic Sciences* 12: 29–47.
- Cheng, Chin-Chuan. 1991. Quantifying affinity among Chinese Dialects. In William Wang S-Y (ed.) *Journal of Chinese Linguistics – Monograph series*, 3: 78–112.
- Cheng, Chin-Chuan. 1993. (With Zhiji Lu) Chinese dialect affinity based on syllable initials. *Studies in the Linguistic Sciences* 15, 127–148.
- Cheng, Chin-Chuan. 1997. Measuring Relationship among Dialects: DOC and Related Resources. *Computational Linguistics & Chinese Language Processing* 2: 41–72.
- Crystal, David. 2000. *Language Death*. Cambridge: Cambridge University Press. ISBN 0-521-65321-5.
- DeFrancis, John. 1984. *The Chinese Language: Fact and Fantasy*. Honolulu: University of Hawaii Press.
- Ding Shengshu, Rong Li. 1956. *Hanyu fangyan diaocha jianbiao*. [Worklist for Chinese Dialectal Research], Beijing: Zhongguo kexueyuan yuyan yanjiusuo [Institute of Linguistics, Chinese Academy of Sciences].
- Ding Shengshu. 1989. *Fangyan diaocha cihui shouce*. [Lexicon Handbook of Dialects Research], *Fangyan(Dialects)*. (2): 91–97.
- Grootaers, Willem A. 1948. Problems of a Linguistic Atlas of China. *Leuvense Bijdragen*, 38: 52–72.
- Grootaers, Willem A. 1958. Linguistic Geography of the Hsuan-hua region. *Bulletin of the Institute of History and Philology*. 29(1): 59–86.
- Groves, Julie M. 2008. "Language or Dialect – or Topolect? A Comparison of the Attitudes of Hong Kongers and Mainland Chinese towards the Status of Cantonese", *Sino-Platonic Papers* 179: 1–103.
- He, Jiuying. 1995. *A History of Chinese Modern Linguistic*. Guangdong Education Press. ISBN 7-5406-3399-9.
- Hou, Jinyi. 1994. *Xiandai hanyu fangyan yiku*. (in Chinese) Shanghai: Shanghai Education Press.
- Hou, Jinyi. 2003. *Xiandai hanyu fangyan yiku*. (CD-Rom version, in Chinese) Shanghai: Shanghai Education Press.
- Huang, Jinhu. 1987. *Hanyu Fangyanxue*. (Chinese Dialectology). Xiamen University Press.
- Hudson, R. A. 1980. *Sociolinguistics*. [M]. Cambridge: Cambridge University Press.
- Kilborn, Omar Leslie. 1917. *Chinese Lessons for First Year Students in West China*. Published by the Union University. (in the University of Toronto Library)
- Ladefoged, Peter. n.d. "Preserving the sounds of disappearing languages." URL: <http://www.linguistics.ucla.edu/people/ladefoge/Preserving%20sounds.pdf>.
- Lau, Chun-fat. 2002. Hanyu Fangyan de fenlei biaozhun yu "Kejiahua" zai Hanyu fangyan fenlei shang de wenti [The Criteria for Chinese Dialect Classification and the Problem of the Position of the "Hakka dialect" in Chinese Dialect Grouping]. *Journal of Chinese Linguistics* 30: 82–96. (in Chinese).
- Li, Fang-kuei. 1937. *Languages and Dialects of China*. The Chinese Yearbook. Shanghai: Commercial Press (Reprinted in *Journal of Chinese Linguistics* 1: 1–13).
- Li, Rong. 1957. *Hanyu Fangyan Diaocha Shouce*, [A Handbook for Research on Chinese Dialects]. Beijing: Science Press.
- Li, Rong. 1987. Languages in China. In: Wurm, Stephen A., Benjamin T'sou, David Bradley, Li Rong, Xiong Zhenghui, Zhang Zhenxing, Fu Maoji, Wang Jun and Dob (eds.), *Language Atlas of China*. Jointly compiled by the Chinese Academy of Social Sciences and the Australian Institute of Humanities. Hong Kong: Longman, Map A-1.
- Li, Rong. 2002. *Xiandai Hanyu Fangyan Dacidian* (Great Dialect Dictionary of Modern Chinese, in Series). Nanjing: Jiangsu Jiaoyu chubanshe [Jiangsu Education Press].
- Mair, Victor H. 1991. "What Is a Chinese "Dialect/Topolect"? Reflections on Some Key Sino-English Linguistic terms", *Sino-Platonic Papers* 29: 1–31.
- Norman, Jerry R. 1988. *Chinese*. Cambridge University Press.
- Ramsey, S. Robert. 1987. *The Languages of China*. Princeton: Princeton University Press.
- Tang, Chaoju and Vincent J. van Heuven. 2007a. Mutual intelligibility and similarity of Chinese dialects, in Bettelou Los & Marjo van Koppen (eds), *Linguistics in the Netherlands 2007*. Amsterdam: John Benjamins, AVT (24): 223–234.
- Tang, Chaoju and Vincent J. van Heuven. 2007b. Predicting Mutual Intelligibility in Chinese Dialects. *Proceedings of 16th International Congress of Phonetic Sciences*. www.icphs2007.de: ICPHSXVI:1457-1460.

- Tang, Chaoju and Vincent J. van Heuven. 2008. Mutual Intelligibility of Chinese dialects tested functionally. *Linguistics in the Netherlands*. AVT (25): 145–156.
- Tang, Chaoju and Vincent J. van Heuven. 2009. Mutual Intelligibility of Chinese dialects experimentally tested. *Lingua*. 119(5): 709–732.
- Tang, Chaoju. 2009. *Mutual Intelligibility of Chinese dialects: An experimental approach*. LOT dissertation series nr. 228. Utrecht: LOT.
- Ting, Pang-Hsin. 1982. Hanyu Fangyan qufen de diaojian. [Phonological features for classification of the Chinese dialects] *Tsing Hua Journal of Chinese Studies* (14): 257–273.
- Wang, Li. 1936. *Zhongguo Yinyunxue*. [Phonology of Chinese]. Beijing: Shangwu Yinshuguan. [The Commercial Press].
- Wang, William S-Y. 1996. Linguistic Diversity and Language Relationships. In: Huang and Li (eds.), *New Horizon in Chinese Linguistics*. 235–267. Boston: Kluwer Academic Publishers.
- Wang, Yude. 1960. Chugokugo dai hogen no bunretsu nendai no gengo nendai gakuteki shitan (Lexicostatistic estimation of time depths of five major Chiense dialects). *Gengo Kenkye* 38: 33–105.
- Yan, Margaret Mian. 2006. *Introduction to Chinese Dialectology*. LINCOM Studies in Asian Linguistics.
- Yang, Cathryn, and Andy Castro. 2009. Representing tone in Levenshtein distance. *SIL International Media Release*.
- Yuan, Jiahua. 1960. *Hanyu Fangyan Gaiyao* [An Outline of the Chinese Dialects]. Beijing: Wenzi Gaige Chubanshe.
- Zhan, Bohui. 1981. *Xiandai Hanyu Fangyan*. [Modern Chinese Dialects] Wuhan: Hubei People's Press.

35 Dialects of Japanese

TAKUICHIRO ONISHI

35.1 Introduction

Japanese is a language isolate, with no proven genetic relation to any other language. Although Chinese and Korean are spoken in areas near Japan and there are many Chinese loan words in Japanese, no language family including both Japanese and either Korean or Chinese has been verified. A historical comparison of Japanese with other East Asian languages, including Ainu (an indigenous language spoken on Japan's northern island of Hokkaido), has been attempted, but there is no clear evidence pointing to extensive similarities or genetic relationship.

Japanese is spoken mainly by about 130 million people in Japan, where it is the sole official language and is spoken by 99 percent of the population (who are overwhelmingly of Japanese ethnicity). Japan spans four main islands (including Honshu, the largest island and site of the capital, Tokyo) and thousands of smaller islands in an archipelago off the northeast coast of Asia. Part of this archipelago is the Ryukyu region, which extends from Japan south toward Taiwan, including Okinawa. This region is politically part of Japan today but was an independent country until 1879. The origin of the history of Ryukyu country (Kingdom of Ryukyu) is not clear, but it is thought to be at least the beginning of the fifteenth century. After the Meiji Restoration, Ryukyu was brought under Japanese rule. This political movement is called *Ryukyu Shobun*, which means “disposition of Ryukyu.” The Amami region was the northern part of the Kingdom of Ryukyu. In 1609, in the Edo era before the Meiji Restoration, the Amami region was conquered by Satsuma-han, a domain with great military strength. Ryukyuan is a language spoken in the Ryukyu region, and it is the only language confirmed to have common origins with Japanese. Although historically, Ryukyuan is an independent language, it has been treated as a dialect of Japanese since the Ryukyu region became part of Japan. The area of the Ryukyuan language includes the Okinawa prefecture, but it is wider than the prefecture, as will be made clear in the discussion that follows.

While dialects of Japanese, including Ryukyuan, are diverse, they have some common features:

- Phonologically, the basic structure of the Japanese syllable is open: (C)V. Most syllables consist of a consonant and a vowel or of a single vowel; combinations of more than two consonants are rare, and single-consonant syllables or syllables closed by a consonant are limited to the cases of *sokuon* (doubled consonants: for example, *tatta* “stood”) and *hatsuon* (syllabic nasal: that is, *nonda* “drank”).

- Word order in sentences is subject, object, verb (SOV), as in, *watashi-ga mizu-o nondai*, lit. "I water drank."
- Verbal conjugation forms express various grammatical meanings: *nom-u* (infinitive of "drink"), *nom-a-nai* (negative of *nomu*), *nom-da* (past tense of *nomu*), *nom-e* (imperative of *nomu*).
- To show grammatical relations, postpositional case markers are used: *watashi-ga* (I, subject), *mizu-o* (water, object), *omae-ni* (you, object).

35.2 The Main Studies and Sources

Research on Japanese dialects is very extensive. The following review, which divides the studies and sources of data by type, is therefore necessarily selective.

35.2.1 Descriptive Studies and Handbooks

Nihonhogen-no Kizyutsuteki Kenkyu. [Descriptive Study of Japanese Dialects.]

Kokuritsu Kokugo Kenkyujo [The National Language Research Institute of Japan (presently, National Institute for Japanese Language and Linguistics)] 1959 ***Nihonhogen-no Kizyutsuteki Kenkyu.*** [Descriptive Study of Japanese Dialects.] Meizishoin, Tokyo.

The National Language Research Institute of Japan has carried out descriptive studies of the dialects of each prefecture since it was established in 1949. All of these are maintained in the institute as manuscripts. Fifteen of them have been selected and published in this book. Each paper was written by a representative dialectologist of the region (i.e., Fukuo Ishigaki of Hokkaido, Kazutami Nishimiya of Nara, and Kan'ichi Itoi of Oita) or of an era (e.g., Takeshi Sibata and Yukio Uemura for the 1950s–1970s). Descriptive objects and fields are basically limited to phonology and grammar, and each paper is 20 pages on average. The descriptions are not long, but basic items are described and the material retains its importance today.

Hogengaku Koza. [Dialectological Handbooks.]

Tojo, Misao. 1961 ***Hogengaku Koza.*** [Dialectological Handbook.] 4 volumes. Tokyodo, Tokyo.

Volume 1 of these handbooks contains introductory material for less advanced students. Volumes 2 to 4 describe dialects according to Tojo's regional divisions: volume 2 for the Tobu [eastern] dialect, volume 3 for the Seibu [western] dialect, and volume 4 for the Kyushu and Ryukyu dialect. Many dialectologists, especially students, have long preferred this series, since each paper is concise and written simply.

Koza Hogengaku. [Handbooks of Dialectology.]

Ryoichi Sato and Kiichi Iitoyo ed. 1982–1986 ***Koza Hogengaku.*** [Handbooks of Dialectology.] 10 volumes, Kokushokankokai, Tokyo.

This series has a very similar name to the previous one but was published 20 years later. In that time, Japanese dialectology developed and young scholars became veterans. With advances in dialect geography and sociolinguistics, new and more detailed information about each dialect became necessary. This new series met these needs. Volumes 1 to 3 treat the guidance, methodology, and perspective of dialectology. Volumes 4 to 10 describe each dialect based on the prefecture where it is spoken: Hokkaido and Tohoku (Vol. 4); Kanto (5); Chubu (6); Kinki (7); Chugoku and Shikoku (8); Kyushu (9); and Amami and Okinawa (10). This series is still available and is used widely as a source of basic data on Japanese dialectology.

35.2.2 Atlases

There are two linguistic atlases that cover all of Japan. These differ in several respects: it is said that the *Linguistic Atlas of Japan* (LAJ) contains interpretational maps, whereas the *Grammar Atlas of Japanese Dialects* (GAJ) contains maps of material.

LAJ

Kokuritsu Kokugo Kenkyūjo [The National Language Research Institute of Japan (presently, National Institute for Japanese Language and Linguistics)] 1966–1974 (reprinted in reduced size 1981–1985) *Nihon Gengo Chizu*. [*Linguistic Atlas of Japan*.] 6 volumes. Okurasho Insatsukyōku, [The Printing Bureau of the Ministry of Finance,] Tokyo.

The *Linguistic Atlas of Japan* (LAJ), with 300 maps, was published between 1966 and 1974 and reprinted in reduced size between 1981 and 1985. The research for LAJ was conducted between 1957 and 1965 at 2,400 locations; the birth year of informants was before 1903. The main items of LAJ are lexical, for example, words for “snail,” “potato,” and “corn,” although some phonetic items are also included. Following publication of LAJ, linguistic geography became mainstream in Japanese dialectology and many atlases of smaller areas, as well as many papers, were subsequently published.

GAJ

Kokuritsu Kokugo Kenkyūjo [The National Language Research Institute of Japan (presently, National Institute for Japanese Language and Linguistics)] 1989–2006 *Hogen Bunpo Zenkokuchizu*. [*Grammar Atlas of Japanese Dialects*.] 6 volumes. Okurasho Insatsukyōku and Kokuritsu Insatsukyōku, [The Printing Bureau of the Ministry of Finance and the National Printing Bureau,] Tokyo.

The *Grammar Atlas of Japanese Dialects* (GAJ), with 350 maps, was published between 1989 and 2006. Its research was carried out between 1979 and 1982 at 807 locations, each represented by one person older than 65. The GAJ’s main contribution was the mapping of grammatical variation: the LAJ had treated few grammatical items and the *Kogoho Bunpuzu*, an earlier grammatical atlas (Kokugochosaiinkai 1903), was limited by inconsistent editing and mapping methods.

GAJ made all aspects of map making, such as data selection and marking up the legends of maps using raw data, systematic, in contrast to LAJ, which was criticized for not having clarified these issues. Beginning with volume 5, the GAJ’s maps were made with computers. Following its publication, all of the GAJ’s maps and related material were made accessible on the Web (see http://www2.ninjal.ac.jp/hogen/dp/dp_index.html). The GAJ data are convenient to analyze and map on personal computers, and the geolinguistic patterns associated with its grammatical variables are clearer than those connected with lexical items. It has inspired many papers about grammatical distributions and analysis of dialect grammars.

35.2.3 Dictionaries

Japanese dialectology has progressed through the efforts not only of professional scholars, such as professors or students in universities, but also of amateur dialectologists in local areas, or dilettantes of dialects and folklore, who have published many dictionaries of local dialects. Some of the most important Japanese dialect dictionaries are listed below.

SDJD

Munemasa Tokugawa and Ryoichi Sato ed. 1989 *Nihon Hogen Daiziten*. [*Shogakukan’s Dictionary of Japanese Dialects*.] 3 volumes. Shogakukan, Tokyo.

Shogakukan’s Dictionary of Japanese Dialects (SDJD) was edited using data from almost 1,000 local dictionaries of dialects (most of these edited by amateur dialectologists), papers on

dialectology, and classical documents treating dialect words. Most of the data for the dictionary were gathered by Masanaka Oiwa. There were two dictionaries (Tojo, 1951, 1954b) before SDJD, but, unlike the SDJD, their data were not complete.

DJD

Hirayama, Teruo. 1992–1994 *Gendai Nihongo Hogen Daiziten*. [Dictionary of Japanese Dialects.] 9 volumes. Meizishoin, Tokyo.

The *Dictionary of Japanese Dialects* (DJD) treats variants of basic vocabulary items from 72 places representing all dialect areas of Japan. Although its material is limited to basic vocabulary, the description of each item includes its meaning, accent, and an example of its use in a sentence. Volume 1 gives a brief description of each of the 72 regional dialects. Volumes 7 to 9 provide an index to the main body of the dictionary (volumes 1 to 6), making it possible to search for particular words.

35.2.4 *Bibliographies*

The first bibliography of Japanese dialectology was published in Tojo (1944). Subsequent reference lists were edited by the *Nihon Hogen Kenkyukai* or Dialectological Circle of Japan (CDJ 1964, 1978). CDJ (1990) contains a complete list, and CDJ (2005) is an updated list with a CD-ROM. In addition, every issue of *Conference Papers of the Dialectological Circle of Japan* includes a list of books (not including papers) on dialectology, and the recent lists (since 2004) can be downloaded from the CDJ's website: <http://dialectology-jp.org/>.

35.3 Divisions and Features of Japanese Dialects

Misao Tojo, author of the *Dialectological Handbooks* mentioned above, is considered the founder of modern Japanese dialectology. He claimed that the purpose of dialectology is the classification of dialects and the mapping of borders between them (Tojo 1944). Although subsequent discussion in the CDJ (1964) focused on arguments for and against various divisions of dialects and many ideas for divisions were reported, divisional studies of dialects have declined since the publication of CDJ (1964). Kato (1990) proposed that the reason for the decline is both an excess of data on the dialects and the unclear methodology of their division. The main focus of Japanese dialectology subsequently shifted to dialect geography. Nevertheless, the standard division of Japanese dialects, shown in the map in Figure 35.1, was established by Tojo (1954a). He proposed a primary division between Hondo (mainland) and Ryukyu dialects. Next, he divided the Hondo group into three main regions, Tobu, Seibu, and Kyushu, each comprising several dialects: Tobu was divided into five dialects, Seibu into five dialects, and Kyushu into three dialects. The Ryukyu region was also divided into three dialects. The remainder of this section describes the phonological and grammatical features of each dialect compared to standard Japanese. It is important to note, however, that geographic divisions of these features usually exhibit clines or transition zones rather than the sharp breaks implied by isoglosses.

35.3.1 *The Hondo Dialect Group*

The Hondo or “mainland” dialects, a group comprising the Tobu, Seibu, and Kyushu regions, have five vowels—/a, i, u, e, o/—and have no distinction between the infinitive and attributive of verbs and adjectives.

35.3.1.1 *The Tobu Dialect Region*

In the Tobu or “eastern” region, vowels are weakly articulated and consonants are strong, thus high vowels between voiceless consonants are sometimes voiceless, as in [sita] “tang,” [kusa] “grass,” and [teika] “underground.” The Tobu accent system is the Tokyo pattern. This

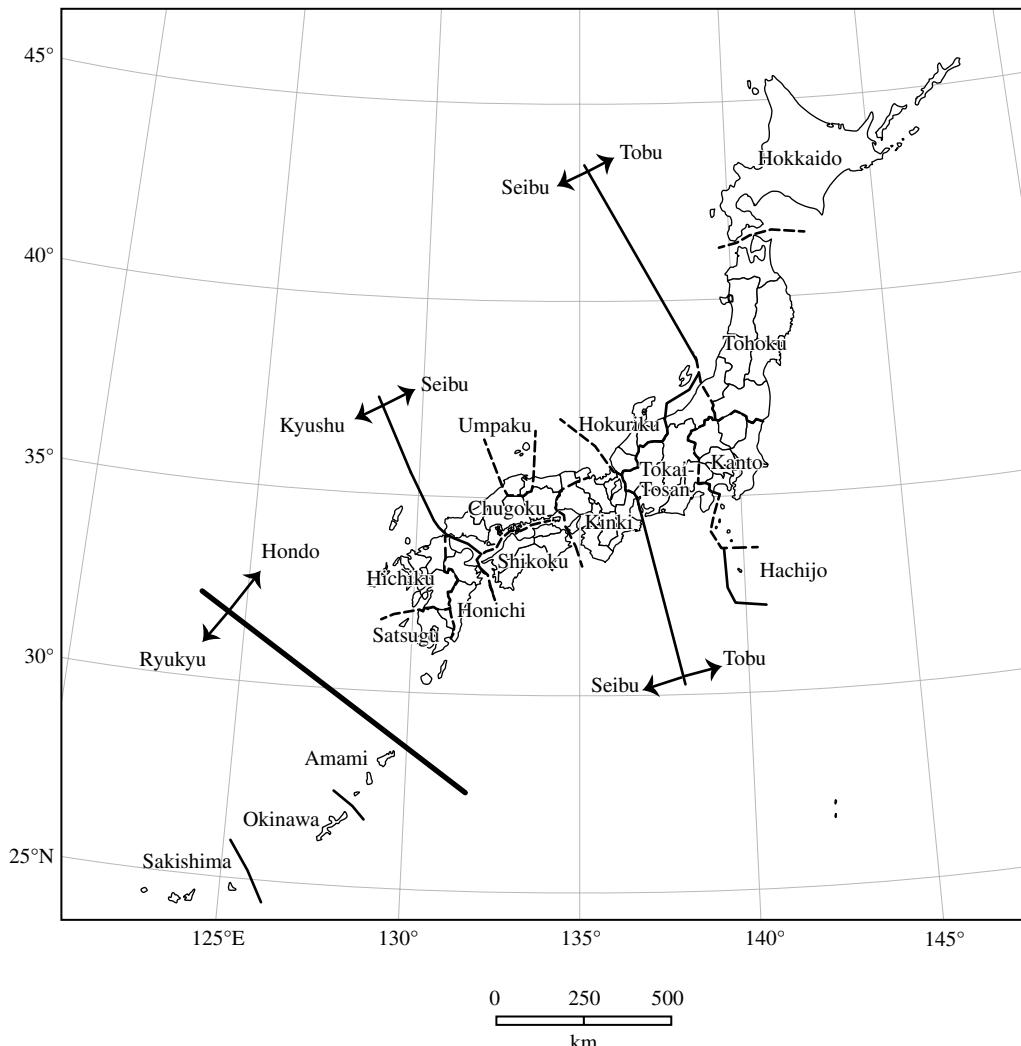


Figure 35.1 Dialect regions of Japan.

has a dropping accent core,¹ which is defined by the position where the accent shifts from high (H) to low (L) and is marked with ['] following the accented syllable: [karada] “body” is pronounced with the LHH accent pattern, [a'sahi] “rising sun” with the HLL pattern, and [tama'go] “egg” with the LHL pattern. The negative suffix /nai/ is realized as [ne:], and the copula is [da] in most places.

35.3.1.1.1 Hokkaido Dialect

Beginning at the northern end of Japan, the Hokkaido dialect is similar to that of the neighboring Tohoku region of Honshu, since Tohoku people immigrated to Hokkaido in the Edo era, introducing new words like *shitakke* “thus” and *ssho* “may” into Hokkaido.

35.3.1.1.2 Tohoku Dialect

Phonetic changes are conspicuous in the Tohoku dialect. The vowels /i/ and /e/ are merged, both having the lower quality, pronounced [e̞] so *i* “stomach” and *e* “picture” are homophones, and the contrast between /i/ and /u/ is neutralized after /s/, so *si* “four” and *su*

"vinegar" are both [su̥i]. The voiceless stops /k/ and /t/ are voiced between vowels, as in [iga] "cuttlefish" and [hada] "flag," whereas voiced stops are nasalized or prenasalized, as in [iziŋo] "berry" and [haŋda] "skin." Directions are shown by the postpositional case marker *sa*, as in *kita-sa* "to the north." Past tense is expressed by *ke*, especially for adjectives such as *takeke* "was expensive."

35.3.1.1.3 Kanto Dialect

The Kanto dialect is very similar to standard Japanese, since the region includes Tokyo, formerly called Edo, and standard Japanese, established after the Meiji Restoration, was based on the Edo dialect. Some local features, nevertheless, differ from standard Japanese. The diphthong /ai/ changes to [e:], as in [take:] "expensive." *Bee* is used to express intention and conjecture: *korekara iku-bee* "I am thinking of leaving now," and *aitsu-mo iku-bee* "I think he will leave." *Bee* was well known as a Kanto dialect word in old times, and the speech of the Kanto people was called *Kantobee* in the classical literature of the Edo era.

35.3.1.1.4 Tokai-Tosan Dialect

The Tokai-Tosan dialect (named after old main roads) is located near the boundary between the Tobu and Seibu regions, so it has some Seibu features and the voiceless vowels do not occur so frequently. The suffixes expressing conjecture are *ra* and *zura*. *Ra* can be combined with verbs and adjectives, but not with nouns, thus:

<i>aitsu-ga</i>	<i>iku-ra</i>
pron. 3-subj.	v. conjecture "go"
"I think he will go."	
<i>ashita-wa</i>	<i>samui-ra</i>
n.-focusing marker "tomorrow"	adj. conjecture "cold"

"I think it will be cold tomorrow."

Zura combines with nouns, verbs, and adjectives:

<i>aitsu-wa</i>	<i>gakusei-zura</i>
pron. 3-focusing marker	n. conjecture "student"
"I think he is a student."	
<i>aitsu-ga</i>	<i>iku-zura</i>
pron. 3-subj.	v. conjecture "go"
"I think he will go."	
<i>ashita-wa</i>	<i>samui-zura</i>
n.-focusing marker "tomorrow"	adj. conjecture "cold"

"I think it will be cold tomorrow."

The difference between *ra* and *zura* is not only in the parts of speech they combine with, but also in their meanings: *zura* expresses a cause for the conjecture, and *ra* does not express a cause so clearly. The copula *da* can directly follow verbs and adjectives: *aitsu-ga iku-da*, *ashita-wa samui-da*, although standard Japanese does not combine both verbs and adjectives with the copula.

35.3.1.1.5 Hachijo Dialect

The Hachijo dialect is spoken on Hachijo and Aogashima Islands, whose population is declining markedly; the dialect is therefore endangered. In Hachijo speech, infinitive and attributive forms of verbs and adjectives are different. *Nomowa* "drink" (infinitive of the verb

"drink") is different from *nomo toki* "drinking time" (attributive of the verb "drink"). *Takakja* "high" (infinitive of the adjective "high") differs from *takake yama* "high mountain" (attributive of the adjective "high"). Conjecture is expressed with *nouwa*. *Nomu-nouwa* means "I think (he) will drink." This, along with other features, makes Hachijo very different from the other dialects in Tōbu. The main reason for classifying Hachijo with other Tōbu dialects is that the features of infinitival and attributive forms described above are similar to those of the ancient eastern dialects recorded as *Azumauta* (song of eastern area) in *Man'yoshū*, which is a songbook edited in the eighth century.

35.3.1.2 The Seibu Dialect Region

Seibu or "western" dialects contrast strongly with those of the Tōbu region. Seibu vowels are strongly articulated, so the voiceless vowels do not occur as frequently. The Seibu accent system consists of "dropping" the accent core and tones. The Kyoto accent pattern (representative of the Seibu region) has both the high tone and low tone. The accent core and tone are combined in the dialect's accent. The high tone can have the HHH pattern, as in [sakana] "fish," the HLL pattern, as in [i'nochi] "life," and the HHL pattern, as in [omo'te] "front." The low tone has the LLH pattern, as in [usagi] "rabbit," and the LHL pattern, as in [tama'go] "egg." The negative suffix is *n* or *hen*, and the copula is *ya* or *zya*.

35.3.1.2.1 Hokuriku Dialect

The Hokuriku dialect shares with the Tohoku dialect the neutralization of the /i-u/ contrast after /s/, so /si/ and /su/ sound the same. This dialect has a characteristic winding intonation, which shows what the segments of a sentence are and can be used after each segment. The Hokuriku dialect has a well-developed system of honorifics. It is not rare to find three stages of difference in honorifics. For example, speakers in Gokoyama, which is in the southern area of the Toyama prefecture, use *nomassharu* as the highest stage of the verb "drink", *nomyaru* as the higher stage, and *nomu* as the normal stage.

35.3.1.2.2 Kinki Dialect

The Kinki region includes big cities such as Osaka, Kyoto, and Kobe, so its dialect is usually taken as representative of the larger Seibu region and is well known, since some Kinki media personalities use it on television and radio even when broadcasting from Tokyo. In the central area of the Kinki dialect, *hen* is used for the negative suffix, and *ya* is used for the copula.

35.3.1.2.3 Chugoku Dialect

Grammatical features of the Chugoku dialect are typical of the Seibu region. The negative suffix is *n*, and the copula is *zya*. Both of them are older than the *hen* and *ya* of the Kinki dialect. Aspects are expressed by compound verbs with *oru* "exist" or "be." For example, the compound verb *chirioru* "falling" is from *chiri* "fall" and *oru*, and the aspectual meaning is that of continuance. When *te* is inserted between the verb and *oru*, as in *chichchoru* (<*chiri-te-oru*>), the aspectual meaning changes to that of result. The Chugoku accent is of the Tokyo type, despite the fact that the Chugoku dialect is classified as belonging to the Seibu region.

35.3.1.2.4 Umpaku Dialect

As in the Tohoku and Hokuriku dialects, /si/ and /su/ are not distinguished in the Umpaku dialect. Another phonological feature is that /r/ is frequently dropped, as in [okiru] "to get up" > [okiu] > [okii]. The copula is *da*, as in the Tōbu region. These features make the Umpaku dialect an unusual variety in the Seibu region.

35.3.1.2.5 Shikoku Dialect

The accent in the Shikoku dialect is very complicated since there are more than two tones that are also found in the Kyoto accent pattern and there is another falling tone.

The differences between /zi/ and /di/ and /zu/ and /du/, which are lost in standard Japanese, are preserved in the southern part of the Shikoku dialect. Thus, [ɸuzi] “Fuji” (the name of a mountain) is different from [ɸudi] “a flower of wisteria,” and [zu] in, for example, [kazu] “number”, is distinguished from [du] in, for instance, [midu] “water.”

35.3.1.3 The Kyushu Dialect Region

Older features of the Seibu region are preserved in the dialects of the Kyushu region. Examples include preserving the bigrade² vowel alternations in verbal conjugations and phonological contrasts between /zi/ and /di/ and /zu/ and /du/. On the other hand, the Kyushu region also shows innovative changes, such as the monograde to quadric /r/ grade change, as in, for example, *okira-n* < *oki-n* “not to get up.” The name of each Kyushu dialect follows old country names; for example, Hichiku is named after Hizen, Higo, Chikuzen, and Chikugo; the Umpaku dialect of the Seibu region (above) is named in a similar way.

35.3.1.3.1 Hichiku Dialect

The Hichiku dialect area includes big cities such as Fukuoka (the central area is called Hakata), Kita-Kyushu, and Kumamoto, so the Hichiku dialect is generally taken to be representative of the Kyushu region. In this dialect, the final element of the infinitive of adjectives is not *i* but *ka*: *yoka* “good” (not *yoi*), *waruka* “bad” (not *warui*), and *kuraka* “dark” (not *kuroi*). This *ka* changed from *ku-aru*; *yoka* is from *yoku* “good” (continuative form of the adjective), and *aru* is historically “be” (a verb meaning existence). The lower bigrade of verbs is well preserved (e.g., *aku-ru* “open”, *ake-n* “do not open”, *uku-ru* “accept”, *uke-n* “do not accept”). On the other hand, the upper bigrade changes to monograde, as in *oki-ru* “get up” (if it was a bigrade, it would be *oku-ru*), and there is a further change from monograde (m) to quadric r (qr). This is by analogy (*okiru* (m): X= *toru* “get” (qr): *toran* (negative), X= *okiran*), since the final form of the monograde and quadric r is the same and the number of words belonging to the quadric r type is very large (the number of monograde verbs is less than 10).

35.3.1.3.2 Honichi Dialect

Older forms are preserved in the Honichi dialect. In the conjugations of verbs, both of the upper bigrade and lower bigrade are kept: *ok-u-ru* “get up,” *ok-i-n* “do not get up,” *ak-u-ru* “open,” *ak-e-n* “do not open.” Another Honichi feature is *kakarimusubi*, a grammatical construction where, following certain postpositions, verbs and adjectives appear in conditional or attributive forms, but not in the infinitive form. An especially telling example is the conditional form corresponding to the postposition *koso*. In most dialects, this form has been lost, but it is still preserved in the Honichi dialect, as demonstrated in the example below.

<i>ware-ko</i>	<i>sogee</i>	<i>hayoo</i>
pron.1-postposition < <i>koso</i>)	adv. “such”	adj. continuative “early”
<i>ok-u-re</i>	<i>hoka-no</i>	<i>mon-wa</i>
v. conditional “get up”	n.-case marker “other”	n.-focusing marker “people”
<i>mada</i>	<i>nechora-ya</i>	
adv. “still”	v. progressive conjectural “may be sleeping” – sentence final particle	

“Even if you get up so early, other people are still sleeping.”

If *ko* (<*koso*) is dropped, a conjunction such as *domo* is needed, but in the *koso*-conjectural structure conjunctions are not required.

35.3.1.3.3 Satsugu Dialect

In the Satsugu dialect, final syllables *ki*, *gi*, *ku*, *gu*, *chi*, *zi*, *tsu*, *bi*, *bu*, and verb-final *ru* become /t/: *kaki*>*kat* “persimmon,” *kagi*>*kat* “key,” *kaku*>*kat* “to write,” *kagu*>*kat* “to smell,” *kachi*>*kat* “victory,” *kazi*>*kat* “fire,” *katsu*>*kat* “to win,” *kabi*>*kat* “mildew,” *tobu*>*tot* “to fly,” and *toru*>*tot* “to get.” As a result, the Satsugu dialect has syllables closed with /t/, a remarkable divergence from the usual Japanese (C)V syllable type. This feature makes the Satsugu dialect difficult to understand for speakers of other Japanese dialects. The accent pattern of Kagoshima, which is representative of the Satsugu dialect, has only tones, without an accent core: it has two tones, which are final rising and final falling. For example, the final rising LH tone is found in the pronunciation of *hana* “flower,” the LLH tone in the pronunciation of *abura* “oil,” and the LLLH tone in the pronunciation of *yomikata* “way to read”; the final falling tone HL is found in the pronunciation of *hana* “nose,” the LHL tone in the pronunciation of *kuruma* “wheel,” and the LLHL tone in the pronunciation of *kamaboko* “a food made from fish.”

35.3.2 The Ryukyu Dialect Group

In contrast to the five-vowel system of the Hondo dialect group, dialects of the Ryukyu group have only three vowels, /a/, /i/, and /u/, with the Hondo mid vowels, /e/ and /o/, raised to /i/ and /u/, respectively, in Ryukyu speech, though in the Amami and Okinawa dialects, as discussed below, new mid vowels have been created from diphthongs. Most dialects distinguish between infinitive and attributive forms of verbs and adjectives. Ryukyu dialects are only spoken on the Ryukyu Islands. Each island has little contact with the other islands, so dialect diversity is considerable in this region.

35.3.2.1 Amami Dialect

The Amami dialect region is in the Kagoshima prefecture, so Amami and Kagoshima (part of the Satsugu dialect discussed above) share a large number of lexical items. The five vowels of the Hondo dialect, /a/, /i/, /u/, /e/, and /o/, correspond to the Amami vowels as follows: /a/, /i/, and /u/ are the same in both dialects, but the Hondo /e/ corresponds to Amami [i], for example, [ami]<*ame* “rain,” whereas the Hondo [o] corresponds to Amami [u], for example [u]ja<*oya* “parent”. In addition, the Hondo diphthongs /ai/ and /ae/ correspond to [ë] in Amami, for example, [fë]<*hae* “fly,” and Hondo /ao/ corresponds to Amami [o]: [o]<*ao* “blue.” As a result of these changes, the Amami dialect has seven vowels: /a, i, ë, u, e, ë, o/.

35.3.2.2 Okinawa Dialect

The Okinawa dialect is considered representative of the Ryukyu region, since Okinawa is the largest island and has the largest population. Correspondences with Hondo vowels are as follows. Hondo /a/ corresponds to [a] in Okinawa: [kutuba]<*kotoba* “language;” Hondo /i/ to Okinawa [i]: [ifi]<*ishi* “stone;” Hondo /u/ to Okinawa [u]: [umi]<*umi* “sea;” Hondo /e/ to Okinawa [i]: [ki]<*ke* “hair;” and Hondo /o/ to Okinawa [u]: [munu]<*mono* “goods.” Diphthongs correspond as follows: Hondo /ai/ and /ae/ correspond to Okinawa [e]: [ke]<*kai* “shell,” [me]<*mae* “front;” and Hondo /ao/ to Okinawa [o]: [so]<*sao* “pole.” Thus, the Okinawa dialect has five vowels: /a, i, u, e, o/. Infinitive and attributive forms of verbs follow the pattern of *kachun* (infinitive of “to write”) and *kachuru* (attributive of “to write”). These final forms of verbs (-*un* and -*uru*) are thought to have originated from the verb *oru* “(animate) exist.” On the other hand, an example of the infinitive form of an adjective is *takasan* “high,” and an example of the attributive form is *takasaru*. The final forms of adjectives (-*an* and -*aru*) are thought to originate from the verb *aru* “(inanimate) exist.”

35.3.2.3 *Sakishima Dialect*

The Sakishima dialect has preserved a [p] corresponding to Hondo /h/. It is thought that both sounds originate from an earlier */p/. Vowels on Yonaguni Island are limited to /a, i, u/, without the new mid vowels created from diphthongs in other Ryukyu dialects, making it the only Japanese dialect with only three vowels. Until recently, there were no detailed descriptions of the Sakishima dialect, especially of its grammar, but since UNESCO registered it as endangered, some excellent descriptions have been done, with new grammatical categories and syntactic structures being reported. These studies inspire researchers in other dialect regions.

35.4 Current Research

35.4.1 *Descriptive Studies*

In 2009, UNESCO designated the Yaeyama and Yonaguni dialects as severely endangered, while the Hachijo, Amami, Kunigami, Okinawa, and Miyako dialects were designated as definitely endangered. Yaeyama, Yonaguni, and Miyako belong to the Sakishima dialect, and Kunigami and Okinawa are part of the Okinawa dialect (UNESCO's Okinawa does not correspond to the divisional dialects of Okinawa). As mentioned in the previous section, studies describing the grammars of Ryukyu dialects began about a decade ago. Izuyama (2005) is the first paper describing evidentiality as a grammatical category in the Sakishima dialect. Shimoji (2006) treats a special kind of aspect in Tarama, which belongs to the Sakishima dialect. Inspired by these papers, many other descriptive studies have been published. Shimoji (2009a) describes adjectives in the Irabu dialect, also a Sakishima dialect, whereas Shimoji (2009b) is about the expression of modality in Irabu.

These studies are excellent but concentrated on traditional-style analyses. For endangered dialects, these traditional studies need to be supplemented by appropriate methods of documentation. Shimoji (2010, 2011) shows how documentation can be done.

When the Great East Japan Earthquake happened on March 11 in 2011, it caused scholars to notice that endangered dialects are not limited to the Ryukyu region. The Dialectological Circle of Japan organized a special session during its Ninety-Fourth conference in May 2012. Some projects are trying to record and preserve dialects in Tohoku (Center for the Study of Dialectology of Tohoku University 2012).

35.4.2 *Geolinguistic Studies*

Geolinguistics was the mainstream in Japanese dialectology in the 1970s and 1980s, with more than 400 atlases and 30,000 maps published (Onishi 2010a). After the 1990s, however, the popularity of geolinguistics declined. The GAJ was first published in 1989 and the database made available online in 2000. The process of making dialect maps using computers began at the same time, making the process much simpler than when maps were prepared by hand and also allowing maps to be revised as needed. At the same time, geolinguistics began to use geographical information systems (GIS). With this advance, large quantities of linguistic data and extralinguistic information can be viewed as layers, and the relationships between the types of data can be verified. Geolinguistics has aimed to clarify the spatial distributions of linguistic phenomena using these calculations (Onishi 2010b).

Real-time comparison of dialect distributions across several intervals or timespans is currently an active field of research. As noted above, Japanese geolinguistics was very popular in the 1970s and 1980s. Interval maps are now used to compare the areas mapped in those times to areas mapped more recently. These make it possible to analyze the changes in dialects and dialectal distributions that have happened over the last 30 to 50 years. This analysis will shed light on the dialect formation process (Onishi in print).

35.4.3 *Sociolinguistic Studies*

Before the geolinguistic developments just discussed, Japanese sociolinguistics conducted local real-time or interval research in some cities. Especially noteworthy is that conducted by the National Institute for Japanese Language and Linguistics (NINJAL, formerly National Language Research Institute: NLRI) in Tsuruoka in 1950, 1971, 1991, and 2012; by Okazaki in 1953, 1972, and 2008; and by Furano in 1959 and 1986. This research used interviewees selected by random sampling as well as those who participated in earlier research. It produced analyses of both language changes in each city in chronological time, and language changes during the lifetime of each informant. Even if the general tendency is a change to standard Japanese, people in cities still pick up and use the local dialects (NLRI1974).

35.5 Japanese Dialectology in the Future

Japanese dialectology was developed by concentrating on the linguistic study of dialects. On the other hand, dialects are spoken in local regions by ordinary people and are influenced by society, history, and other extra-linguistic factors. Dialectologists need to study these additional elements that are external to language, in addition to language variation and dialectical changes.

Updated methods of documentation are needed to deal with the problem of endangered dialects. Traditional linguistics mistakenly believed that developing models or theories through descriptive studies was the most important purpose of their research, but this is not the most important thing for endangered dialects: what they require most is records and documentation, which can provide a refreshed sense of the value of linguistics in the modern world (Nakayama 2009).

GIS in the context of geolinguistics connects dialectal distributions and nonlinguistic factors, such as demographic data, elevation, roads, and railways, as well as highlighting dialectical distributions in mathematical and statistical ways. In the same way, sociolinguists study both language data and nonlinguistic factors using statistics. The most important viewpoint is the relationship between social networks and language variation and change, which should continue to be a main focus of both geolinguistics and sociolinguistics.

Though many linguists believe that the purpose and goal of dialectology is only in its contribution to general linguistics, in my opinion, dialectology has concentrated on linguistics too much. In the near future, dialectology, including descriptive, geographic, and socio-logical studies, has a chance to establish itself as a separate discipline in the humanities. Dialectology should aim to elucidate the language of ordinary people and their life. Japanese dialectology can realize this objective by continuing to build on and pursue the long-term accumulation of knowledge by talented scholars.

NOTES

1 The accent core is a distinctive feature of an accent system. For instance, [karada] is distinguished from [a'sahi] and [tama'go] in excluding or including the accent core, and [a'sahi] and [tama'go] are distinguished by the locations of the accent core ([a'sahi] has its accent core on the first syllable, and [tama'go] has it on the second syllable). Several kinds of accent cores are known (i.e., dropping, rising, and falling), and they are different for each dialect.

- 2 Japanese grammar traditionally classifies the conjugations of verbs according to vowel alternations. The bigrade has two alternant vowels, as in the contrasts between *ake-n* “not to open,” *ake-ta* “opened,” *aku-ru* “to open,” and *ake-ro* “open!” In these forms, /e/ and /u/ alternate, which is called the lower bigrade, whereas an alternation of /i/ and /u/ is called the upper bigrade. A quadric grade has four alternant vowels, as exemplified in *dasa-n* “not to take out,” *dasi-ta* “took out,” *dasu* “to take out,” and *dase* “take out!” In this example, /a/, /i/, /u/, and /e/ alternate. A monograde has no alternations, and only one vowel appears, as in, for example, *mi-n* “not to see,” *mi-ta* “saw,” *mi-ru* “to see,” and *mi-ro* “see!”

REFERENCES

- Itzuyama, Atsuko. 2005. Ebidenshariti: Ryukyu, Sakishima Hogen-no baai. [Evidentiality in Sakishima dialect.] *Nihongogaku*. 24(14): 56–66.
- Kato, Masanobu. 1990. Hogenkukaku-no rekishi. [History of division of dialects.] Nihon Hogen Kenkyukai [Dialectological Circle of Japan (CDJ)] 1990 *Nihon Hogenkenkyu-no Ayumi*. [History of Japanese Dialectology.] Kadokawashoten. Tokyo: 173–188.
- Kokugochosaiinkai. 1903. *Kogoho Bunpuzu*. [Atlas of Grammar]
- Nakayama, Toshihide. 2009. Shinzidai-no kijutsugengogaku. [Descriptive linguistics at the new stage.] *Gekkan Gengo*. 38(7,8): 66–73, 68–75.
- National Language Research Institute. 1974. *Chiiki Shakai-no Gengoseikatsu: Tsuruoka-niokeru 20nenmae-tono Hikaku*. [Language Survey in Tsuruoka City, Yamagata Pref.: After 20 Years from the Preceding Survey.] Shuei Shuppan, Tokyo.
- Nihon Hogen Kenkyukai. [Dialectological Circle of Japan (CDJ)] 1964. *Nihon-no Hogenkukaku*. [Division of Japanese Dialects.] Tokyodo, Tokyo.
- Nihon Hogen Kenkyukai. [Dialectological Circle of Japan (CDJ)] 1978. *Nihon Hogen-no Goi*. [Lexicon of Japanese Dialects.] Sanseido, Tokyo.
- Nihon Hogen Kenkyukai. [Dialectological Circle of Japan (CDJ)] 1990. *Nihon Hogenkenkyu-no Ayumi*. [History of Japanese Dialectology.] Kadokawashoten. Tokyo.
- Nihon Hogen Kenkyukai. [Dialectological Circle of Japan (CDJ)] 2005. *Nizisseiki Hogenkenkyu-no Kiseki*. [Japanese Dialectology in 20th Century.] Kokushokankoukai, Tokyo.
- Onishi, Takuichiro. 2010a. Mapping the Japanese language. *Language and Space: An International Handbook of Linguistic Variation: Language Mapping*. Mouton De Gruyter. Berlin: 333–354.
- Onishi, Takuichiro. 2010b. Analyzing Dialectological Distributions of Japanese. *Dialectologia, Special Issue 1*: 123–135.
- Onishi, Takuichiro. In print. The Relationship between Area and Human Lives in Dialect Formation. *Congress Papers of 7th SIDG* (tentative). Preasens Verlag, Vienna.
- Shimoji, Kayoko. 2006. Ryukyu Taramajima hogen-no paafekuto-no keishiki. [Perfect forms in the Tarama Island dialect of Ryukyu.] *Nihongo-no Kenkyu*. 2(4): 76–90.
- Shimoji, Michinori. 2009a. The adjective class in Irabu Ryukyuan. *Nihongo-no Kenkyu*. 5(3): 33–50.
- Shimoji, Michinori. 2009b. Epistemic modality in Irabu Ryukyuan. *Shigen: Language Description Papers of Tokyo Foreign Language University*. 5: 25–42.
- Shimoji, Michinori. 2010. Firudowaku-ni dekakeyo. [Let's go to fieldwork.] *Nihongogaku*. 29(12): 26–30.
- Shimoji, Michinori. 2011. Bunpo Kijutsu-niokeru tekusuto-no juuyousei. [Importance of text in the description of grammar.] *Nihongogaku*. 30(6): 46–59.
- Tohoku Daigaku Hogen Kenkyu Sentaa. [Center for the Study of Dialectology of Tohoku University] 2012. *Hogen-wo Sukuu, Hogen-de Sukuu*. [Helping Dialects, Helped by Dialects.] Hitsuj Shobou, Tokyo.
- Tojo, Misao. 1944. *Hogen-to Hogengaku*. [Dialect and Dialectology.] Shun'yodo, Tokyo.
- Tojo, Misao. 1951. *Zenkoku Hogen Ziten*. [Dictionary of All Japanese Dialects.] Tokyodo, Tokyo.
- Tojo, Misao. 1954a. *Josetsu*. [Introduction.] *Nihon Hogengaku*. [Dialectology of Japan.] Yishikawakobundo, Tokyo: 1–86.
- Tojo, Misao. 1954b. *Bunrui Hogen Ziten*. [Thesaural Dictionary of Japanese Dialects.] Tokyodo, Tokyo.

36 Dialects of Malay/Indonesian

ALEXANDER ADELAAR

36.1 Malayic Languages and Dialects: Basic Facts

Malay, or Indonesian, as it is called in its standardized form in Indonesia, is arguably the most important language in South East Asia. It is the national language of Indonesia, Malaysia, Brunei, and Singapore. It is spoken by more than half of the inhabitants of Indonesia, Malaysia, and Brunei, which together number almost 280 million. It is also the mother tongue of more than 1 million Malays in the south of Thailand and of diaspora groups in Kampuchea, Sri Lanka, the Netherlands, and Australia.

Although of crucial importance as an official language and as a second language, Malay is only one of the many languages in the area. In Indonesia alone there are 719 languages, and in Malaysia 140. In fact, the number of Indonesians who speak a form of Malay as their first language is relatively small, and Malay mother tongue speakers are easily outnumbered by first language speakers of Javanese or even Sundanese, who together make up for almost half the population of Indonesia.

Malay belongs to the Austronesian language family, which has more than 1,200 members, most of which (including Malay) belong to the Malayo-Polynesian branch of Austronesian and are spoken in (mainly) insular South East Asia, the Pacific, and Madagascar. (Other Austronesian branches have only a very few other members, which are all spoken in Taiwan.) Within Malayo-Polynesian, Malay belongs to the Malayic subgroup, which includes Malay proper and a large variety of Malay dialects and Malay-like languages; this subgroup is defined by a set of phonological changes that its members have commonly undergone since Proto-Malayo-Polynesian (cf. Adelaar 1992: 2). Within the Malayic group there are several separate languages: some are distantly related to Malay, such as Iban and Kanayatn in western Borneo; others are much closer to Malay but structurally divergent enough to be considered languages in their own right, such as Minangkabau and Kerinci (both in Sumatra), Banjarese (South Borneo), Kelantan Malay (north east corner of West Malaysia), and Urak Lawoi' (Southwest Thailand), to name a few (see map in Figure 36.1 and discussion further below). The Malayic language group seems to be most closely related to Cham in Vietnam and Cambodia, and to Achehnese in North Sumatra (Thurgood 1999).

Nowadays, varieties of Malay are spoken throughout Malaysia, Brunei, Singapore, South Thailand, and Indonesia. However, Malay civilization traditionally belongs to the coastal regions of the South China Sea, and more particularly, Southeast Sumatra, the Malay peninsula, and West Borneo. The cradle of Malay civilization is generally taken to be South East Sumatra (in the regions including the cities of Jambi and Palembang). However, historical linguistic evidence indicates that the prehistorical homeland of the Malayic languages is in West Borneo (Adelaar 2004; Collins and Awang Sariyan 2006).

Sociolinguistically, the many varieties of Malay can be categorized into standard varieties, vernaculars, and regional lingua franca varieties (with some categorical overlap for some



Figure 36.1 Location of Malay and Malayan varieties referred to in this section.¹

varieties. Standard Malaysian Malay and Indonesian are basically the same language, although there are differences in vocabulary, pronunciation, and morphological valency. Standard Brunei Malay is very similar to Standard Malaysian Malay but uses different terms of address and reference. Vernacular Malay varieties are spoken in the traditional Malay regions. They generally show a greater typological diversity than the other two categories. Regional lingua franca varieties exist mainly in and around urban areas. They include among others Banjar Malay (in South Borneo), Palembang Malay (in South Sumatra), Kupang Malay (West Timor), Manado Malay (North Sulawesi), Ambon Malay (Ambon Island and surrounding areas), Papuan Malay (West Papua), Brunei Malay, and Jakarta Indonesian (which is to be distinguished from Jakarta Malay, see below). Some lingua franca varieties appear to be even more vibrant than the national languages in their respective regions, such as Kupang Malay, Jakarta Indonesian, Banjarese Malay, and possibly other regional linguae francae (see further below). Many of these varieties (especially the ones spoken outside the traditional Malay-speaking regions) have some specific typological features setting them off from the literary and national standards as well as from many vernacular Malay varieties. These features seem to be due to intensive contact with non-Austronesian languages in the distant past. The varieties in question have been labeled as Pidgin Derived Malay varieties by Adelaar and Prentice (1996) and as Vehicular Malay by others (cf. Paauw 2008), in reference to their past—if not current—role as a tool for interethnic communication.

From a historical perspective, then, some of the lingua franca varieties grew out of Malay vernaculars, whereas other varieties (including all eastern Indonesian ones) are forms of Vehicular Malay. Vehicular Malay typology consists of a configuration of features which are at least partly reflected in its individual representatives. It includes the following features (Adelaar 2011):

1. plural pronouns are based on *orang* “person, human being”;
2. possessive constructions involve a linker and are in a Possessor-linker-Possessed order;
3. progressive aspect is expressed with *ada*, the original existential marker;
4. *pigi* or *pi* (derived from *pergi*, a verb meaning “to go”) is used as a directional preposition “to”;
5. the demonstratives *ini* “this” and *itu* “that” have become reduced to *ni(h)* and *tu(h)* respectively and function as definite markers;
6. periphrastic causative constructions have replaced the original causative derivations;
7. much of the inherited Malayic morphology was lost;
8. significantly, this includes the loss of the original “symmetric” voice system².

Any categorization of Malay varieties is bound to be a loose one on account of influences due to regional contiguity and interaction between acrolect and basilect varieties. Throughout the centuries, Standard Malay varieties have had a constant influence on vernacular and vehicular varieties, and the latter have also exerted influence on one another and on standard varieties. Moreover, several lingua franca varieties have become mother tongues, especially in eastern Indonesia. This has inevitably had a blurring effect on the boundaries between these categories.

Malay has played a major role in the history of insular South East Asia. Standardized forms of Malay have been used as literary languages and official and/or national languages in many places. They have also been the main vehicles for the transfer and teaching of various forms of Indic, Islamic, and Christian religions. Such forms of Malay have often been used for the expression of religious, cultural, and scholarly ideas wherever these religions were established. In the meantime, vehicular forms of Malay became a means of interethnic communication all over insular South East Asia, and some of the vehicular and vernacular varieties developed into regional koines. Considering its application as both a cultural

language and a lingua franca, the role of Malay in insular South East Asia has arguably been more encompassing than that of Latin in Medieval Europe.

Malayic languages other than Malay such as Iban and Kanayatn are genetically distant members within the Malayic group. This distance also appears from the different structure and lexicon each of them has. However, there are also varieties that belong to the Malay group and are genetically much closer to Malay but nevertheless deserve a separate linguistic status because of some far-reaching structural peculiarities that they exhibit. This certainly applies to Kerinci Malay in Sumatra, which has a very different morphosyntax and may have undergone the effects of a non-Austronesian substratum, and to Sri Lanka Malay, which has developed a typical SOV structure through contact with Tamil and more recently Sinhala. The language/dialect distinction is sometimes random: Minangkabau, spoken on Sumatra, is considered a separate language, for which there is some justification linguistic grounds, but Minangkabau is in fact no more different from standard forms of Malay than are the Malay varieties spoken in its vicinity, with which it clearly forms a dialect continuum. By contrast, Kerinci Malay is usually classified as a Malay dialect in spite of its idiosyncratic morphosyntax (Steinhauer 2002). The separate language status of Minangkabau is primarily due to history, as it used to be the language of an important separate sovereignty in West Sumatra, and less to its linguistic distinctiveness. On the basis of the latter criterion, many other so-called Malay dialects could justifiably be considered languages in their own right. However, their speakers are generally united in a common identity, which is largely determined by a common history, religion and culture, with the fifteenth-century Islamic kingdom of Malacca and its court language as common reference points.

Malay civilization began to develop in Sumatra. The oldest written evidence for the existence of Malays consists of seventh-century inscriptions found in South Sumatra and on nearby Bangka Island. This area used to be controlled by Srivijaya, a seaborne empire. The political center of this empire was presumably situated in the region in or around the area where the city of Palembang is currently located (see map in Fig. 36.1), until it moved to the Malay peninsula and eventually to Malacca in the early second millennium.

The South Sumatran inscriptions are in a Brahmi-derived Indic script. Early Malay had a large percentage of Sanskrit vocabulary, part of which was gradually replaced by Arabic and (to a lesser extent) Iranian lexicon after the rulers of Malacca converted to Islam. In colonial times, Malay adopted many words from Portuguese, Dutch, and English. It also adopted a fair number of Chinese loanwords. At present, Indonesian and standard Malaysian Malay are still borrowing words from English and Arabic and have used Sanskrit and Arabic for coining new vocabulary. Sanskrit, Arabic, and European languages have also had an impact on the grammatical structure of Malay.

The Malay literary language at the court of Malacca in the fifteenth century AD and later on at the court of the Riau-Johore sultanate became a literary standard for written Malay elsewhere and ultimately also for the national language standards in Indonesia (Indonesian) and in Malaysia and Brunei (standard Malay). It may have its roots in an even older literary language, which presumably originated at the Malay courts of South Sumatra in the seventh century.

In the context of colonial history, in Indonesia the implementation of Indonesian has been remarkably successful (see below). But in spite of its general acceptance as the national language, to most Indonesians standard Indonesian sounds rather formal and unnatural in everyday conversation. An informal counterpart to this language has arisen in the form of Jakarta Indonesian. Basically a form of Indonesian, it tends to use emblematic features at various linguistic levels from Jakarta Malay or Betawi, the traditional Malay dialect of the Indonesian capital. The most important of these features are:

1. maintenance of historical schwa in last syllables (elsewhere in the Malayic subgroup historical schwa and */a/ have merged to a in this position);

2. loss of initial */s/ in grammatical function words;
3. merger of */-a/, */-ah/, and */-ay/ to /è/ [e] (for example, Jakarta Indonesian has *ampè* “until,” *ajè* “only, just,” *udè* “already (perfective marker),” and *amè* “with,”, where standard Indonesian has the corresponding forms *sampay*, *saja*, *sudah*, and *sama*);
4. use of the Hokkien-derived pronouns *gue* and *lu* as 1sg. and 2sg. pronouns, respectively;
5. substitution of a single suffix *-in* for the applicative suffixes *-kan* and *-i*;
6. a reduced morphology.

Jakarta Indonesian is a vibrant and fast growing non-formal counterpart to official Indonesian, and it is probably the most prominent non-standard variety in the country. It has become a prestige dialect even outside Jakarta and Java island, leaving a mark on regional Malay varieties elsewhere in Indonesia. It is also becoming a literary language in its own right (Djenar 2008). It is distinct from Jakarta Malay, although it is often confused with it because of the latter’s marked influence. Meanwhile, Jakarta Malay itself has become seriously endangered (Grijns 1991: 14).

The acceptance of standard Malay in Malaysia has been more problematic. Many Malaysians are ethnically Chinese or Indian and prefer English, which is also used among some urban Malays. It remains attractive as a global means of business and global communication. In the aftermath of a civil war in the 1950s, racial tensions erupted in the late 1960s. After negotiations between the several power groups within the country (mainly between Malays and Chinese), a compromise was reached in which the Malays were able to solidify their position and to enforce the use of their language in education and administration. The implementation of Standard Malay was even problematic among Malays themselves, as it was a highly engineered language and rather difficult to understand for L1 speakers of vernacular varieties.

In Brunei, Standard Brunei Malay is an adaptation of Malaysian Malay (see above) and has become widely accepted as an official and written language. At the same time, local Brunei Malay remains unchallenged as an everyday language among Bruneians. It is the mother tongue of 80% of the population and a lingua franca among Bruneians in general. It is also used in local ceremonies. Most Malays speak it, although a minority among them speaks Kadayan, which is slightly different from (local) Brunei Malay (Martin 1998).

36.2 Main Studies and Sources

The documentation of Malay and Malayan varieties is rather patchy. Fairly comprehensive overviews of the variety and geographic spread of Malay are included in the linguistic atlases by Wurm and Hattori (1981–1983) and Wurm, Mühlhäusler, and Tryon (1996). Nearly exhaustive accounts of the linguistic literature are the Malay bibliographies by Collins (1990, 1995a, 1995b, 1996) and the bibliography of Bornean languages by Blust and Smith (2014).

Considering the scope of Malayan varieties, Malay dialectology and dialect geography have generally remained rather underdeveloped sub-fields, although various elementary surveys were made by the Pusat Bahasa (Indonesia’s national institute of language planning) in the last three decades of last century. Three works stand out as exceptions: Grijns’ (1991) dialect geography of Jakarta Malay; Lauder’s (1993) dialect study of the Tangerang region near Jakarta; and Tamsin Medan’s unpublished dialect geography of Minangkabau (1980).

Turning to an overview of reference works concerning the individual members of the Malayan subgroup, we can begin with Sneddon’s (2010) Indonesian reference grammar and Stevens and Schmidgall-Tellings’ (2010) Indonesian-English dictionary, which are standard English-medium reference works for Indonesian. Reference works for Malaysian are

available through Malaysian Malay including the Kamus Dewan dictionary (2005) and the grammar by Nik Safiah Karim *et al.* (1993). No such sources exist for Brunei Malay: the small Brunei Malay–Malaysian Malay dictionary (Kamus 1991) is not comprehensive, mainly describing distinctive Brunei Malay words. For Iban there is an excellent dictionary by Richards (1981), a grammar by Asmah Haji Omar (1981) and a grammar sketch by Steinmaier (1999). There is also a detailed grammar description of Mualang (Tjia 2007), which is very closely related to Iban. The Salako (or Badameà) dialect of Kanayatn is accessible through Adelaar's (2005) grammar sketch and lexicon. Recent Minangkabau reference works are Moussay's (French-medium) grammar (1981) and dictionary (1995) as well as Nurlela Adnan, Ermitati and Rosnida M. Nur's (1994) Indonesian-medium dictionary. For Banjar Malay there is a Banjarese Malay–Indonesian dictionary by Abdul Jebar Hapip (2006), but no grammar in published form. The most comprehensive source for Jakarta Indonesian to date is Sneddon (2006). Grammars of Jakarta Malay are Muhadjir (1982) and Ikranegara (1980), and there is a Jakarta Malay–Indonesian dictionary by Abdul Chaer Mad'ie (1976).

Much progress has recently been made in the study of Vehicular Malay varieties. Substantial sources are Van Minde (Amboin Malay, 1997), Stoel (Manado Malay, 2005), Nordhoff (Sri Lanka Malay, 2009), Litamahuputty (Ternate Malay 2012), and Kluge (Papua Malay, 2014). Adelaar (1996) gives a short outline of Cocos Malay, spoken on the Cocos Keeling Islands (West Australia).

36.3 Past and Current Research Questions in the Study of Malay and Indonesian

36.3.1 “High” Versus “Low” Malay?

In Indonesia, the social stratification of Malay and the choice of a suitable sociolect to serve as a standard have been ongoing issues since early colonialism. They have manifested themselves in various ways throughout the Malay-speaking world.

The difference between “high” and “low” already appears in correspondence between the earliest European colonials and regional Indonesian rulers (Sneddon 2003: 62–65). After the Dutch had established one of their main South East Asian strongholds in Amboin City (on Amboin Island) in the seventeenth century, colonial administrators and missionaries were faced with the question of what form of Malay should be used for a Bible translation. A major controversy broke out between proponents of literary Malay, who argued that its elevated style was best suited for conveying the biblical message, and supporters of local Ambonese Malay, who countered that this Vehicular Malay variety was understood much more widely. The choice fell on literary Malay, with the inevitable result that many Christians were not able to fully grasp the biblical message. If anything, the language of the Bible mainly became appreciated as a liturgical language.

Forms of literary Malay (often with elements from Vehicular Malay) also became the language of colonial administration and education in the nineteenth and early twentieth centuries. In the meantime, Vehicular Malay varieties were used as linguae francae in everyday communication in the cities. The latter were largely populated by Chinese and other migrant populations as well as by Eurasians. Many of them would end up speaking these varieties as a mother tongue.

The nineteenth century saw various tendencies moving in the direction of a unified language, such as the development of a (vehicular) Malay press, the emergence of a colonial administrative jargon, the progressive use of Malay as a medium language in education, and the appearance of literary works. The latter were mostly in vehicular Malay but also include the works of Abdullah bin Abdul Kadir Munshi on the Malay peninsula (nowadays West

Malaysia), who was the first author to write modern prose in literary Malay, and of Raja Ali Haji in Sumatra. At the turn of the twentieth century, the Dutch colonial administration designed a standardized Malay language through its bureau for language planning, called Balai Pustaka (founded in 1908). The new standard was primarily based on the literary Malay of the twin courts of Riau in Sumatra and Johore on the Malay peninsula, on the current border between Indonesia and Malaysia. This border area was (somewhat mistakenly) considered to be the traditional heartland of literary Malay. The standard underwent strong influence from Minangkabau as a result of Balai Pustaka's policy to employ Sumatran teachers for the production of modern literature in it.

As already mentioned, the implementation of this standard (which already in colonial times came to be known as "Indonesian") as a national language in Indonesia is generally considered a success in post-colonial history. Several factors favored its acceptance. One was that Malay was neither the language of a majority (which was Javanese by a long shot), nor of a leading elite. Another was that Malay was not a socially stratified language, in contrast to Javanese, which requires the use of different vocabulary depending on the social status and/or formal distance between speaker and hearer. Furthermore, Malay was widely understood as a second language. Finally, the general acceptance of Indonesian was officially promoted in the last decennia of colonialism and enforced by the Japanese in WWII.

However, the preference for a "pure" and authentic form of literary Malay from Sumatra over the widely spoken vehicular Malay varieties had the undesirable long-term effect of creating a sociolectal diglossia, since the newly designed standard was rather different from both the vernaculars spoken in traditional Malay communities, and the Vehicular Malay varieties spoken in places where Malay was originally imported. Its acceptance has by no means led to a replacement of vehicular Malay, nor has it prevented the rise of Jakarta Indonesian, which has now become the informal counterpart of standard Indonesian. Notwithstanding the success of its implementation, the direction of its further development has been the subject of much controversy since Indonesia's independence. On one end of the spectrum there is the Pusat Bahasa, a government agency and basically the continuation of the colonial Balai Pustaka. It tries to guide the adaptation of Indonesian to the needs of modern Indonesian society in a fast-globalizing world. It assumes a conservative approach in grammar codification and vocabulary building by trying to reduce European influence and use elements from Sanskrit, Arabic, Old Javanese, and Classical Malay instead, especially in the creation of terminologies. On the other end there are more modernist forces favoring a *laissez-faire* approach to the progress of Indonesian, with less stringent grammatical rules and no elimination of European vocabulary if it is already integrated in the language or is part of global terminology.

A special mention should be made of the enormous influx of English vocabulary since Indonesia's independence. The Pusat Bahasa and some individuals see this as a threat to the integrity of Indonesian, whereas other people consider it a natural tendency in its current development. Whatever the case may be, the use of English vocabulary is widespread among Indonesians, including politicians, journalists, TV presenters, and the like. It occurs even in cases where Indonesian words are also available. Action taken against this practice by the Suharto government in the 1980s and 1990s had only a limited effect (Sneddon 2003: 173–195). Moreover, critics of the Suharto regime tended to interpret this effort as an unhealthy form of social engineering by an authoritarian government. Some modernists are also more open to the use of non-standard varieties of Malay as non-formal counterparts to standard Indonesian, which is generally felt to be too formal and stilted.

High versus low controversies have also marred standardization efforts in Malaysia and the Malay diaspora. In Malaysia, a standard language was designed around the same time as in Indonesia. British colonial officers by and large used the same literary criteria as their Dutch colleagues, with whom they often cooperated in matters of standardization. Here too, the

resulting standard was rather different from the local vernacular and vehicular varieties in use. It was also worked on at the teacher training college in Tanjung Malim, where the applied linguist Zainul Abidin (or “Za’ba”) played a major role in its further development. However, its acceptance and implementation was delayed and difficult compared to Indonesian in Indonesia. It only started from the late 1960s onward, after Malaysia had experienced political turmoil caused by racial tension, and a compromise was reached between the main factions, which were ethnically aligned. Standard Malay was accepted as an official language and made compulsory as a medium in administration, education, and the army, yet some Malaysians continued to resist using the language. Another problem is the lack of confidence Malaysians (including Malays) have in using Malay in education, science, and industry, to the extent that in 2003 the Malay-led government under Mahathir re-instated English as a medium language for teaching science and mathematics in higher education, in a bid to keep Malaysia internationally competitive. English has also remained the medium in legal courts.

In Sri Lanka, where Malay acquired an entirely different grammar, the local Malay community is divided over the form of Malay to take as a basis for standardization. Some Sri Lankan Malays are in favor of adopting standard Malaysian Malay, citing criteria of linguistic purity and aspiring to a stronger connection with metropolitan Malays. Others support the local Sri Lankan Malay variety, wishing to maintain its uniqueness and taking pride in it because of its heritage value. They also fear the emergence of two very divergent registers and hence an undesirable state of diglossia. A comparable situation existed (or still exists) on the Cocos Keeling Islands, where the local vehicular Malay variety became influenced by more standardized forms of Malay on the advice of outsiders upholding Malaysian or Indonesian as a norm (Adelaar 1996).

In the Netherlands, there is a community of Moluccans who came to the country in the early 1950s in the aftermath of the Indonesian war of Independence against the Dutch. Most of them spoke a predominantly eastern Indonesian form of vehicular Malay, which has continued to be influenced by Dutch ever since. Efforts to standardize this language in the last decades of the twentieth century also evolved around the choice of a suitable sociolect. The spoken language (perceived as “low” Malay) lacked prestige, whereas the liturgical language used in church (“high” Malay) was not sufficiently mastered by the community. Meanwhile, for political and historical reasons, the adoption of Indonesian used to be anathema among community members.

As discussed previously, the implementation of standard Malay in Brunei Darussalam has been relatively unproblematic in comparison to the situation in other countries.

As to the resilience of non-standard Malay, Errington (2014) points to the development of Malay koines in some Indonesian urban settings such as Kupang (capital of West Timor) and Jakarta, and possibly elsewhere. These are local (vernacular or vehicular) Malay varieties or (in the case of Jakarta Indonesian)³ informal forms of Indonesian enriched with local Malay features. Although officially rejected, these koines have become increasingly popular, especially among the young. They are more important than official Indonesian in everyday communication. Their rise in prominence is in tandem with the emergence of regional middle classes.

36.3.2 *Authentic or Creolized?*

The sociohistorical division of Malay by Adelaar and Prentice (1996) and Adelaar (2011) into literary, vernacular, and vehicular varieties was based on the assumption that vehicular varieties derive from a simplified language that came about through contact, combining Malay vocabulary with a grammar partly based on that of another language, possibly a form of Chinese. Many grammatical features of vehicular Malay are also found in Chinese varieties, and contacts

between Malays and Chinese are very old. According to Andaya and Andaya (2001) the Chinese were already well established in Srivijaya when this city lost its significance as a Malay center in the twelfth century AD.

Other authors question the contact origins of vehicular Malay. As early as 1980, Collins argued that Ambonese Malay was not a creole language and that a creole origin could only be established on sociohistorical grounds. He also maintained that typological features were not critical because some of these could also be found in languages with no demonstrable creole history. Collins (1980) was also reacting against Hall's (1966: 13, 18) misguided assumption that Indonesian is based on a creole language (an assumption that unfortunately has found much resonance in the sociolinguistic literature).

McWhorter (2001) believes that creole grammars are simpler than grammars of other languages and considers vehicular Malay varieties as creoles because of their structural simplicity. In McWhorter (2008a,b) he asserts that many perceived non-creole varieties of Malay are in fact also creoles through the combination of two factors, namely language contact and the wholesale use of these varieties as a second language.

Gil (2001) argues that McWhorter's parameters of linguistic simplicity are defined too narrowly. The correlation between creole languages and grammatical simplicity is only unilateral: whereas creole grammars are generally simpler, the grammars of non-creole languages vary between simple and complex. Furthermore, the perceived simplicity of vehicular Malay varieties also applies to some vernacular Malay varieties. While acknowledging the importance of language contact and second language use, Gil (2008) points out that these are not the only reasons for the "simplicity" of Malay, another one being intrinsic developments at the Proto-Malayic level, long before the emergence of Malay as a hegemonic trade language.

Gil's argument that grammatical simplification can also be triggered from within a language is essentially correct. Discussants in the debate about the sociolinguistic origins of Malay varieties tend to use arguments based only on general linguistic typology and sociolinguistics. However, historical linguistic evidence clearly shows that Malayic grammars underwent considerable simplification due to two sound changes, namely the merger of schwa and */a/ in final syllables to *a* and the neutralization of vowels to schwa in antepenultimate syllables (Adelaar 1992). These changes together brought about the merger and loss of many affixes in the history of Malayic varieties. They also caused the erosion of the original Austronesian verbal morphosyntax, reducing its morphological four-voice system to one only opposing agent and patient voice.

Historical linguistic evidence also shows that many features contrasting vehicular Malay varieties to other forms of Malay are typologically at odds with more general Austronesian grammatical patterns, such as the total loss of the original Malayic voice system (which was in itself a reduced version of the original Austronesian voice system), and the loss of Malayic plural pronouns. In short, apart from being simple, the structure of vehicular Malay varieties is also rather atypical from an overall Austronesian (and Malayic) typological and historical perspective. This is another strong argument for a creole analysis of vehicular Malay varieties.

NOTES

1 Thanks to Chandra Jayasuriya (cartographer at Melbourne University) for providing this map.

2 In symmetric voice systems, prevalent in many Austronesian-speaking regions, including Malaysia and western Indonesia, actor and undergoer voice alternations are marked on the verb, so that neither voice is clearly the base form (Himmelmann 2011:112ff).

3 Errington does not follow Grijns' (1991) distinction between Jakarta Malay and Jakarta Indonesian.

REFERENCES

- Abdul Chaer Mad'ie. 1976. *Kamus dialek Melayu Jakarta – Bahasa Indonesia*. Jakarta: Nusa Indah.
- Adelaar, Alexander. 1992. *Proto-Malayic: the reconstruction of its phonology and parts of its morphology and lexicon*. Pacific Linguistics C-119, Canberra: Research School of Pacific Studies, Dept. of Linguistics, A.N.U. (Revised edition of Proto-Malayic, 1985).
- Adelaar, Alexander. 1996. Malay in the Cocos (Keeling) Islands, in Bernd Nothofer (ed.), *Reconstruction, Classification, Description. Festschrift in honor of Isidore Dyen*, 167–198. Hamburg: Abera Verlag (Asia Pacific).
- Adelaar, Alexander and Prentice, D.J. 1996. Malay: its history, role and spread, in S.A. Wurm, P. Mühlhäusler and D. Tryon (eds): *Atlas of languages of intercultural communication in the Pacific, Asia and the Americas*. 673–693. Berlin/New York: Mouton de Gruyter.
- Adelaar, Alexander. 2004. Where does Malay come from? Twenty years of discussions about homeland, migrations and classifications, *Bijdragen tot de Taal-, Land en Volkenkunde*, 160/1: 1–29.
- Adelaar, Alexander. 2005. Salako or Badameà. *Sketch grammar, texts and lexicon of a Kanayatn language (West Kalimantan)*. Wiesbaden: Harrassowitz.
- Adelaar, Alexander. 2011. Structural diversity in the Malayic subgroup, in Alexander Adelaar and Nikolaus P. Himmelmann (eds): *The Austronesian languages of South East Asia and Madagascar*. (2nd edition; 1st edition 2005). London: Routledge. 202–226.
- Andaya, Barbara Watson and Andaya, Leonard Y. 2001. *A history of Malaysia*. (2nd edition). Hampshire: Palgrave.
- Asmah Haji Omar. 1981. *The Iban language of Sarawak: a grammatical description*. Kuala Lumpur: Dewan Bahasa dan Pustaka.
- Blust, Robert A. and Alexander D. Smith. 2014. *Bibliography of the languages of Borneo (and Madagascar)*. Borneo Research Council Reference Series Volume 2. Phillips (Maine): Borneo Research Council.
- Collins, James T. 1980. *Ambonese Malay and creolization theory*. Kuala Lumpur: Dewan Bahasa dan Pustaka.
- Collins, James T. 1990. *Bibliografi dialek Melayu di pulau Borneo*. Siri Monograf Bibliografi Sejarah Bahasa Melayu. Kuala Lumpur: Dewan Bahasa dan Pustaka.
- Collins, James T. 1995a. *Bibliografi dialek Melayu di pulau Sumatera*. Siri Monograf Bibliografi Sejarah Bahasa Melayu, Kuala Lumpur: Dewan Bahasa dan Pustaka.
- Collins, James T. 1995b. *Bibliografi dialek Melayu di pulau Jawa, Bali dan Sri Lanka*. Siri Monograf Bibliografi Sejarah Bahasa Melayu, Kuala Lumpur: Dewan Bahasa dan Pustaka.
- Collins, James T. 1996. *Bibliografi dialek Melayu di Indonesia Timur*. Siri Monograf Bibliografi Sejarah Bahasa Melayu. Kuala Lumpur: Dewan Bahasa dan Pustaka.
- Collins, James T. and Awang Sariyan (eds). 2006. *Borneo and the homeland of the Malays*. Kuala Lumpur: Dewan Bahasa dan Pustaka.
- Djenar, Dwi Noverini. 2008. On the development of a colloquial writing style: examining the language of Indonesian teen literature, *Bijdragen Tot de Taal-Land-en Volkenkunde*, 164(2–3), 238–268.
- Errington, Joseph. 2014. In search of Middle Indonesian: linguistic dynamics in a provincial town, in Gerry van Klinken and Ward Berenschot (eds): *In search of Middle Indonesia. Middle classes in provincial towns*. 199–219. Leiden/Boston: Brill.
- Gil, David. 2001. Creoles, complexity and Riau Indonesian, *Linguistic Typology* 5: 325–371.
- Gil, David. 2008. Why Malay/Indonesian undressed: contact, geography, and the roll of the dice. Paper: International Seminar of Malay and Indonesian Linguistics, Leiden University, June, 2008.
- Grijns, C. D. 1991. *Jakarta Malay. A multidimensional approach to spatial variation* (two volumes). Verhandelingen 149. Leiden: Koninklijk Instituut voor Taal-, Land- en Volkenkunde.
- Hall, Robert A. 1966. *Pidgin and creole languages*. Ithaca (NY): Cornell University Press.
- Ikranagara, Kae. 1980. *Melayu Betawi Grammar*. Nusa 9, Jakarta: Badan Penyelenggara Seri Nusa.
- Kamus Bahasa Melayu Brunei*. 1991. Bandar Seri Begawan: Dewan Bahasa dan Pustaka.
- Kamus Dewan* (4th edition). 2005. Kuala Lumpur: Dewan Bahasa dan Pustaka.
- Kluge, Angela. 2014. *A grammar of Papuan Malay*. Utrecht: LOT Netherlands Graduate School of Linguistics.
- Lauder, Multamia R.M.T. 1993. *Pemetaan dan distribusi bahasa-bahasa di Tangerang*. Seri

- ILDEP. Jakarta: Pusat Pembinaan dan Pengembangan Bahasa/Leiden: Leiden University, Dept of Languages and Cultures of South East Asia and Oceania.
- Litamahuputty, Betty. 2012. *Ternate Malay. Grammar and texts*. Utrecht: LOT Netherlands Graduate School of Linguistics.
- Martin, Peter W. 1998. A sociolinguistic perspective on Brunei, *International Journal of the Sociology of Language* 130/1:5–22.
- McWhorter, John. 2001. The world's simplest grammars are creole grammars, *Linguistic Typology* 5: 125–166.
- McWhorter, John. 2008a. Why does a language undress? Strange cases in Indonesia, in F. Karlsson, M. Miestamo and K. Sinnemäki (eds): *Language complexity: Typology, contact, change*. 167–190. Amsterdam: John Benjamins.
- McWhorter, John. 2008b. The diachrony of Malay: what “just happens,” Paper: International Seminar of Malay and Indonesian Linguistics, Leiden University, June, 2008.
- Moussay, Gérard. 1981. *La langue minangkabau*. Paris: Association Archipel.
- Moussay, Gérard. 1995. *Dictionnaire minangkabau – indonésien – français* (two volumes). Cahiers d'Archipel 27. Recherches Asiatiques. Paris: l'Harmattan/Association Archipel.
- Muhadjir. 1981. *Morphology of Jakarta dialect: affixation and reduplication*. Nusa 11, Jakarta: Badan Penyelenggara Seri Nusa.
- Nik Safiah Karim, Farid M. Onn, Hashim Haji Musa, and Abdul Hamid Mahmood. 1993. *Tatabahasa Dewan Edisi Baharu*. Kuala Lumpur: Dewan Bahasa dan Pustaka.
- Nordhoff, Sebastian. 2009. A grammar of Upcountry Sri Lanka Malay (three volumes). Utrecht: LOT Netherlands Graduate School of Linguistics.
- Nurlela Adnan, Ermitati and Rosnida M. Nur. 1994. *Kamus Bahasa Indonesia – Bahasa Minangkabau* (two volumes). Jakarta: Pusat Pembinaan dan Pengembangan Bahasa.
- Paauw, Scott. 2008. *The Malay contact varieties of eastern Indonesia: a typological comparison*. PhD thesis, State University of New York (Buffalo).
- Richards, Anthony. 1981. *An Iban–English dictionary*. Oxford (UK): OUP, Clarendon Press.
- Sneddon, James. 2003. *The Indonesian language. Its history and role in modern society*. Sydney: University of New South Wales Press.
- Sneddon, James. 2006. *Colloquial Jakartan Indonesian*. Canberra: Pacific Linguistics.
- Sneddon, James. 2010. *Indonesian. A comprehensive grammar*. (2nd and revised edition by Alexander Adelaar; Dwi Noverini Djenar; Michael Carter Ewing). London: Routledge; St Leonards (New South Wales, Australia): Allen and Unwin.
- Steinhauer, H. 2002. More (on) Kerinci sound-changes, in K.A. Adelaar and R. Blust (eds): *Between worlds. Linguistic papers in memory of David John Prentice*. 149–176. Canberra: Pacific Linguistics.
- Steinmaier, Otto. 1999. *Jalai Jako' Iban. A basic grammar of the Iban language of Sarawak*. Kuching (Sarawak): Klasik Publishing House.
- Stevens, Alan M. and A. Ed. Schmidgall-Tellings. 2010. *A comprehensive Indonesian–English dictionary*. Athens (Ohio): Ohio University Press (1st edition 2004).
- Stoel, Ruben. 2005. *Focus in Manado Malay. Grammar, particles, and intonation*. Research School of Asian, African and Amerindian Studies, Leiden University. Leiden: CNWS Publications.
- Tamsin Medan. 1980. *Dialek-dialek Minangkabau di daerah Minangkabau/Sumatra Barat (Suatu Pemerian dialektologis)*. Jakarta: Pusat Pembinaan dan Pengembangan Bahasa.
- Thurgood, Graham. 1999. *From ancient Cham to modern dialects: two thousand years of language contact and change*. Oceanic Linguistics Special Publication 28. Honolulu: University of Hawai'i Press.
- Tjia, Johnny. 2007. *A grammar of Mualang, an Ibanic language of Western Kalimantan, Indonesia*. Utrecht: LOT Netherlands Graduate School of Linguistics.
- Van Minde, D. 1997. *Melayu Ambong: phonology, morphology, syntax*. Research School of Asian, African and Amerindian Studies, Leiden University. Leiden: CNWS Publications.
- Wurm, Stephen A., and Shirô Hattori (eds.). 1981–1983. *Language atlas of the Pacific area* (two volumes). Canberra: Pacific Linguistics.
- Wurm, Stephen A., and Peter Mühlhäusler; Darrell Tryon (eds). 1996. *Atlas of languages of intercultural communication in the Pacific, Asia and the Americas* (three volumes). Berlin: Mouton de Gruyter.

Index

- AAE, *see* English
AAVE, *see* English
abbreviation, 374–376, 378, 551
absolute distance, 335, 342
Abstandssprache, 205
Académie Française, 479–480
accent, 20, 26, 75, 112, 119, 151, 171, 219–228
 and dialect, 4–5
 pitch accent, *see* pitch
accommodation, 34, 61, 118, 143–149, 151, 370, 380
acoustic phonetics, 9, 58, 62, 86, 109, 195, 198, 221, 225, 226, 236, 314–329, 334, 337, 416, 452
affix, 94, 538, 542
Africa, 150, 227, 439, 440, 445, 446, 478, 481, 499–501, 503–504, 523, 525, 527, 530, 532
African English, *see* English
African American English, *see* English
age, 29, 42, 51, 61, 66, 94, 110–113, 116, 147, 153, 170, 183, 195, 199, 207, 222, 226, 245, 247–249, 271, 275, 278–279, 291, 359, 360, 370, 385–388, 390–391, 394–396
aggregation, 400–414
ain't, 300–301
AIS (*Sprach- und Sachatlas Italiens und der Südschweiz*), 255, 264, 284, 353, 489
Albanian, 318
Alemannic, 162, 164, 165, 168, 170, 172, 173, 305
allophone, 58, 70, 74, 78, 80–81, 100, 151, 162, 288, 317
allophonic differentiation, 317
All-Slavic Linguistic Atlas (Obščeslavjanskij lingvističeskij atlas, OLA), 359, 510
all-word comparison, 335
ALPHAMALIG, 338
Alsace, 164, 166
Alsatian, 160, 164
ambisyllabicity, 95
American Dialect Society, 30
American English, *see* English
American Speech, 30
ANAE, *see* *Atlas of North American English*
analogy, 28, 88
Andorra, 409
Apulia, *see* Italy
Arabic, 5, 147–148, 207, 307, 322
 Classical, 525
 dialects, 523–532
 Egyptian Arabic, 529–530
 Maghrebi Arabic, 524, 525, 527, 530, 532
 Mesopotamian Arabic, 529
 Modern Standard Arabic, 525
 Syro-Lebanese Arabic, 530–531
archaeology, 18, 487, 539
areal linguistics, 29, 57, 69, 126, 129, 305–306, 308–309, 348–350, 354–357, 359–361
areal structure, 348, 350, 354, 355, 357, 359, 361
'armchair' method, 91, 92
Armenian, 350, 352
arrow map (map, directed line), 354
articulation rate, 321
Ascoli, Graziadio Isaia, 489, 494
aspiration, 74, [numerous references from p.319 onwards, inc. to pre-aspiration]
association, 114, 195, 199, 221–222, 226, 336–337, 339, 372–373, 375–376, 418
asylum process, 227–228
Athens (Georgia), 140
Athens (Greece), 23, 25
Atlanta, 456
Atlante Linguistico Italiano (ALI), 490
atlantometry, 19, 123
atlas, *see* dialect atlas
Atlas lingüístico de la Península Ibérica (ALPI), 500

- Atlas Lingüístico Diatópico y Diastrático del Uruguay (ADDU)*, 359
- Atlas linguistique de la France (ALF)*, 6, 29, 62, 63, 69, 71, 124–134, 137, 138, 140, 254, 284, 288, 353–354, 356, 464, 465, 474, 475, 479, 481
- Atlas linguistique de la Gascogne (ALG)*, 127–128, 482
- Atlas Linguistiques et Ethnographiques de la France par régions (ALFR)*, 479
- Atlas of English Dialects (AED)*, 348–349
- Atlas of North American English (ANAE)*, 10, 19, 61, 73, 85, 86, 107, 288, 296, 314, 318, 350, 360, 452–458
- attitudes, 3, 20, 24–26, 32, 39–41, 50, 107, 118, 119, 171, 177, 188–195, 197–200, 204, 207, 209, 212–213, 258–259, 262, 265, 310
- audience design, 144, 380
- auditory coding, 314
- Ausbausprache*, 205
- Australian English, *see* English
- Austria, *see* German
- Austronesian, 152, 363, 364
- authenticity, 20, 127, 143, 152, 153, 160, 173, 220, 225, 226, 307, 331
- autocorrelation
- pitch tracking, 322
 - spatial
 - global, 418–423, 431
 - local, 421–426, 430–431
- automatic speaker recognition (ASR), 228–229
- Avis, Walter, 31, 452
- Bailey, Nathaniel, 30
- barriers, 1, 6, 237, 370, 537
- Basque Country, 173, 181, 182, 342
- Bavarian, *see* German
- Bayesian inference, 377
- BBC (British Broadcasting Corporation), 108, 223, 224, 249, 262
- BBC Voices survey, 108, 262, 441
- beam map, 129, 133, 354
- Beijing, 549
- Belarusian dialects, 515–516
- Belfast, 5, 113–114, 258, 306–307
- Belgian Dutch, *see* Dutch
- Belgian French, *see* French
- Belgium, 53, 125, 151, 160, 207
- Benchmark Database of Phonetic Alignments in Historical Linguistics and Dialectology (BDPA)*, 339
- Berlin, 11, 30, 163, 165, 323
- Berliner Lautarchiv*, 30
- big data, 18, 67, 70, 372
- bigram, 308, 336
- binary variable, 384, 389, 409–410, 418–420
- bipartite spectral graph clustering (BSGC), *see* clustering, bipartite spectral graph
- Birmingham (Alabama), 195
- Birmingham (UK), 228
- Black English, *see* English, African American
- blocking variable, 398
- blog, 368
- Bloomfield, Leonard, 69, 74, 400
- Bonferroni correction, 389, 398
- bootstrap clustering, *see* clustering, bootstrap
- borders, 10, 60, 61, 120, 137, 139, 151, 161, 163–164, 166, 171, 179, 188, 207, 213, 259, 318, 320, 361, 409
- Bosnian/Croatian/Serbian dialects, 205, 518–520
- Boston, 452, 455
- boundary
- dialect, *see* isogloss
 - personal, *see* fieldworker, boundary
 - physical, 236
 - political, 5, 42, 63
- Brabantian/Brabantic, *see* Dutch, Brabantian
- Brazilian Portuguese, *see* Portuguese
- Britain, *see* Great Britain
- British English, *see* English
- British Isles, 31, 34, 186, 223, 244, 262, 324, 332
- British National Corpus (BNC)*, 304
- Brunei Malay, *see* Malay
- Bulgarian dialects, 337–339, 520
- Bullockar, William, 24, 25
- bundling (of isoglosses), *see* isogloss
- California, 186, 207, 457–458
- Canada, 31, 61, 62, 146–147, 186, 248–250, 255, 257, 270, 274–277, 279–280, 282, 318
- Canadian English, *see* English
- Canadian French, *see* French
- Canadianism, 31, 278
- Canadian Raising, 98, 151, 276, 458
- Canadian (Vowel) Shift, 458
- Cantonese, *see* Chinese
- Cardiff, 144–145
- Carpathian Dialect Atlas (OKDA)*, 350, 351
- cascade model, *see* diffusion
- Cassidy, Frederick, 8, 30, 42, 451
- Castellano, *see* Spanish
- Catalan, 131, 207, 357, 409
- Majorcan, 320
 - Valencian, 320
- Catalonia, 126, 140, 173, 181, 182, 409
- categorical design, 408
- categorical reasoning, 92
- Celtic, 4, 11, 24
- chain shifts, 19, 61, 85–86, 96, 110, 114, 197, 200, 317–318, 378, 454–458
- Chambers, J. K., 62, 106, 111, 112, 116, 117, 146–147, 249, 268–282
- Charmey, 30, 61, 111, 481
- Chaucer, Geoffrey, 24

- chi-squared (χ^2) test of independence, 110, 237
 Chicago, 195, 374, 376, 378, 454, 457
Chicano English, *see* English
 Chinese
 dialects and languages, 5, 152–153, 207–209,
 547–556
 Cantonese, 5, 549
 Mandarin, 549
Putonghua (Standard Modern Chinese),
 547–549
 Yue dialects, 549
 Chomsky, Noam, 19, 69, 75, 85, 89, 92
 chorochromatic map, 355, 356
 choropleth map, 132–135, 137, 356, 363
 Cincinnati, 457
 class, *see* social class
 Classical Arabic, *see* Arabic
 closed question, 253, 277, 280
 close-knit network, *see* network
 close tie, 111, 117
 cluster analysis, 115, 183–185, 342
 clustering
 bipartite spectral graph (BSGC), 406
 bootstrap, 402
 complete link *see* clustering, furthest neighbor
 furthest neighbor, 402
 hierarchical, 183, 209, 401–402
 k-means, 183, 401–403
 nearest neighbor, 402
 noisy, 342, 402–404
 single-link clustering, *see* clustering, nearest
 neighbor
 spatial, 415, 419–425, 427, 431
 stochastic, *see* clustering, noisy
 code-switching, 34, 370
 co-difference (COD), 131, 133, 137
 codification, 23, 34, 39, 91, 160–162, 205
 cognates, 29, 43, 79, 82, 207–211, 214, 215, 336
 cognitivist models, 99–101
 Cohen's d, 237
 collocation, 42, 46, 52, 262, 309
 colloquialism, 53, 222, 263
 collostruction, 406
 Colombia, 33
 colonisation, 4, 33, 34, 148, 154, 159, 450, 475, 498
 [and a number of other mentions later in
 the text]
 commensurability, *see* comparability
 commodification, 173
 community of practice, 114, 115, 117
 comparability, 50, 51, 97, 112, 125, 127, 146, 234,
 248, 253, 262, 265, 286
 comparative method of reconstruction, 7
 comparative philology, 27, 28
 Competing Grammars model, 98, 102
 complete link clustering, *see* clustering,
 complete link
 complex (adaptive) system, 58–60, 62, 67, 70
 comprehensibility, 40, 177, 179, 196, 200, 212, 238
 comprehensibility, *see* intelligibility
 mutual comprehensibility, *see* intelligibility,
 mutual
 computational linguistics, 236
 computers, 10, 33, 57, 63, 67, 69, 70, 92, 107, 115,
 128, 129, 131, 186, 229, 236, 250, 270, 281,
 308, 330–347, 370, 376–378, 470, 568
 confound, statistical, 234, 387
 connotation, 23, 52, 151, 212, 262
 consciousness, 117, 119, 127, 144, 195, 199, 200,
 221, 235, 271, 277, 281
 consistency, 62, 68, 254, 308, 340–341, 431
 consonant
 voiced, 319, 320
 voiceless, 151, 319, 320
 constraint, 18, 81, 89–90, 93, 98–102, 108, 148, 150,
 305–307, 309, 398
 ranking, 90, 101, 305, 306
 contact effects, 9–11, 20, 34, 50, 89, 107, 114, 117,
 138, 143–155, 162, 163, 204, 207, 213, 226,
 324, 331, 333, 339, 505
 contingency table, 385–387
 contour plot, 410
 contrast variable, 384, 388, 394
 Helmert contrast, 385
 sum contrast, 385, 391, 394
 treatment contrast, 385, 392
 convergence (of dialects), *see* dialect leveling
 conversation, 23, 49, 50, 71, 94, 107, 127, 250, 256,
 262, 285–289, 294–295, 305
 conversation analysis, 194–195, 301
 Copenhagen, 11, 107, 112, 115, 119, 120, 169, 170,
 199, 208, 213, 463, 469
 cophenetic distance, 402–403
 core area, 350, 356
Corpus Gesproken Nederlands (CGN), 305
Corpus of Contemporary American English (COCA),
 300, 301, 308
Corpus of Global Web-Based English (GloWbE), 301
 corpus linguistics, 235, 300–315
 correctness, 27, 60, 191–193, 195, 199, 222–225,
 279–280
 political, 40
 correlation, 282, 286, 318, 331–332, 335, 337,
 340–342, 403–405, 421
 correspondence analysis, 98, 406
 coverage, 235, 241, 245–246, 248–249, 269–271,
 274–275, 281, 284, 302, 304, 332, 356,
 435, 436
 creole, 35, 357, 532, 578–579
 criterion variable, *see* dependent variable

- Croatian dialects, *see* Bosnian/Croatian/Serbian dialects
- Cronbach's alpha (α), 340
- cross-validation, 409
- crowdsourcing, 19, 53, 69, 234
- Cumberland, 30, 245–246, 444
- Cumbria, 187
- cutoff
- maximum distance cutoff, 424
- Czech, 161, 510, 517
- Dallas, 456
- Danish, 11, 107, 109–114, 116–120, 159, 169, 171, 199, 205, 207–210, 212–214, 270, 321, 463, 466, 469
- DARE, *see* Dictionary of American Regional English data
- categorical data, 237, 331, 384–399, 404, 406, 408, 416, 419
 - collection, 6, 10, 31–32, 40, 47–49, 62, 73, 107, 124, 127, 229, 233–236, 238, 394, 400, 464
 - matrix, 129–131, 133, 333, 340–341, 398, 418
 - mining, 305, 307
 - numerical data, continuous data, 19, 33, 97, 123, 128, 236, 333, 356, 404, 419, 420
- data-driven (statistical) analysis, 372
- decline (of dialects), *see* dialect leveling
- dedialectalization, 111, 115
- Defoe, Daniel, 26
- degree of difference method, 84, 177, 181–185, 192
- degree of freedom, residual degrees of freedom, 388–389, 391–392, 396, 398
- dendrogram, 349, 402–403
- Denmark, *see* Danish
- dependence
- non-stationary spatial, 421
 - spatial, 417, 419–421
 - stationary spatial, 421
- dependent variable, 109, 111, 407, 408
- Derby, 319, 445
- Derbyshire, 246
- destandardization, 164, 171, 172
- Detroit, 112, 114, 378, 457
- Deutscher Sprachatlas (DSA), 6, 30, 60, 139, 165–167, 169, 170, 254, 269, 284, 356, 359, 464
- Deutscher Wortatlas, 269, 464
- deviation
- standard deviation, 236, 237, 321, 356, 402, 421, 422
- Diachronic Electronic Corpus of Tyneside English (DECTE), 304
- diachrony, 2, 17, 75, 91, 92, 96, 97, 100, 126, 130, 301, 305–306, 359, 379–380
- diaglossia, 101, 164, 165, 173, 469
- dialect
- atlas, 6, 8, 19, 29, 30, 57–71, 73, 76, 80, 109, 123, 139, 168, 233–235, 247, 249, 257, 269, 307, 310, 350, 359, 360, 451, 463–466, 479, 482, 489–490, 500, 554, 561, 568, *see also All-Slavic Linguistic Atlas (OLA), Atlante Linguistico Italiano (ALI), Atlas lingüístico de la Península Ibérica (ALPI), Atlas Lingüístico Diatópico y Diastrático del Uruguay (ADDU), Atlas linguistique de la France (ALF), Atlas linguistique de la Gascogne (ALG), Atlas Linguistiques et Ethnographiques de la France par regions (ALFR), Atlas of English Dialects (AED), Atlas of North American English, Carpathian Dialect Atlas (OKDA), Deutscher Sprachatlas, Deutscher Wortatlas, Digitaler Wenker Atlas, Grammar Atlas of Japanese Dialects, Language Atlas of China, Historischer Südwestdeutscher Sprachatlas, Linguistic Atlas and Survey of Irish Dialects, Linguistic Atlas of Chinese Dialects, Linguistic Atlas of Japan (LAJ), Linguistic Atlas of New England (LANE), Linguistic Atlas of the Gulf States (LAGS), Linguistic Atlas of the Middle and South Atlantic States (LAMSAS), Linguistic Atlas of the United States and Canada, Linguistic Atlas of the Upper Midwest (LAUM), Linguistischer Atlas des dacorümäischen Sprachgebietes, Mittelrheinischer Sprachatlas, Nouvel atlas linguistique de France, Schwäbischer Dialektatlas, Sprachatlas der deutschen Schweiz, Sprach- und Sachatlas Italiens und der Südschweiz (AIS), Taalatlas van Nord- en Zuid-Nederland*
- awareness, 24, 259, 310
- boundaries, 6, 62, 77, 80, 83–84, 117, 183, 199, 250, 400
- coaching, 219–229, 435
- contact, 20, 143–155
- continuum, 5, 163, 164, 206, 282, 437, 463, 466–469, 499, 511, 518, 537–538, 542, 544
- convergence, *see* leveling
- decline, *see* leveling
- dictionary, 6–8, 19, 29–31, 35, 39–54, 160, 244, 451–452, 466, 479–480, 490–491, 561–562, *see also Dictionary of American Regional English (DARE), Dictionary of Canadianisms on Historical Principles (DCHP), Dictionary of Japanese Dialects (DJD), Dictionary of Newfoundland English (DNE), English Dialect Dictionary, Fāngyán (Chinese Dialect Dictionary), Ømålsordbogen (Dictionary of the Insular Dialects), Shōgakkan's Dictionary of Japanese Dialects (SDJD)*

- dialect (*cont'd*)
 distance, 308, 310, 330–331, 333–343, 376–377,
 400–404, 408–410, 470, 493, 552
 divergence, 1, 20, 77, 143, 144, 149, 152, 159–164,
 166, 172–173, 207, 337–338, 377, 400
 ethnic (ethnolects), 2, 9, 11, 27, 34, 42, 111, 115,
 159, 161, 173, 194, 205, 222, 302, 321, 451, 458
 geography, 6, 32, 43, 85, 86, 99, 106, 107, 111,
 184, 268, 285, 297, 309, 324, 357, 360, 401,
 463–466, 479, 489–490, 500, 554, 561, 568
 grammar, 88–92, 307
 ideology, 2–3, 18, 118, 163, 437, 487, 526
 and language, distinction, *see* language and
 dialect, distinction
 lexicography, *see* dialect dictionary
 leveling (and convergence, decline and loss), 5,
 10–11, 20, 33, 94, 97, 110, 144–147, 149–155,
 159–173, 210, 227, 306, 458, 462–463, 488,
 548, 568
 loss, *see* leveling
 maps, 10, 306, 348–364, 404–405, 416–417,
 423–426, 431, 469, 511, 568
 origins of, 1–2, 33
 perception, 20, 60, 62, 118, 177–200, 259, 262,
 310, 335, 340–343, 354
 reduction, 219, 223–225, 229
 rural, 2, 7, 8, 75
 shift, 144, 146, 147, 149, 163, 165
 societies, 7, 30–31
 surveys, 8, 228, 241, 242, 244–250, 254, 268–282,
 325, 362, 428
 traditional, 6–8, 11, 17, 18, 29, 94, 111, 150, 152,
 154, 155, 159, 160, 163, 170, 173, 183, 241,
 242, 248, 301, 303
 urban, 2, 8–9, 11, 35, 42, 162, 163, 324
dialectology
 forensic, 225
 history of, 5–10, 17–18, 28–35, 60–63
 social, 3–4, 8–9, 35, 39, 106–122, 221–225, 233,
 258, 301, 348, 411, 505, 543
 structural, 9, 19, 73–86, 454
 urban, 9, 42, 106, 111, 301, 314
dialectometry, 10, 19, 20, 33, 53, 123–140, 206, 208,
 237, 307–308, 330–347, 356, 376, 400–414,
 417, 431
Dialect Topography (of Canadian English),
 249–250, 270, 275–278, 280, 282, 452
diaphone, 80–82
diasystem, 33, 77, 81–83
DiaTech, 342
dictionary, *see* dialect dictionary
Dictionary of American Regional English (DARE), 8,
 30–31, 42, 248, 316, 317, 451–453
Dictionary of Canadianisms on Historical Principles
 (DCHP), 31, 452
Dictionary of Japanese Dialects (DJD), 562
Dictionary of Newfoundland English (DNE), 452
Dieth, Eugen, 8, 29, 242, 246, 254–256, 289, 440
diffusion, 1, 6, 10–11, 46, 101, 116–119, 125, 126,
 137, 152–154, 161–163, 233, 237, 360, 368,
 371, 378–379
cascade model, 32
contagious, 163
counterhierarchical, 32
fan, 360
hierarchical, 163
parachuting, 163
ring, 360
Digitaler Wenker Atlas (DiWA), 60, 102, 265,
 356–360, 363
diglossia, 2, 100, 164, 165, 355, 462, 469, 487, 488,
 510, 525, 541, 544, 576–577
dimension reduction, 403–406
diphthong, 70, 80, 82, 149, 150, 162, 163, 166,
 168–169, 195, 223, 315, 332
diphthongisation, 163
direct method, *see* data collection
discourse, 23, 50, 57, 109, 119, 144, 194, 195, 200,
 205, 236, 250, 289, 290, 301, 302, 305, 308,
 309, 361, 374
disguise, 20, 219, 225–227, 319
dispersion, spatial, 419, 420, 425
distance matrix, 98, 129, 131, 139, 333, 340–341,
 402–406, 418–430
ditransitive, 304
divergence, *see* dialect divergence
dot-density map, 350
draw-a-map method, 185–188
duration
 phonetic, 319–321, 323, 324, 404, *see also*
 vowel length
 of survey, 274, 275
Dutch, 7, 33, 39–41, 43, 46, 49, 53, 93–96, 102, 159,
 161, 163, 164, 179, 181, 206–208, 210, 211,
 213, 301, 309, 314, 318, 320, 321, 331–332,
 335, 337–338, 356–358, 401, 407–410
Belgian, 39, 40, 53, 301
Brabantic, 42, 43, 53, 160, 177, 178, 465, 468
dialects, 7, 33, 39–41, 43, 49, 53, 93–96, 159, 161,
 164, 179, 206–208, 210, 301, 309, 318,
 320–321, 331–332, 335, 337–338, 357–358,
 400–411, 436, 462–463, 465–468
Flemish, 40, 42, 43, 53, 54, 151, 210, 234, 318,
 405, 468, *see also* Belgian Dutch
Netherlandic, 40, 301, 314
Dynamic Model (DM), 33, 34
East Anglian Fens, 151, 306
Eckert, Penelope, 114, 115, 117, 119
edit distance, 140, 208, 211, 214, 236, 330, 333–343, 406
Edmont, Edmond, 29, 62, 125, 244, 256, 269–270,
 276, 288

- education, 3–4, 17, 20, 23, 27, 40, 42, 48, 51, 62, 112, 152, 161, 172, 183–184, 193, 194, 205, 219, 221–229, 234, 241–242, 245, 247, 277, 280, 291, 297, 302, 310, 359, 370, 411
- effect size, 237
- Egypt, *see* Arabic
- eigenvalue, 405
- E-language, 89
- electropalatography, 320
- elicitation, 18, 32, 48–50, 52, 61, 63, 65, 71, 73, 86, 94, 127, 199, 234–235, 238, 243–244, 246–248, 253–263, 269–282, 285–286, 288–290, 292, 294–297
- Ellis, Alexander, 7, 29, 244, 349, 440
- elocution, 20, 220, 223
- endangerment, 41, 554, 564, 568, 569, 575
- England, 4, 7, 8, 11, 23–26, 28–30, 32, 34, 52, 75, 140, 146, 148, 149, 160, 163, 179, 185, 187, 188, 222, 223, 225, 244, 255, 263, 287, 301, 303, 305–306, 308, 319–320, 322, 324, 348–349, 355, 439–445, 452, 455–457
- English, 3–5, 7, 11, 24–30, 32–35, 39, 41, 49, 60, 63, 74, 75, 79, 82, 107, 110, 117, 123, 129, 146, 148, 173, 210, 220–227
- as a world language, 11, 30, 577–578
- African English, 33, 40
- African American English, 27, 194, 371–372, 389, 451
- American English, 4, 5, 8, 19, 27, 30–31, 42, 57, 61, 69, 73, 80–81, 85, 86, 96, 107, 146, 181–200, 279, 289, 291, 296, 300–301, 308, 314, 316, 320–321, 331, 348, 350, 359–361, 364, 371, 372, 404, 407, 415–433, 450–458
- Australian English, 82, 153, 439, 442, 445–446
- British English, 4, 11, 24–27, 63, 80, 147, 148, 153–154, 187, 222–228, 261–262, 277, 301–302, 307–308, 319–321, 332, 439–449
- Canadian English, 11, 31, 98, 146–147, 151, 249, 276–281, 450–453, 458
- Caribbean English, 47
- Chicano English, 27
- dialects, *see* individual regions under this entry
- General American English, 5, 458
- Irish English, 11, 24–26, 32, 39, 148, 225, 334, 441
- Latino English (United States), 451
- modern, 24–25, 27, 37, 110, 153–154, 222
- Native American English, 451
- Newfoundland English, 31, 452
- New Zealand English, 148, 149, 159, 439–440, 442, 445–446
- North American English, 3, 19, 30, 61, 73, 85, 86, 96, 107, 248, 277–278, 288, 450–458
- North-South divide (England), 130, 469, 501
- Northern British English, 444–445
- Old English, 23, 28
- Pakistani English, 301
- Scottish English, 11, 24, 149, 187, 221, 223, 226, 259, 262, 270, 320, 324, 434–444
- South African English, 40, 439–440, 442, 445–446
- Southern Hemisphere English, 148, 436, 439–440, 442, 445–446
- Standard American English, 5, 458
- Standard British English, 4
- Standard English, 4–5, 26, 28, 75, 79, 222–224, 305–306, 308–310, 439, 575
- Welsh English, 11, 82, 441
- World Englishes, *see* English as a world language
- English Dialect Dictionary* (EDD), 7, 29, 35, 244, 255, 440
- English Dialect Society, 7, 30
- enregisterment, 160, 161, 173, 310
- entropy, 341, 343
- enumerator effect, 274–276
- Estonian Dialect Corpus*, 305
- ethnicity, *see* dialects, ethnic
- ethnolect, *see* dialects, ethnic
- Ethnologue, 53, 206, 355, 362, 540
- etymology, 23, 25, 28, 30, 40, 46, 53, 54, 79, 96, 126, 210, 262
- Euclidean distance, 333, 403, 427–428
- Europe, 5–7, 11, 12, 20, 24, 29, 33, 39, 40, 60, 75, 101, 125, 159–174, 205, 207, 226, 287–288, 362
- Exemplar Theory, 99, 100
- existential non-agreement, 280
- expectation maximization, 378
- expert consensus, 340–341
- extra-linguistic factors, 50, 51, 83, 100, 101, 206–215, 305–306
- extrapolation, 76, 309, 356–357, 362
- Facebook, 249, 368–369, 380
- face-to-face interaction, 235, 253
- factor analysis (FA)
- factor loading, 405
 - factor rotation, 406
 - factor score, 405
- familiarity, 210, 213, 220, 223, 227, 292, 293
- family tree, 7, 28, 537–539
- Fāngyán* (Chinese Dialect Dictionary), 547
- feature frequency method (FFM), 331–333
- feature string comparison, 334
- Fens, *see* East Anglian Fens
- F1~F2 space, 109, 110, 316
- field interview, 68, 235, 254, 272–274, 284–297
- field record, 234, 275, 317, 322
- fieldwork, 6, 29, 39, 40, 42, 47–50, 52–54, 62–63, 68, 69, 71, 76, 80, 81, 97, 109, 125, 174, 194, 228, 244–248, 253–257, 262, 268–276, 279, 282, 284–295, 297
- bias, 272, 276
- boundary, 288
- effect, 297

- Finland, 282, 318, 362, 463
 Finnish, 161, 333
 First Germanic Consonant Shift (Grimm's Law), 7, 28
 First Nations English (Canada), 451
 Fisher's Linear Discriminant (LDA), 339
 fixed effect, 407–408, 412
 Flemish, *see* Dutch
 Florida, 186, 227, 247, 271, 456
 focal area, 348
 focussing, 33, 34, 146, 149, 151
 folk linguistics, 20, 177
 forced alignment, 100–101, 324, 335, 342
 Forced Alignment and Vowel Extraction (FAVE), 100–101, 324, 342
 foreign (a) nativization (in English), 458
 forensic phonetics, 219, 225–229
 formality, 32, 39, 40, 62, 114, 115, 118, 151, 161, 171–173, 193, 234, 241, 248, 253, 255, 256, 262, 288, 290, 368, 380, 408
 formant, 86, 109, 197, 198, 236, 315–319, 337, 404, 406, 412
 trajectory, 315–316, 318
 Fowlkes and Mallows Index, 341
 fractal, 19, 59, 67
 France, 6, 24, 28, 29, 53, 61–63, 65, 69, 71, 123–128, 138, 140, 159, 166, 173, 186, 188, 207, 244, 269, 354, 356, 474–482
 Franconian, *see* German
 free alignment, 335
 Freiburg English Dialect Corpus (FRED), 49, 302–304, 307–309, 332
 French, 4, 6, 24, 29, 30, 33, 39, 41, 58, 60–63, 100, 124, 125, 127, 129–131, 139, 148, 164, 186, 207, 210, 256, 270, 279, 305, 320, 356, 436, 474–482
 Belgian French, 41, 481
 Canadian French, 477–478, 480, 481
 dialects, 4, 6, 474–482
 Langue d'Oc (Languedoc, southern French), 24, 137, 138, 139, 476, 479
 Langue d'Oil (Languedoc oil, northern French), 24, 137, 138, 476, 479
 Quebec French, *see* Canadian French
 Swiss French, 481
 frequency
 count, *see* token frequency
 distribution, 58–59, 70, 112, 130, 137–138, 235
 formant frequency, *see* formant
 fundamental frequency, 144, 226, 321
 profile, 58–59, 70, 333
 relative, 373
 token frequency, 44, 60, 67, 89, 100, 110, 116, 168–169, 172, 233, 258, 301, 304, 307–309, 330–333, 342, 373, 377–380, 385–387, 406, 408, 410, 416, 444
 Friesland, 463
 Frisian, 40, 210, 405, 463
 fudged lects, 355
 functional testing, 206, 208
 fundamental frequency, *see* frequency
 furthest neighbor clustering, *see* clustering
 Gabmap, 342, 364, 405
 Galicia, 173, 181, 182
 Gauchat, Louis, 30, 61, 111
 gender, 29, 39, 42, 51, 61, 110, 111, 116, 127, 222, 225, 226, 234, 243, 247, 249, 262, 302, 303, 370, 385–387, 390–391, 394–397, 407, 408, 411, 442, 469
 grammatical, 117, 150, 343, 527–531
 General American English, *see* English, American
 generalized additive modeling, 408–411
 generalized linear model, 388
 generative grammar, 8, 19, 33, 69–70, 75, 85, 89–92, 95, 97–98, 100–102, 306
 genetically related languages, 7, 18, 28, 75, 209, 336, 337, 436
 geographical distance, 207, 306–308, 310, 407, 409
 Geographic Information Systems (GIS), 57, 186
 geolinguistics, 19, 123–128, 139, 305–308, 310, 364
 feature area, 125–129
 geolocation, 370, 372, 377, 379
 Georgia (USA), 140, 181, 183, 247
 geo-statistics, 237, 408, 415–433
 German, 6, 7, 29–30, 40, 46, 53, 60–61, 69, 90, 117, 129, 140, 161–172, 188, 206–208, 210, 211, 238, 320, 361, 400, 462–471
 Austrian German, 39, 53, 96, 161, 462
 Bavarian German, 11, 160–163, 324, 468
 Berlin German, 323
 dialects, 5, 6, 11, 29–30, 53, 60–61, 69, 79, 90, 161, 163–164, 188, 206–208, 211, 244, 254, 268, 305, 354, 360–361, 407, 462–471
 Dresden German, 323
 Düsseldorf German, 60, 71, 269, 324
 Franconian, 161, 168, 172, 405, 462, 468, 471
 High German, 162, 164, 165, 269, 359
 High German Consonant Shift, 359, 466–467
 Low German, 39, 47, 159, 164, 165, 269, 462, 463, 467, 468
 Lower Saxon, 405, 407
 Middle German, 163–164, 359, 467–468
 Middle High German, 359
 Old High German, 360
 Standard German, 60, 161, 162, 164, 168, 171–172, 254, 269
 Swabian German, 11, 161, 162, 167–169, 172, 468
 Swiss German, 82, 84, 86, 140, 147, 159, 161, 164, 464
 Upper German, 165, 172, 467–468

- Upper Saxon, 162, 163, 468
 urban German, 324
 Viennese German, 324
- Germanic languages, 6, 7, 11, 20, 28, 110, 126, 160, 161, 173, 359, 436, 437, 462–471
- Germany, 5, 6, 11, 29–30, 39, 58, 61, 160–172, 186, 188, 245, 254, 268, 305, 341, 354, 356–357, 359, 407, 462–471
- Getis-Ord G, 421–427
- Gilliéron, Jules, 6, 8, 29, 61–63, 65, 67, 101, 124–127, 254–256, 265, 269–271, 273–274, 276, 282, 284–285, 287–288, 291, 297, 479
- Global Positioning System (GPS), 371
- glottalization, 163, 319
- gold standard, 341
- gradience, 92, 99, 207
- grammar, 18, 27, 29, 39, 46, 58, 62, 70, 74, 75, 79, 82, 85, 88, 89, 95, 98–102, 109, 117, 126, 160, 184, 205, 207, 212, 219–220, 261–262, 276, 279–280, 282, 302, 307–308, 310, 359, 440
- Grammar Atlas of Japanese Dialects* (GAJ), 561
- gravity hypothesis, 238, 378, 400, 407
- Great Britain, 4, 8, 11, 26, 69, 150, 220, 261, 285, 303, 305–306, 317, 439–449
- Great Lakes region (US), 270, 317–319, 454, 457
- Greek, 28, 33
- Ancient Greek, 23, 26
- grid, 58, 124, 125, 127, 130, 137, 138, 241–242, 275, 284, 291, 297, 428–430
- Grimm, Jacob, 8, 27–28, 400
- Grimm's Law, *see* First Germanic Consonant Shift
- /h/, 29, 223, 568
- /h/-dropping, 223, 306
- hand-pick, 237
- Harvard Dialect Survey*, 270–271
- Hebrew, 28
- Helmert contrast, *see* contrast
- heterogloss *see* isogloss
- hierarchical clustering, *see* clustering
- High German, *see* German
- High German Consonant Shift, *see* German
- Hindi, 5, 74, 148, 150, 151, 536, 538–545
- dialects, 148, 150, 538–540, 542–543
 - Khari Boli (Hindi dialect), 542–543
 - Modern Standard Hindi, 543–544
- Hispanic American, 248, 370, 500
- historical dialectology, 233, 362
- historical linguistics, 1, 7–8, 18, 23–35, 107, 284, 302, 306, 337, 339, 354, 357, 359–360, 363, 490, 547
- historical corpus linguistics, 301
- Historischer Südwestdeutscher Sprachatlas* (HSS), 350
- Hochdeutsch, *see* German
- Hogengaku Koza* ([Japanese] Dialectological Handbooks), 560
- homophony, 27, 81, 86, 199, 444, 563–564
- Hondo dialects, *see* Japanese
- Hong Kong, 549
- Honshū, 180, 559, 563
- horizontal convergence, 159, 161–163, 167, 173
- Houston, 456
- H-variety, 161
- identity, *see* social identity
- ideology, *see* dialect ideology
- idiolect, 76, 80, 82
- I-language*, 89
- imitation, 219–221, 225
- immigration, 34, 159, 173, 227
- impersonation, 20, 219, 227
- implicit association test (IAT), 199
- imposed norm hypothesis, 212–213
- independence
- spatial independence, 417
 - statistical independence, 110, 237, 387, 388, 392, 398, 417
- indexicality, 100, 173, 200, 259
- India, 5, 7, 11, 148, 150, 151, 154, 499, 535–545, 575
- indirect method (data collection), 49–51, 464
- Indo-Aryan dialects, 535–544
- Indo-European, 1, 7, 27, 28, 74, 486, 494, 510, 535
- Indonesia, 571–579
- Indonesian/Malay dialects, *see* Malay / Indonesian
- industrialization, 6, 10, 48, 165, 248
- inflectional morphology, 279
- informant, 8, 29, 31–32, 47–52, 58, 62, 63, 68, 70, 71, 81, 107–108, 110–114, 127, 165–168, 177–180, 185, 186, 188, 191, 193, 195–198, 235, 241–249, 253–256, 258–263, 270, 272, 275, 284–289, 291–297, 303, 306, 309
- folk informant, 284
- inherent value hypothesis, 212, 213
- inherited words, 208, 211–212
- innere Kausalität*, 93
- instability, 402, 404
- intelligibility, 10, 20, 40, 143, 177, 179, 181, 196, 200, 204–216, 319, 335, 471, 544
- asymmetric, 5, 207, 213
- criterion, 206, 215
- mutual, 1, 5, 11, 143, 204, 206–212, 215, 238, 243, 463, 471, 474, 499, 516, 529, 535, 538, 540, 542, 543, 545, 549, 553, 555, 556
- partial, 5, 207, 211, 215
- testing, 181, 204, 209, 214, 215
- threshold, 206, 215
- interaction effects, 389–390, 408
- higher-order interaction, 390, 397
 - predictor interactions, 392, 408
- intercept, 388, 390–391
- interdialect, 131, 151

- internal evidence, 91, 97
- International Dialects of English Archive (IDEA), 32
- International Phonetic Alphabet (IPA), 28, 62, 82, 221, 284, 334
- internet, 10, 32, 53, 67, 69, 99, 108, 204, 206, 220, 228, 234, 249–250, 270–271, 281, 289, 360, 363, 368, 370, 470–471
 internet survey, *see* online survey
- interpoint map, 136, 138–139, 354
- interpolation
 inverse distance interpolation, 428–429
 spatial interpolation, 428–431
- interval algorithm, 133, 137, 139
- interview, 6, 8, 29, 31–32, 49–50, 58, 62–63, 68, 76, 86, 107, 125, 146, 165, 166, 168, 171, 172, 174, 195, 225–228, 234–235, 238, 241–249, 253–256, 258, 263, 265, 270–276, 284–297, 301, 303, 322, 332, 409
 field interview, 68, 235, 254, 272–274, 284–297
 telephone interview, 107, 234, 243, 248, 253, 286–289, 296–297, 318, 360
- intonation, 112, 228, 260, 314, 322–324
 phrase, 322
- intuition, 18, 48, 91, 99, 206, 211, 284–286, 289, 300, 349, 361, 373
- invariant *be*, 389
- Ireland, 11, 24–26, 33, 34, 223, 225, 260, 305, 332, 439–441, 452
- Irish English, *see* English
- Isle of Man, 80, 440
- isogloss, 6, 29, 35, 61, 83–84, 116–117, 126, 128, 129, 133, 138–139, 166–170, 206, 250, 305–306, 349, 353–354, 356, 360–362, 364, 400–410, 417, 422, 428, 430
 bundling, 6, 138–139, 206, 306, 400
 non-overlapping, 83–84, 128, 138, 206, 250, 400, 404, 417, 481–482, 562
- isopleth map, 133, 136, 139
- Italian, 58, 62, 131, 160, 164, 171, 172, 207, 215, 323, 353, 355, 357, 360, 400, 410, 411, 486–496
- Italic, 7, 486, 494, 495
- Italy, 5, 25, 58, 125, 126, 140, 160, 162, 173, 207, 255, 355, 436, 437, 462, 486–496
 Apulia dialect, 488–489
 Calabrian, 160, 486
 dialects of, 25, 125, 160, 162, 255, 323, 353, 355, 357, 486–496
 Turin Italian, 487, 489
 Tuscan Italian, 488, 495
 Veneto Italian, 160, 488–489, 495
- Jakarta, *see* Indonesian
- Japan, 11, 146, 150, 177–181, 186, 320, 436, 437, 559–570
 Japanese, 148, 150, 181, 194, 436, 559–570
 dialects, 148, 150, 179–181, 194, 559–569
 Hondo dialects (Japanese mainland), 562–567
 Kyūshū dialects (far-western Japan), 566–567
 Ryukyu dialects (southwestern Japanese islands), 559, 567–568
 Seibu dialects (western Japan), 565–566
 Tobu dialects (eastern Japan), 562–565
- Javanese, 436, 571, 577
- join count statistic, 419–420
- Junggrammatiker*, *see* Neogrammarians
- Jutland, 111–113, 117, 120, 169–171, 466, 469
- Karelia, 362
- Khari Boli, *see* Hindi
- k-means clustering, *see* clustering, k-means
- koineization, 34, 149, 151, 152, 162
- Korea, 188
 North, 186
 South, 186
- Korean, 436, 555, 559
- kriging, 428–431
 ordinary, 428–431
- Kurath, Hans, 8, 61–64, 69, 70, 75–76, 80–81, 84, 85, 242, 244, 246–247, 250, 272, 288, 297, 359, 451–453
- Kuwait, 147–148, 526
- Kyūshū dialects, *see* Japanese
- Labov, William, 8–10, 19, 30, 31, 49, 61, 69, 76, 85, 88, 97, 106–112, 114–116, 118, 153, 195, 234, 243, 247–248, 258, 271–272, 288–289, 296–297, 314, 317–319, 360, 378–379, 389, 416, 451, 454–455
- Lancashire, 11, 29, 30, 439, 444–445
- LANCHART, 107, 110, 112–113, 119, 120, 169
- language
 change, 1, 7–8, 20, 28, 32, 33, 35, 88, 89, 91–93, 95–98, 100, 101, 107, 108, 110, 112, 116, 118, 128, 143–145, 152, 154, 155, 169, 211, 233, 242, 244, 248, 259, 271, 277, 280–281, 291, 296, 305–306, 319, 337–339, 354, 359–361, 378–380, 386, 391, 397, *see also* sound change
- contact, 34, 50, 117, 138, 204, 207, 213, 324, 331, 333, 339, 357, 445, 505
- community, *see* speech community
- crimes, 225, 227
- and dialect, distinction, 5, 205–206, 437, 486, 535, 541–544, 549, 555–556, 574
- evolution, 18, 126, 138, 162
- family, 1, 7, 20, 27, 28, 322
- national, 4, 26, 143, 205
- planning, 20, 204, 205, 542, 575, 577
- policy, 437, 543
- relatedness, *see* genetically related languages

- language analysis for determination of origin (LADO), 227–228
- Language Atlas of China*, 554
- langue*, 69, 70, 74, 102
- Langue d’Oc, *see* French
- Langue d’Oil, *see* French
- latent
- structure, 236
 - variable, 377
- Latin, 486, 494
- Latin American Spanish, *see* Spanish
- Latino English (United States), *see* English
- layer, 67, 78, 127, 306, 350
- leaf-path ancestor, 333
- Lebanon, *see* Arabic, Syro-Lebanese
- Leeds, 29, 244, 245, 320, 441
- Leipzig, 28, 29, 163, 468
- lemmatization, 236
- level of variable, 385, 389–390
- leveling, *see* dialect leveling
- Levenshtein
- algorithm, 211, 214, 333–338
 - distance, *see* edit-distance
 - three-dimensional L. distance, 337
- lexical
- distances, 208–210, 215, 331
 - frequency, 100, 169, 233, 404, 408, *see also* frequency, token
 - set, 27, 29, 58, 147, 278, 444
 - variable, 273, 330, 361, 368, 371–373, 375, 404, 409, 411, 441, 453, 454, 470
 - variation, 3, 75, 79, 169, 233–235, 245–246, 255, 262–264, 269, 271–273, 275, 277, 285, 296, 310, 331, 342, 357, 400, 410, 440–442, 451, 453, 464, 466, 491
- lexicography, 6–7, 24, 26, 31, 39–54, 126, 269, 465, 466, 480, 491
- lexicon, 18, 31, 39–54, 57, 58, 73, 75–76, 79, 80, 82, 84, 88–90, 94–97, 99, 100, 111, 115, 128, 129, 146–148, 169, 184, 194, 205–212, 219, 378, 380, 465, 470, 479, 548, 553, 554, 574, 576
- lexical phonology, 90, 94–97, 99
- lexical variable, 58, 146
- lexical variation, 73, 194, 205, 219, 470, 479, 525, 553–554
- likelihood
- estimation, 409
 - ratio, 388–389, 391
- Likert scale, 52, 193
- Limburg, 43, 95, 173, 467, 468
- Limburgian, 39, 43, 161, 465
- linear predictive coding (LPC), 315
- linear relation, 407–409
- Linguistic Atlas and Survey of Irish Dialects*, 334
- Linguistic Atlas of Chinese Dialects*, 554
- Linguistic Atlas of Japan (LAJ)*, 561
- Linguistic Atlas of New England (LANE)*, 8, 63, 64, 81, 288, 289, 297, 359
- Linguistic Atlas of the German Empire*, *see* Deutscher Sprachatlas
- Linguistic Atlas of the Gulf States (LAGS)*, 68, 69, 247–248, 285–286, 288–289, 291, 294–295, 297
- Linguistic Atlas of the Middle and South Atlantic States (LAMSAS)*, 58–59, 63, 65–68, 233, 247, 285, 288–289, 297, 335, 356
- Linguistic Atlas of the United States and Canada*, 257, 270, 274, 451
- Linguistic Atlas of the Upper Midwest (LAUM)*, 270–275
- linguistic
- anthropology, 310
 - atlas, *see* dialect atlas
 - distances, 20, 128, 144, 208, 209, 216, 308, 330, 341, 356, 407
 - embedding, 109–111, 116
 - insecurity, 193, 223, 224
 - reconstruction, 7, 27, 28, 165, 174
- Linguistic Survey of India (LSI)*, 539–540
- Linguistic Survey of Scotland (LSS)*, 84, 244–246, 441
- Linguistischer Atlas des dacoromanischen Sprachgebietes*, 254
- link function, 388, 409
- literacy, 50, 204, 214, 270, 276, 277, 281, 486, 487, 543, 544
- programs, 204
- Lithuania, 77, 79
- little arrow method, 178–179, 181, 470
- Liverpool, 222, 441, 444–445
- loanwords, 164, 208, 210–212, 279, 458, 559, 574
- phonology, 458
- local incoherence, 331, 341
- logistic
- regression, *see* regression transformation, 373
- logit, 388–389, 392, 409
- log likelihood, 373
- log-linear, 388, 398
- log odds, 388, 409
- London, 11, 24, 25, 27, 32, 110, 153–154, 161, 220–223, 226, 244, 306, 439
- loose-knit networks, *see* network
- Los Angeles, 457–458
- loss (of dialects), *see* dialect leveling
- loudness, 321
- low-back vowel merger (in English), 4, 360, 454–455
- Lower Saxon, *see* German

- Low German, *see* German
 Luxembourg, 164, 462, 466
 /l/-vocalization, 162, 244, 306, 442, 446, 504
- Macedonian dialects, 520
 machine learning, 237–238
 macrostructure, 39–42, 47, 48
 Madrid, 33, 181, 182, 500, 501
 Maghreb (North Africa), *see* Arabic
 Malay/Indonesian
 Brunei, 575
 dialects, 571–579
 Jakarta Indonesian, 574–575
 Standard, 571, 573, 575, 577, 578
 Vehicular Malay, 573, 576–579
 Mandarin, *see* Chinese
 maps
 area, 354–357
 colors, 554
 layer, 67, 350, 568
 line-related map
 directed line map, 354, 357
 undirected line map (beam map), 354
 point-related map, 350–353, 362
 map task, 193, 260
 Martinet, André, 85, 454, 480
 matched guise method, 118, 188, 192, 199, 213, 228
 matrix
 contiguity, 418
 distance, 98, 129, 131, 139, 333, 340–341, 376, 398, 402–406
 spatial weight matrix (SWM), 418–425
 maximum likelihood estimation, 373
 McDavid, Raven, 63, 69, 70, 76, 80–82, 84, 241, 242, 244, 246–247, 276, 285, 288, 297
 McFarlane Guidelines, 228
 media, (storage media), 363
 communication, 3, 11, 19, 108, 119, 155, 161, 171, 194, 205, 227, 235, 237, 249, 268, 368–380
 social, *see* communication
 medieval period, *see* Middle Ages
 merger, 4, 27, 85–86, 110, 147, 149, 152, 153, 197, 261, 263, 291, 296, 317–318, 350, 360, 445, 446, 452, 454–458, 476, 502, 512–516, 537, 544, 552, 567, 568, 575, 579
 Mesopotamia, *see* Arabic
 metadata, 63, 64, 68, 97, 369, 371
 metalexicography, 39–42, 47, 50
 metropolitan statistical area, 373, 375–377
 Michigan, 181–182, 184–186, 191–193, 195–198, 200, 270, 318, 457
 microblog, 368–369
 Micronesia, 148, 150
 microstructure, 39–42, 44, 46–47
 Middle Ages, 24, 126, 463, 474, 475, 574
 Middle East, 159, 227, 352, 523–532
 Midland (US), 60, 184, 185, 188, 318, 320, 452, 453, 455, 457, 458
 Midlands (UK), 60, 304, 308, 440
 Milroy, James, 113–114, 116–118
 Milroy, Lesley, 113–114, 116, 118
 Mind Research Repository, 411
 minimal pair, 80, 86, 260, 263, 291
 Minneapolis-St. Paul, 458
 minority languages, 6, 30, 39, 53, 205, 458, 462, 489, 541, 554
Mittelrheinischer Sprachatlas (MRhSA), 359, 360, 364, 464, 470
 mixed-effects model, *see* regression, mixed-effects
 mixed lects, *see* fudged lects
 mobility, 10, 18, 47, 111, 117, 143, 145, 152, 153, 155, 166, 194, 224, 241–242, 309, 359, 440, 464
 modal (verb), 235, 261, 285, 297, 305, 308
 multiple modals, 285
 model
 fit, 116, 387–388, 392–395
 statistical, 19, 237, 368, 377, 387–397, *see also*
 generalized additive model, regression,
 mixed-effects modeling, and geo-statistics
 stochastic, 90, 98–100, 377, 402–404
 Modern Import Words in the Nordic Countries (MIN), 107
 modularity, 89, 92
 monophthongization, 184, 195–196, 198–199, 223, 454, 456–458, 468
 Montreal, 458, 481, 482
 Moran's I, 419–422, 425, 426
 local, 421, 425, 426
 morphology, 4, 17, 47, 48, 52, 74, 79, 82, 94, 98, 125, 128, 129, 144, 154, 168, 206, 212, 234–236, 249, 255, 257, 261, 269, 277, 279, 300, 302, 310, 343, 360–361, 441, 454, 464–466, 469–471, 493, 526–530, 573, 575
 morphosyntax, 47, 99, 212, 261–262, 286, 291, 305, 308, 332, 376–377, 481, 499, 502–505, 574, 579
 mouthing, 214–215
 multi-collinearity, 388–389, 394, 398
 multi-dimensional scaling (MDS), 349, 356–357, 364, 403–405, 411
 multiple causes, principle of, 237
 multiple sequence alignment, 338
 multiway table, 388
 mute maps (*cartes muettes*), 125–127, 353
 mutual intelligibility, *see* intelligibility, dialect
 MySpace, 368
 naive discriminant learning, 339–340
 natiolect, 39–40, 53
 Native American English, *see* English
 naturalistic data, 127, 300, 331–332

- nearest neighbor clustering, *see* clustering
 negation, 4, 150, 172, 261, 304–309, 377, 408, 481,
 493, 495, 504, 505
 multiple negation, 4, 305–309
 Neogrammarians, 7–9, 28, 32, 57, 60–63, 69, 70,
 88, 96–97, 284, 297, 305, 400, 465
 nesting (of models), 389
 Netherlands, 5, 28, 39, 40, 53, 164, 173, 177–179,
 207–208, 301, 318, 321, 350, 405, 408–410,
 462, 463, 465, 468, 571, 578
 network
 close-knit, 111, 114, 117
 loose-knit, 117–118
 scale-free, 19, 58–60, 67
 social, 112–118, 146–148, 243, 369
 neutralization, 528, 565, 579
 Newcastle upon Tyne, 222, 223, 319–320, 445
 new dialect formation, 20, 33–34, 143–155
 New England, 62–64, 186, 192, 242, 244, 246–247,
 250, 272, 288, 359, 424, 450–452, 455–457
 Newfoundland, 31, 249, 450, 452, 458
 Newfoundland English, *see* English
 New Regionalism, 159, 173
 New York City, 9, 107, 111, 112, 118, 153, 191, 192,
 199, 243, 247, 258, 276, 289, 371, 374–375,
 378, 385, 450–452, 454–457
 New Zealand, 33, 34, 146, 148, 149, 159, 439, 440,
 442, 445, 446
 New Zealand English, *see* English
Nihon Hōgen Kenkyūkai (NHK; Dialectological Circle of Japan), 562, 568
 noise, 75, 234, 278, 319, 322, 325, 402, 404
 noisy clustering, *see* clustering
 nonlinear distribution, 19, 58–60, 67, 70, 89,
 215, 409
 non-mobile older rural female (NORF), 166
 non-mobile older rural male (NORM), 47–48, 111,
 166, 241, 242, 309, 359–360, 440, 441
 non-standard dialects, *see* standard and non-standard dialects
 Nordic countries, *see* Scandinavia
 normal distribution, 90, 268, 391–392
 normalization, 276, 308, 316–318, 324, 332, 371
 normalized distance, 335
 normalized formant, 316–318
 normalized Pairwise Variability Index (nPVI), 321
 North America, 8, 19, 30, 61, 75, 85, 86, 107, 146,
 248–249, 278, 296, 318, 436, 440, 450–458,
 475, 478, 549
 North American English, *see* English
North American Regional Vocabulary Survey (NARVS), 249, 452
 North Carolina, 27, 318
 Northern Cities (Vowel) Shift (NCS, American English), 61, 114, 197, 200, 318, 378, 454, 457
 Northern Ireland, 225, 305
 Northumberland, 26, 245–246, 440, 443–445
 Norway, 111, 117, 145–146, 149, 153, 164, 341,
 407, 463
 Norwegian, 145, 146, 149, 159, 205, 212, 321,
 341–342, 463, 466, 468–469
 not contraction, 415–416, 420–431
Nouvel atlas linguistique de la France (NALF), 127, 479
 nugget, 427
 null-subject cycle, 361
 object nonconcord, 280
 observation
 areas, 416
 points, 416, 429
 observer's paradox, 31, 32, 49, 50, 234, 272,
 274–275, 293–294, 297
Obščeslavjanskij lingvističeskij atlas, see All-Slavic Linguistic Atlas
 occlusion, 319–320
 Odder (Denmark), 112, 117, 120, 169, 170
 Odense, 115, 120
 offset, *see* vowel
 Ohio, 270, 316–318, 452, 457
 Oklahoma, 32, 181, 286, 290–291, 296, 297, 451
 Old English, *see* English
Ømålsordbogen (Dictionary of the Insular Dialects),
 107, 109, 466
 online survey, 249, 260, 262, 263, 281
 onomasiology, 42–46, 51–53, 302
 onset, *see* vowel
 open(-ended) question, 253, 269, 277–278, 280
 opinion testing, 206, 208, 209
 optimality theory (OT), 89–90
 oral investigation, 49–50
 oral tract, 89, 109, 221–222, 316
 orthography, 27, 36, 39, 41, 46, 49, 50, 204, 205,
 207, 210, 213–214, 263, 269, 301–303, 308,
 356–358, 374–375, 479, 491, 531, 556
 Orton, Harold, 29, 244–246, 254–257, 289, 306,
 440–441
 Oxford, 29, 30, 439
Oxford English Dictionary (OED), 223, 224
 paired-samples *t*-test, 274
 panel study, 112, 117
Pan-Slavic Linguistic Atlas, see All-Slavic Linguistic Atlas
 PARAFAC, 406
 parameter, 42, 48, 51, 83, 89, 98, 133, 135, 137, 208,
 228, 260, 265, 302, 314, 373, 387–392,
 394–396, 398, 549, 579
 aliased, 394
 map, 129, 133, 135, 137–138
 Paris, 124, 127, 161, 255, 270, 474, 475, 480
 Paris, Gaston, 61, 62, 71, 124, 125, 400, 481–482

- parole*, 70, 74
 part of speech, 303, 308, 333
 part-of-speech tags, 333
patois, 41, 61–62, 474–475, 479
 peak delay, 323
 Pederson, Lee, 57–58, 60, 62, 69–71, 297
 Pennsylvania, 195, 196, 362, 371, 378, 456–457
 perception, 20, 60–62, 70, 118, 177–200, 211, 221,
 228, 253, 259, 261–262, 319, 324, 334–335,
 340, 341, 362
 Perceptual Assimilation Model (PAM), 211
 perceptual dialectology, 10, 20, 118, 177–200, 259,
 262, 310, 354, 470
 Pfeffer corpus, 172
 phenograms, 33
 Philadelphia, 107, 118, 195, 371, 374–375, 378, 385,
 450, 452, 454, 456
 philology, 17, 18, 23, 27–28, 60–61, 71, 124–126
 philosophy, 17, 28, 43, 92
 phonation, 220, 222, 322, *see also* voice
 phone frequency method (PFM), 331
 phoneme, 58, 70, 73–82, 84–86, 110, 147, 210, 211,
 214, 215, 259, 272, 276, 404
 phone string comparison, 334
 phonemic merger, *see* merger
 phonetics, 4, 6, 9, 17, 19, 26–31, 49, 58, 61–63, 73,
 76, 78–82, 84, 90, 92, 96–97, 99–101, 106,
 107, 109, 111, 118, 125, 128, 129, 132, 151,
 184, 195, 206, 208, 210–215, 219–229, 236,
 244, 246–247, 253–254, 256, 263, 269–270,
 276, 278, 282, 284, 287–288, 302, 304–305,
 310, 314–329, 331–340, 359, 375, 378, 406,
 452, 493, 501–504, 554
 distances, 208, 210–211, 214–215, 331–340
 transcription, 6, 28, 58, 62–63, 68, 73, 76, 81,
 109, 125, 211, 228, 236, 256, 278, 288, 331,
 336, 340, 435, 455, 491, 499
 phonological space, *see* vowel, space
Phonologie du Français Contemporain, 305, 480
 phonology, 4, 7, 17, 18, 46, 48, 50, 52, 58, 61, 73,
 74, 76, 78–85, 89–90, 95, 97–100, 102, 109,
 110, 144, 146–148, 151, 154, 162, 166, 195,
 206, 208, 210–212, 215, 219, 223, 253, 255,
 257–263, 269, 276, 279, 282, 291, 295–296,
 302, 305, 314, 319, 323–324, 349–350, 372,
 379–380, 408, 439–440, 456, 457, 465, 470,
 480–482, 501–504, 526, 531, 547–549,
 554, 560
 phonological process, 276, 279
 pidgin, 35, 357, 573
 pilot study, 47, 62, 255
 pitch, 144, 221, 226, 319, 321–324, 519
 accent, 322–324
 excursion, 324
 nuclear pitch accent, 323–324
 pre-nuclear pitch accent, 324
 Pittsburgh, 457
 pluricentrism, 39, 40
 poetry, 526, 543
 point-diagram map, 350, 352
 point pattern analysis, 431
 point-symbol map
 proportional, 350
 qualitative, 350
 point-text map, 353
 pointwise mutual information (PMI), 336–337, 341
 PMI Levenshtein, 336–337
 Poland, 77, 79, 516
 polarity
 negative/positive, 386–387, 389–397, 408, 478
 Polish, 77, 84, 516, 518
 dialects, 516–517
 polytomous variable, 384, 389, 398
 postal survey, *see* survey, mail
 Portugal, 5, 499, 501, 503–505
 Portuguese, 5, 33, 39, 159, 207, 322, 436, 437,
 498–505, 574
 Brazilian, 159, 322, 501, 504–505
 dialects, 5, 33, 39, 207, 437, 498–501, 503–505
 European, 322, 499, 500, 503, 504
 pragmatics, 43, 119, 194, 302, 309, 361, 374, 469, 505
 Prague School, 74, 78, 510
 pre-aspiration, 319
 predictor variable, 169, 385–386, 388–389, 392,
 398, 407–409, 478
 prejudice, 3, 4, 207, 208, 212, 228
 prescriptive grammar, 279–280, 480
 prescriptivism, 23, 25–26, 34, 224, 278–280, 480, 505
 prestige, 3, 4, 8, 91, 159–163, 165, 222, 279, 310,
 439, 446, 477, 510, 525, 541, 549, 575, 578
 Preston, Dennis, 20, 60
 presupposition, 119, 194–195
 prevocalic /r/, *see* rhoticity
 primary determinants, 111–112, 116
 principal component analysis (PCA), 404–406
 prior distribution, 377
 probabilistic map, 362
 probabilistic reasoning, 92, 97, 98, 100, 308, 362,
 373, 376–378, *see also* model, statistical
 pronoun, 59, 79, 94, 99, 110, 305, 308, 361, 371,
 493, 494, 498, 502–504, 513, 515, 527, 529,
 531, 573, 575, 579
Pronunciation of English in the Atlantic States
 (PEAS), 69, 80, 404, 451
 prosody, 90, 117, 119, 211, 220, 225, 314, 319,
 321–324, 343, 463, 469, 470, 511, 518
 prosodic rhythm, *see* rhythm
 protocol, 86, 234–235, 288–290, 296, 441, 480
 Proto-Indo-European (PIE), 7, 28, 533
 proto-languages, 7, 28, 359, 463, 468, 571, 579
 prototype, 51, 111, 160, 161
 Provençal, 161, 493, 494

- Alps Provençal, 493, 494
 Franco-Provençal, 138, 139, 355–356, 476, 477,
 479, 481, 490, 491, 493, 495
 pseudo-passive, 304
 psychology, 91, 97, 99, 113, 118, 188–191, 195
 evolutionary, 18
 social, 113, 118, 188, 191, 195
 Punjabi, 541
Putonghua, *see* Chinese
 p-value, 236–237, 391–392, 407, 421
Pygmalion, 6, 223
- quantification, 19, 33, 109, 128, 208, 214, 343
 quantile-quantile normal plot (QQ-normal plot),
 392–394
 quantitative (analysis), 9, 290, 301, 306–307, 309,
 323, *see also* dialectometry and models,
 statistical
 Quebec, 186, 458, 478–481
 Quebec French, *see* French, Canadian French
 questioning, 255, 256
 questionnaire, 6, 18, 29–31, 49–51, 58, 60, 62, 97,
 125, 127, 165–167, 173, 174, 235, 238,
 244–250, 253–265, 268–282, 287, 289–292,
 295–296
 correspondence, 270
 formal, 256
 postal, 62, 245–246, 249, 254, 270–273, 275, 287,
 440–442
 sociolinguistic, 247
 web questionnaire, 235, 238, *see also* online
 survey
 question type, 51–52, 256–257
 closed, 253, 277, 280
 danger-of-death, 290
 fill-in-the-blank, 256, 280, *see also* open
 questions
 open-ended, 269, 277–278, 280
 shotgun, 294
- R (statistical computing environment), 390,
 411, 415
 /r/, *see* rhotic, rhoticity
 r-lessness, *see* rhoticity
 race, 111, 154, 222, 243, 370, 378, 575, 578
 Rand Index, 341
 random effect, 407–409, 412
 randomness, 31, 94, 137, 242–243, 255, 286, 288,
 297, 318, 341, 372, 377, 402, 407–408, 412,
 417, 420, 421
 rapid anonymous interview, 31
 reallocation, 151–152
 Received Pronunciation (Standard British
 English), 3, 4, 222, 223, 440, 442
 recorded text testing (RTT) method, 204
 Reddit, 368
- Regensburg-Salzburg dialectometry (RS-DM),
 128–140
 regiolect, 162, 163, 166, 167, 169, 170, 172
 regionalization, *see* dialect leveling
 regional language, 39, 41, 47, 53, 138, 462,
 463, 470
 register, 31–32, 52, 207, 225, 262, 300–301, 380, 408
 regression, 238, 273
 design, 343, 407–412
 logistic, 237, 307, 384–399, 407–410
 mixed-effects, 116, 398, 407–412
 model, *see* regression design
 spatial, 431
 surface, 410
 regularity, 115–116, 128, 130, 137, 165, 242–243,
 261, 417
 sound change, 7–8, 28, 32, 337, 339, 343, 359
 structural, 97, 150, 195, 256–257, 279
 regularization, 302, 373
 relative distance
 relative distance value (RDV_{jk}), 133
 relative edit distance, 335
 relative frequency, *see* frequency, relative
 relative identity value (RIV/RIV_{jk}), 131, 133,
 138–139, 330
 relic area, 306, 339
 religion, 243, 543, 544, 574
 replicability, replicable, 236, 314, 417, 430, 431
 respondent, *see* informant
 response variable, 385–386, 389, 407–410
 rhotacism, *see* rhoticity
 rhotic (consonant), 149, 320, 441–443, 455–456,
 477–478
 rhoticity, 29, 146, 149, 184, 404, 441–446,
 455–456, 493
 rhyme, 81, 86, 260, 263, 276
 test, 260, 263, 278
 rhythm, 321–322
 Romance
 geolinguistics, 123
 languages, 5, 33, 123, 125–126, 128–130, 132,
 137–138, 140, 207, 353, 356, 436, 474–505
 rural dialects, 2, 6–8, 75
 Russian dialects, 511–515
 Russian Federation, 362, 511
 Ryukyu dialects, *see* Japanese
- St. Louis, 457
 salience, 117, 160–162, 168, 172, 173, 184, 192, 195
 same-word comparison, 335
 sampling, 18, 70, 75, 94, 107, 110, 113, 124, 143,
 153–154, 233–238, 241–250, 257, 270–271,
 274, 276, 288–289, 297, 300–303, 308–309,
 317–318, 332, 340, 371, 377, 388, 398,
 400–401
 convenience, 242–243, 398

- sampling (*cont'd*)
 judgment, 243
 quasi-random, 243
 quota sample, *see* stratified sample
 random, 94, 242–243, 286, 288, 297, 318, 407, 412
 sample size, 237–238, 340, 402
 speech, 53–54, 62, 184, 185, 195, 198, 214, 221,
 225–226, 228–229, 243, 247, 257–259, 261,
 321, 324, 331, 334
 stratified, 242
 systematic, 243, 297
 Sapir, Edward, 74, 91
 Saussure, Ferdinand de, 69, 70, 74, 75, 89, 102
 scale-free network, *see* network
 Scandinavia, 5, 28, 107, 115, 117, 205, 208–211,
 254, 357, 462–471
 Scandinavian dialects, 463, 466, 468
 Schleicher, August, 28, 32
 Schmidt, Johannes, 32
 Schneider, Edgar, 33, 34, 140
Schwäbischer Dialektatlas, 168, 169
 Scotland, 34, 84, 149, 173, 187, 223, 244–246, 250,
 254, 259, 262, 270, 272, 303, 305–306, 320,
 322, 324, 439–445
 Scottish English, *see* English
 scree plot, 404, 405
 Seattle, 458
 second dialect acquisition, 143, 144, 146–149, 151
 Second Germanic Consonant Shift, 466–467
 second language learning, 4, 11, 152, 204, 211,
 320, 333
 SED, *see* *Survey of English Dialects*
 Séguy, Jean, 123, 127–128, 139, 330, 334, 400, 407
 Séguy's curve, 407
 Seibu dialects, *see* Japanese
 self-observation, 48–49
 self-report, 268, 286–287, 289, 290, 296
 semantic field, 255, 263, 269
 semasiology, 42–46, 52
 semivariogram, 426
 Sense Relation Network (SRN), *see* network
 separation of interviewing and transcription,
 288, 297
 Serbian dialects, *see* Bosnian/Croatian/Serbian
 dialects
 Serbo-Croatian, *see* Bosnian/Croatian/Serbian
 dialects
 sex, 9, 66, 111–113, 116, 198, 243, 245, 275, 277,
 315, 324, 406–407
 sexuality, 44
 Shakespeare, William, 24, 222
 Shaw, George Bernard, 6, 223
 Sheridan, Thomas, 24, 26, 220, 223
Shōgakkan's Dictionary of Japanese Dialects (SDJD),
 561–562
 shrinkage, *see* regularization
 significance, statistical, 95, 109–110, 193, 208–210,
 215, 237–238, 242, 387–389, 391–392,
 394–395, 398, 420–421, 423
 similarity map, 129, 133–134, 137, 349, 356–357, 363
 similarity values, 131, 134, 354
 simplification, 27, 150, 162
 Singapore, 571
 single-link clustering, *see* clustering, nearest
 neighbor
 skew, 135, 138, 235, 244, 271, 279–280, 309, 400
 slang, 115, 263, 374–375, 379
 Slavic language and dialects, 84, 173, 359, 401,
 510–520
 /s/-lenition, 389
 Slovak dialects, 517–518
 Slovenian dialects, 518
 smart phone, 234, 238
 smoothing, 377, 409–410, 422–424
 social class, 1–4, 8–9, 29, 110–112, 116, 144, 145,
 173, 221, 222, 227, 243, 248, 289, 300,
 302–303, 348, 359, 370, 378
 social correlate, 280–281
 social dialectology, *see* socio-dialectology
 social dialects, *see* dialect, social
 social embedding, 111, 112, 114–118, 181
 social identity, 2, 9, 11, 20, 34, 114, 116–118, 155, 169,
 186, 197, 226–228, 259, 262, 272, 310, 380
 social media, *see* media, communication
 social network, *see* network
 social stigma, *see* stigma
 socio-dialectology, 9, 19, 99, 106–120, 233, 235,
 237, 249, 257–259, 270, 281, 301, 390, 408,
 411, 441, 451
 socio-indexical, 100, 173, 200, 259
 sociolect, *see* dialect, social
 sociolinguistics, 2, 8–10, 18–20, 29, 30, 33–35, 48,
 49, 53, 58, 69, 73, 85, 88, 91, 92, 95, 99, 100,
 106–109, 111–112, 114–115, 117–119, 148,
 152, 155, 160, 172, 173, 204, 219, 222, 226,
 228, 234, 237, 242, 258–259, 276, 281, 309,
 357–359, 378, 380, 411, 451–452, 481, 482,
 505, 526, 569
 comparative, 305, 306
 icons, 117
 variationist, 9, 19, 106, 109, 115, 118, 155,
 301, 307
 sociophonetics, 9, 97, 106, 227, 314–329,
 454–458, 482
 sonorant, 319, 332, 335
 Sorbian dialects, 518
 sound change, 7–8, 28, 32, 61, 84, 85, 96–97,
 99–101, 211, 248, 259, 263, 296, 337–339,
 343, 378–379
 sound correspondence, 337, 339, 343
 South (US), 59–60, 181, 247, 285, 296–297,
 317–318, 321, 371, 378, 450, 456

- South Africa, 40, 150, 227, 439, 440, 442, 445, 446
- South African English, *see* English
- Southern Hemisphere English, *see* English
- Southern States *see* South (US)
- Southern States (American) English, 456
- Southern (Vowel) Shift (American English), 454–456
- Spain, 4, 5, 173, 181, 182, 255, 498–502
- Spanish, 4, 5, 33, 39, 148, 159, 207, 215, 318, 320
- Argentinian Spanish, 320
 - Caribbean Spanish, 389
 - Castellano, 4, 501
 - Cuzco Spanish, 322, 324
 - dialects, 4, 5, 33, 39, 207, 318, 320, 322–324, 389, 437, 498–503
 - Dominican Spanish, 320, 323, 498, 502, 503
 - Latin American Spanish, 500, 502–503
 - Lima Spanish, 322, 324, 500
 - Mexican Spanish, 4, 320, 323, 503
 - Peninsular Spanish, 323, 501–503
- spatial autocorrelation, 67, 237, 373, 415, 417–425, 430–431
- spatial regression, *see* regression
- spatial weights matrix (SWM), 418–425
- binary, 418–420, 423, 425
 - continuous, 418–420, 422
 - inverse distance, 418–420, 422–425
 - nearest neighbor, 418–419, 422, 424
 - non-standardized, *see* standardized
 - non-symmetric, 418
 - row-standardized, 418–419, 421
 - standardized, 418–420, 422
- speaker comparison, 225, 228
- speaker profiling, 225, 227, 228
- spectrogram, 315, 321, 323, 334
- speech community, 1–2, 9, 35, 39–41, 47, 85, 88–91, 108, 116, 146, 148–149, 154–155, 180, 205, 226, 258, 261, 263, 268, 368–370
- speech rate, 321, 324
- spelling, *see* orthography
- spline, 409–410
- split
- dialect, 160, 163, 437, 453, 468, 494, 548, 550
 - phonemic, 146, 147, 151, 318, 454
 - tone, 550, 552
- spontaneous speech, 48, 49, 91, 165–168, 170, 234–235, 255–256, 258–259, 261–262, 265, 286, 295
- Sprachatlas der deutschen Schweiz* (SDS), 362, 464
- Sprachatlas des deutschen Reichs*, *see* Deutscher Sprachatlas
- Sprach-und Sachatlas Italiens und der Südschweiz* (SSAIS), 489
- Stack Exchange, 368
- Stammbaum* (family tree), 7, 28, 475, 537, 539
- standard
- English, *see* English
 - language, 2–4, 26, 39–42, 46–48, 50, 52–54, 61, 62, 95, 111, 131, 159–161, 163–165, 171, 172, 224, 233, 238, 242, 244, 254–256, 269, 280–281, 284, 300–301, 310, 318, 342, 348, 359, 411, 462, 463, 469–471, 482, 510, 515, 516, 518–520, 577
 - language ideology, 2–4, 41, 54, 224
 - and non-standard dialects, 2–4, 462–463, 474–475, 486–489, 492, 501, 503, 525–526, 542, 544, 549, 564, 573–577
- Standard American English, *see* English
- standard deviation, 236–237, 321, 356, 402, 421–422
- standardization, 46, 111, 115, 127, 161, 162, 164, 171, 172, 204, 205, 223, 281, 409
- Standard Malay, *see* Malay
- Standard Modern Chinese, *see* Chinese, Putonghua
- starburst method, 185
- statistical significance, *see* significance, statistical
- statistics
- exploratory, 238
 - spatial, 67, 237, 415–432
- stereotype, 3, 6, 149, 192, 195, 197, 199, 200, 212, 226, 228, 285, 477
- stigma, 3, 9, 20, 27, 204, 212, 222, 223, 228, 291, 292, 474, 487, 545
- stochastic clustering, *see* clustering, stochastic
- stochastic model, *see* model, stochastic
- stochastic reasoning *see* probabilistic reasoning
- Stockholm, 11, 161, 162, 213, 463
- straatzaal, 173
- stratified sample, 94, 242
- stress, 41, 75, 84, 95, 96, 109, 224, 321–322
- in MDS, 403
 - timing, 321–322
- string-edit distance *see* Levenshtein distance
- structural dialectology, 9, 73–87, 454
- structuralism, 8, 27, 33, 69, 73–75, 77, 79, 81, 85
- structural linguistics, 74–76, 83, 85
- style, 9, 32, 35, 61, 63, 65, 73, 74, 77, 114, 115, 119, 144, 207, 256, 258, 262, 290, 368, 470
- subconscious bias *see* enumerator effect
- sub-linear curve, 400, 407
- substandard, 40, 46, 48, 53, *see also* standard and non-standard dialects
- substrate, 164, 165, 320, 322, 324, 475, 477, 494, 531, 535
- Südwestdeutsche Sprachatlas* (SSA), 165–170, 174, 350, 464
- sum contrast, *see* contrast variable
- surface map, 356–358, 362

- survey
 dialect, 6, 8, 268–281
 mail, 6, 235, 248–249, 272–273, 275, 289
 written, 126, 234–235, 238, 268–281, *see also*
 questionnaire
- Survey of Anglo-Welsh Dialects* (SAWD), 441
- Survey of Canadian English*, 277, 452
- Survey of English Dialects* (SED), 8, 29, 63, 69, 80, 241–242, 244–246, 250, 254–257, 261, 263, 274, 289, 306, 440, 441, 444, 445
- Survey of Oklahoma Dialects* (SOD), 286, 287, 290, 291, 295–297
- Survey of Regional English* (SuRE), 262, 441
- Survey of Scottish Dialects*, 245, 254, 270
- Swabian, *see* German
- Sweden, 162, 213, 318–319, 356, 463, 466, 469
- SweDia*, 236, 314, 318
- Swedish, 159, 162, 205, 207, 209, 210, 212–214, 314, 318, 321, 356, 406, 436, 463, 466–469
 dialects, 159, 162, 318, 356, 406, 463, 466, 469
- Sweet, Henry, 6, 28, 30
- Swiss French, *see* French
- Swiss German, *see* German
- Switzerland, 30, 62, 111, 124, 125, 140, 147, 159, 164, 185, 186, 188, 227, 363, 462, 464, 468, 477, 479–481, 489, 493
- syllabicity, 335
- syllable
 elision, 321
 offset, 232
 onset, 232
 peak, 323
 rhyme, 93
 timing, 321
- symbolization, pitfalls of, 362–363
- synchronous
 near-synchronous communication, 368
- synchrony, 17, 33, 47, 75, 84, 92, 100, 112, 284, 301, 303, 305, 359–360, 378
- syncretism, 79, 361, 513, 519
- syntax, syntactic, 17, 31, 32, 48, 79, 89–91, 98–100, 110, 206, 212, 234, 236, 238, 249–250, 255, 257, 261–262, 277, 279, 286, 291, 305–308, 331–333, 343, 361, 375–376, 400, 407
 syntactic turn, 361
- Syro-Lebanese Arabic, 530–531
- systematic correspondences, 7, *see also* sound correspondence
- Taalatlas van Noord- en Zuid-Nederland* (TNZN), 350, 465
- Taiwan, 152–153, 194, 556, 559, 571
- tap (alveolar), 149, 170, 320, 445, 446
- tape recorder, 68, 248, 297, 452, 499
- tape recording, 68, 250, 256, 288, 291, 297
- taxatation, 129–131, 133, 330
- telephone, 107, 234, 253
 cellular, 289
 survey, 243, 286–289, 296–297, 318, 360, 452
- Telsur* (telephone survey connected with *Atlas of North American English*), 31, 107, 248, 452
- Tennessee, 456
- tensor reduction, 406
- Texas, 181, 186, 188–190, 199, 243, 247, 292, 297, 456
 /t/-flapping, 145–147, 226
- Thai, 150, 151
- Thailand, 150, 151, 571
- TH-fronting, 32, 445
- three-dimensional Levenshtein distance,
see Levenshtein
- Tobu dialects, *see* Japanese
- Tōjō, Misao, 562
- Tōkyō, 559, 564
- tone
 edge, 322, 324
 in Chinese dialects, 551–552
 lexical, 322, 343
 ToBI (Tones and Break Indices), 322
- toponym, 96, 376
- Toronto, 304, 458
- traditional dialects, *see* dialects
- transcriber, 49, 109, 228, 288
 effect, 276, 288
- transcription, 6, 28, 47, 49, 58, 62, 63, 68–69, 73, 76, 81, 107, 109, 125, 211, 228, 236, 250, 256, 270, 278, 284, 287–288, 297, 300–304, 308, 324, 331–334, 340, 353, 440, 455, 479, 480, 491, 499, 501
 broad, 211, 455
 narrow, 256
- transition zone, 10, 12, 29, 206, 339, 350, 356, 361, 400, 422, 457, 458, 562
- transmission, 17, 39, 41, 42, 154, 487, 488, 525
- treatment contrast, *see* contrast variable
- trend study, 112, 481
- triangle inequality, 400
- trigram, 333, 336, 339
- trill, 320, 477, 502, 504
- triphthong, 315
- Trubetzkoy, Nikolai, 74, 78–79, 510
- Trudgill, Peter, 33–34, 48, 76, 106, 146, 149, 155, 200, 206, 233, 243, 258, 301, 309, 348, 355, 360, 378, 407
- Turin, *see* Italy
- Turkey, 185, 523, 531
- Turkic, 207, 555
- Turkish, 324, 436, 529
- Tuscan, 357, 410
 dialect (Italian), *see* Italy

- Tuscany, 162
 Twitter, 233, 249, 369–380
 typology, linguistic, 354–355, 579
- Ukrainian dialects, 515
 ultrasound (analysis), 320, 324
 umlaut, 167–170
 unbiasedness condition, 430
 United Kingdom (UK), 60, 154, 173, 226, 228, 229, 248
 United States of America (USA), 4, 5, 8, 9, 30, 31, 34, 74, 80, 159, 181, 186, 227, 234, 242, 244, 249, 255, 279, 287–289, 308, 317–318, 321, 370–374, 378, 380, 416, 430, 450–458
 Universal Grammar (UG), 89, 97, 99
 unmonitored speech, *see* spontaneous speech
 unscripted speech, *see* spontaneous speech
 Unweighted Pair Group Method Using Arithmetic averages (UPGMA), 402
 unwritten language varieties, 39, 40, 46
 Upper Saxon, *see* German
 Upton, Clive, 69
 urban dialectology, *see* dialectology
 urban dialects, *see* dialect
 Urdu, 543–544
 USA, *see* United States of America
 usage-based linguistics, 99, 100, 169, 300–301, 306, 308
 phonology, 99, 100
 uvular (consonant), 445, 477
- Vancouver, 458
 VARBRUL, 307, 390, 408
 variable
 acoustic, 416
 alternation, 416–417, 420–421
 individual, 237, 405, 417
 spatially referenced, 416–417, 419–422, 425, 426, 428–429
 usage variable, 279–280
 variant, 271–274, 280, 302
 lexical, 147, 269, 277, 335
 phonetic, 305, 319, 335, *see also* variant,
 pronunciation
 pronunciation, 269, 278, 282
 variation
 systematic, 234
 unsystematic variation, 234, 236
Variation and Change in Dublin English, 32
 variationist linguistics, 234, 237, 411, 481
 variationist sociolinguistics, *see* sociolinguistics
 variogram, 415, 426–428
 empirical, 426–427, 430
 theoretical variogram, 426–427, 429–430
 variogram cloud, 427
- Vehicular Malay, *see* Malay
 velarization, 320
 Veneto dialect (Italian), *see* Italy
 verbal repertoire, 88, 101, 144, 159, 162, 164, 165, 173, 220, 224
 vernacular universal, 333
 Verner’s Law, 28
 Visual DialectoMetry (VDM), 131, 133, 137, 341–342
 visualization, 124, 126, 133–137, 139, 307, 337, 341, 348–350, 354–355, 357, 359, 361–363, 428, 430
 vocabulary *see* lexicon
 voice
 breathy, 319
 creaky, 322
 parade, 228
 quality, 225, 226
 voice onset time (VOT), 320
 voicing *see* consonant, voiced
 Voronoi geometry, 133, 138, 139
 vowel
 advancement, 316
 backness, 334, 404
 classes in English, 455
 duration, 319–320, 404
 formant, *see* formant
 height, 86, 109, 316, 331, 334, 384, 404, 406
 length, 27, 77, 78, 80, 93, 210, 319, 444, 456, 477, 511, 517, 519, 527, 539, *see also* vowel
 duration
 offset, 315, 323
 onset, 315, 323
 quality, 9, 80, 314–315, 317–318, 324, 406, 454, 456, 501
 rounding, 316, 334, 404
 shifts, 454–458
 space, 86, 109, 110, 454, 456, 457
- Wales, 24, 173, 186, 225, 304, 322, 439–441
 Walker, John, 24, 26, 27, 220
was/were, 385–388, 390, 391, 396, 397, 408
 waveform, 321, 322
 wave theory, 32
 weighted identity value (WIV), 330–331
 weighted pair group method using arithmetic averages (WPGMA), 402
 weighting schemes, 331, 334
 Weinreich, Max, 59
 Weinreich, Uriel, 9, 33, 73, 74, 76–79, 81–85, 181, 454
 well-formedness, 91
 Wells, John, 4, 268, 334, 439, 444, 454, 455
 Welsh, 82
 Welsh English, *see* English

- Wenker, Georg, 6, 29–30, 58, 60–61, 63, 71, 139, 165–166, 173–174, 244–245, 254–255, 265, 268–270, 276–278, 281–282, 284–285, 287–289, 296–297, 356, 359–360, 400, 463–465, 470, 471
- Widdowson, John, 31, 200
- Wikipedia, 368, 554
- Wisconsin, 30, 318, 320, 457
- Word Geography of the Eastern United States*, 451
- wordlist, 31, 32, 40, 47, 243–244, 248, 259–260, 282, 289–291, 296, 318, 337
- working map, 129–136
- worksheet, 272, 289, 290, 292, 295–296
- World Englishes, *see* English, as a world language
- World War I, 171, 487
- World War II, 8, 29, 30, 69, 127, 165, 288, 466, 488
- Wright, Joseph, 7, 29, 35, 244, 255, 440
- xbigram, 336
- Yakutia, 362
- y'all, yall*, 59, 79, 287, 290, 296, 371–372, 375–376
- Yang Xiong, 547
- Yiddish, 77, 79, 463, 466
- yinz*, 371–373, 375–376, 380
- yod (palatal glide), 226, 271, 279
- York (UK), 228, 244, 304, 385–387, 390–391, 396, 408
- Yorkshire, 11, 187, 221, 244–246, 440–441, 444–445
- Yue dialects, *see* Chinese

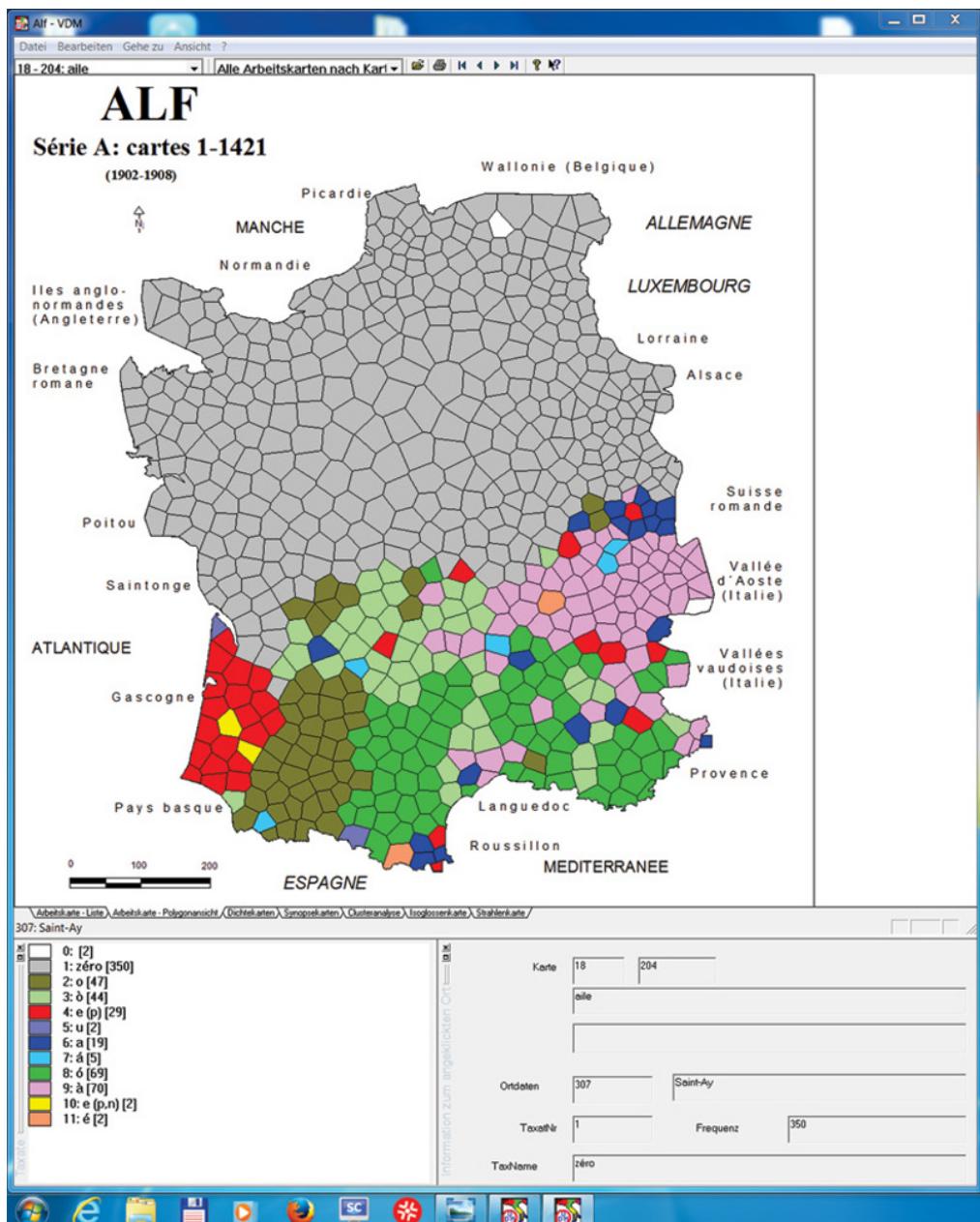


Plate 1 Sample of a *phonetic working map*: spatial distribution of the Gallo-Romance results of final -A in the Latin etymon ÁLA (< Fr. *aile*) “wing” (following ALF 204 *aile*). Cartographic status: *qualitative choropleth map*. See Section 7.7.1, pp. 129–131.

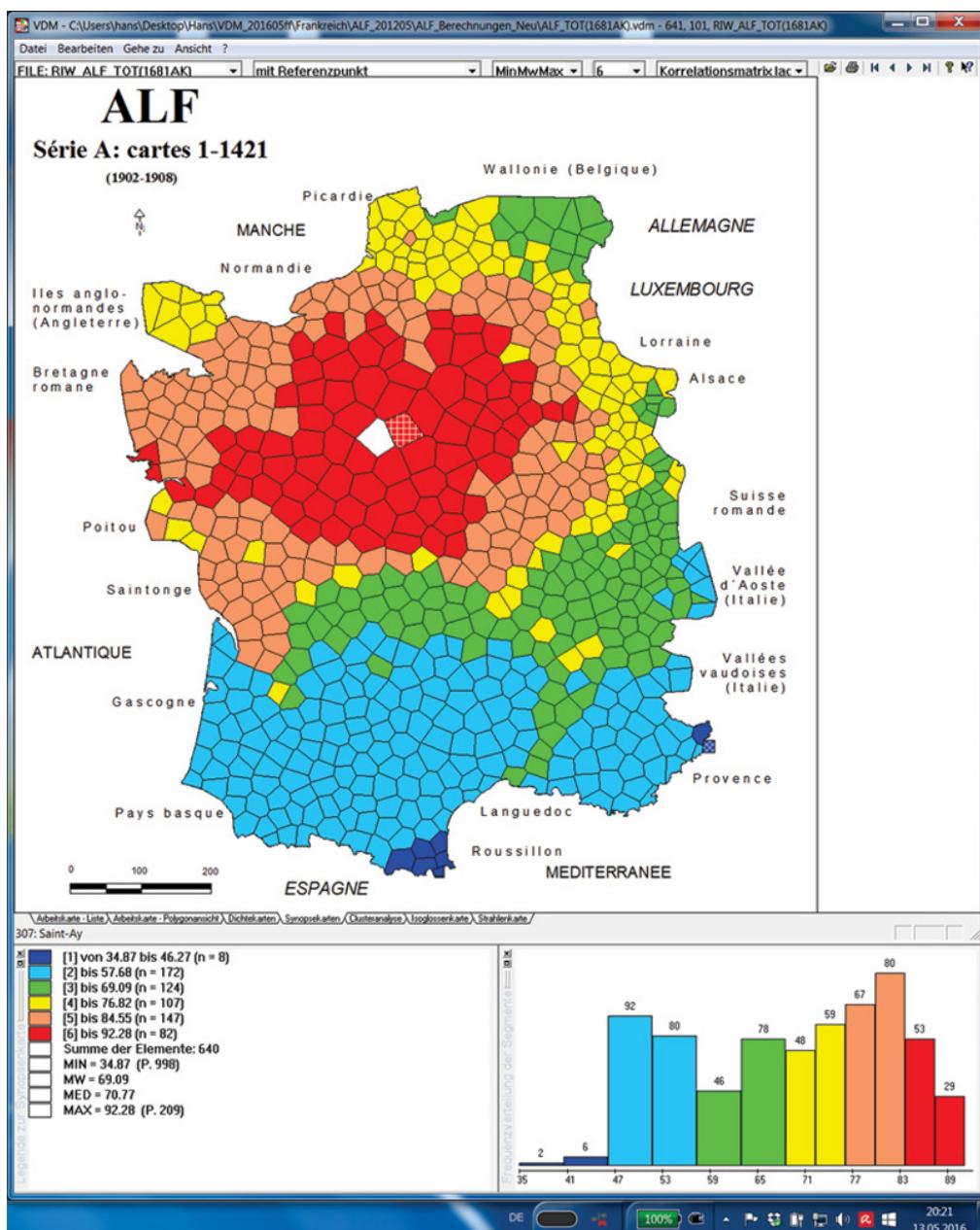


Plate 2 Sample of a similarity map: spatial distribution of the similarity values referring to ALF-point 307 (Saint-Ay, Département Loiret). Similarity index: RIV_{307,k}; corpus: 1681 working maps, all linguistic categories; algorithm of visualisation: MINMWMAX 6-tuple. Cartographic status: *quantitative choropleth map*. See Section 7.7.4, pp. 133–139.

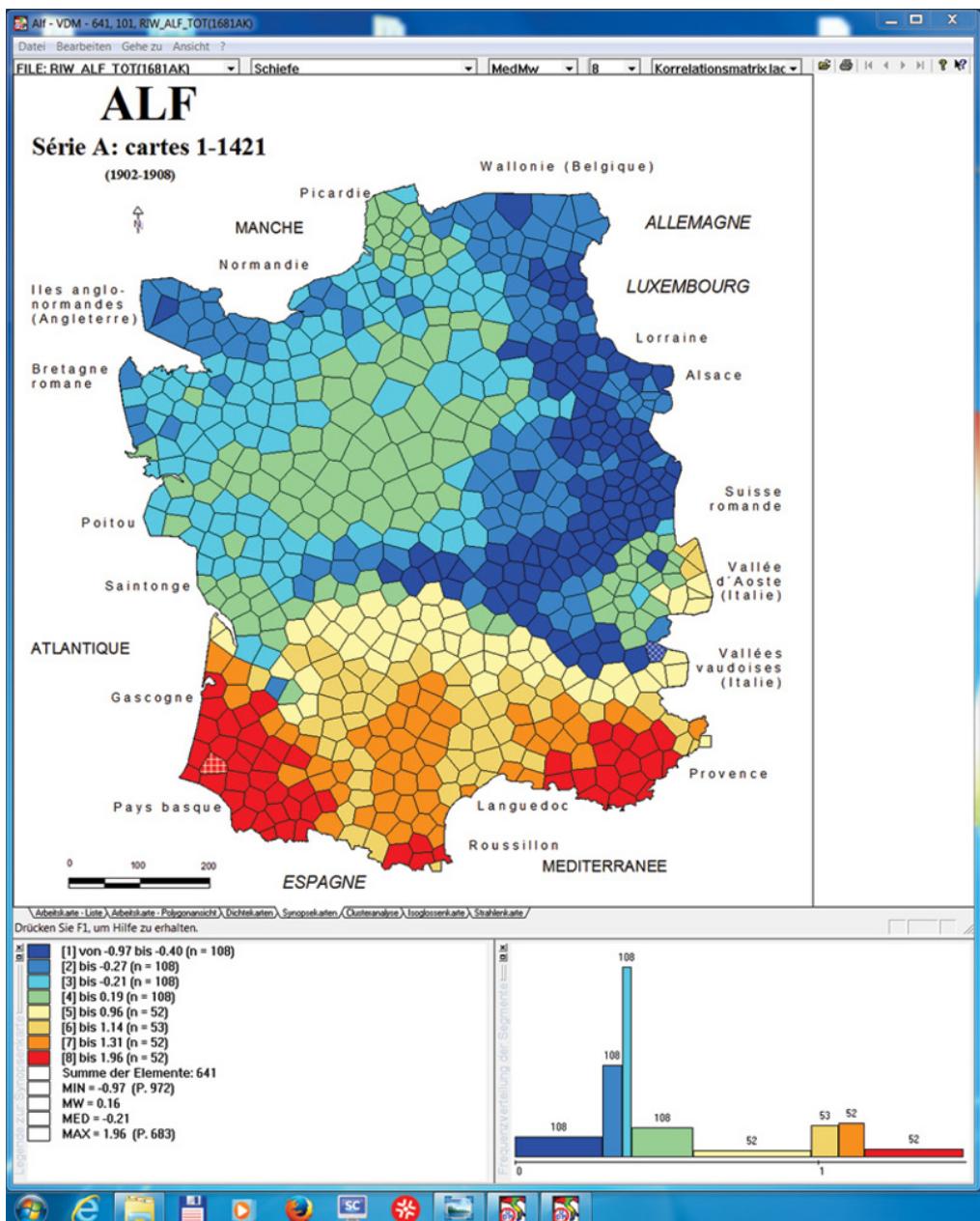


Plate 3 Sample of a parameter map: synopsis of 641 skewness values (according to the asymmetry index of R. A. Fisher). Similarity index: RIV_{jk} ; corpus: 1681 working maps, all linguistic categories; algorithm of visualisation: MEDMW 8-tuple. Cartographic status: *quantitative choropleth map*. See Section 7.7.4, pp. 133–139.

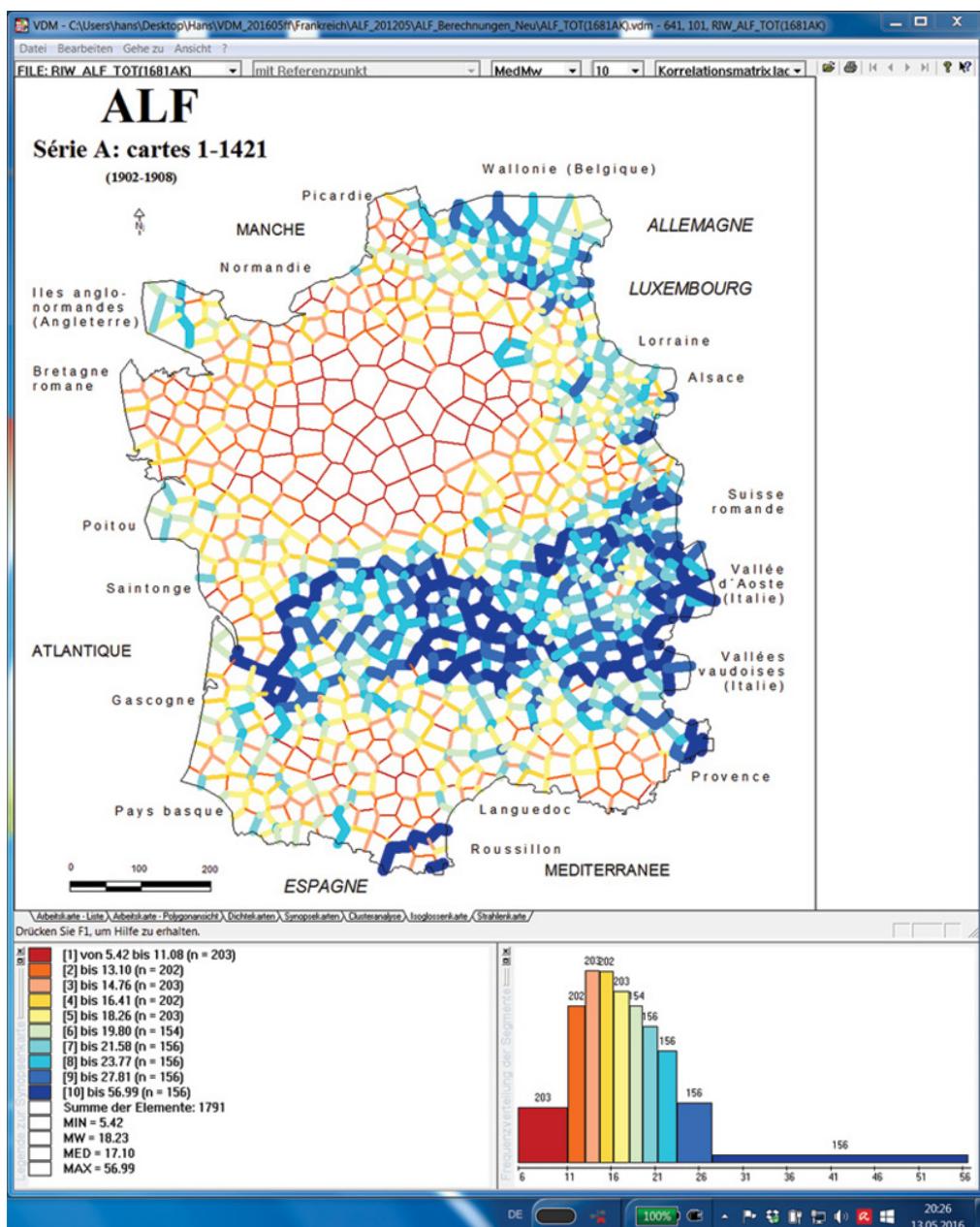
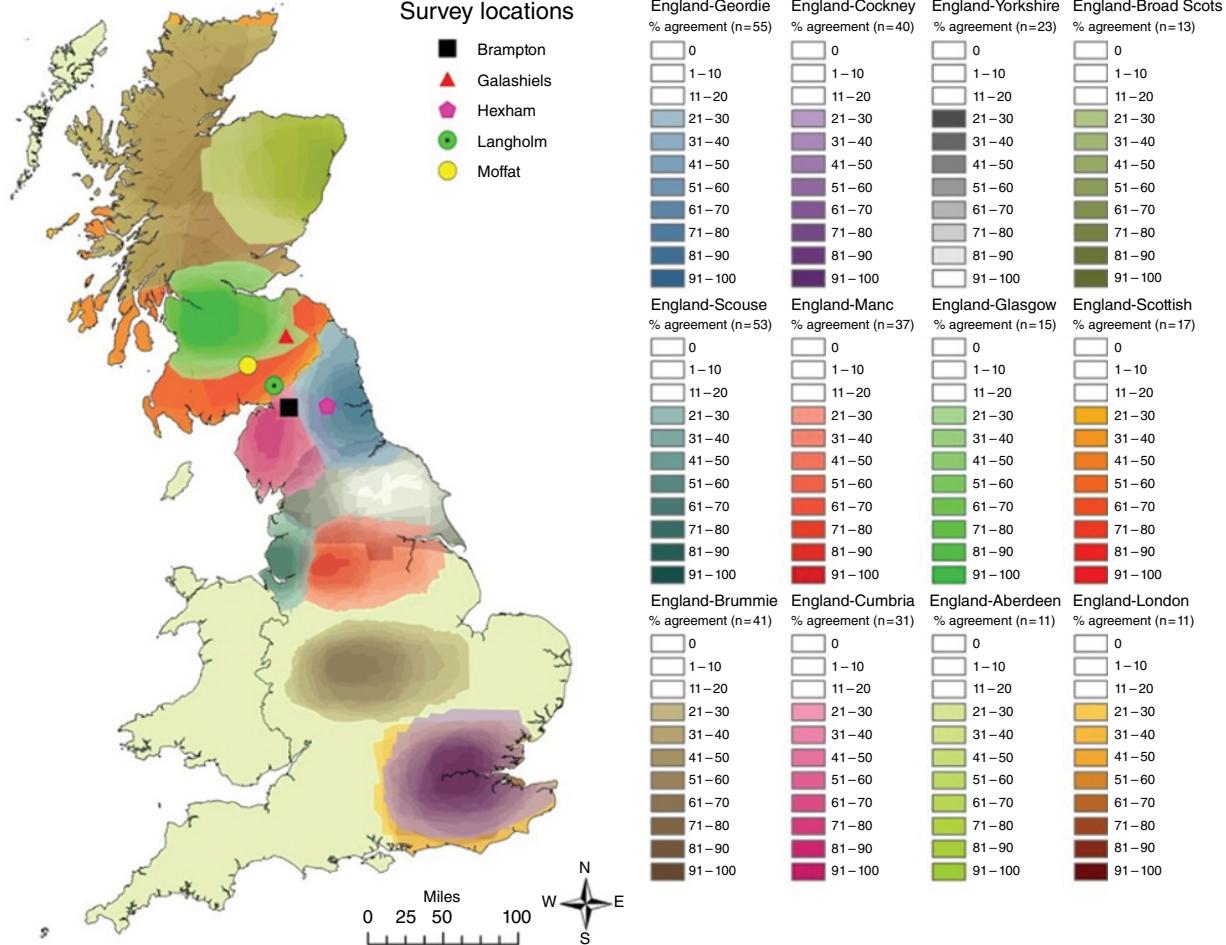


Plate 4 Sample of a interpoint map (honeycomb mode): synopsis of 1791 distance values (according to RDV_{jk}). Distance index: RDV_{jk}; corpus: 1681 working maps, all linguistic categories; algorithm of visualisation: MEDMW 10-tuple. Cartographic status: *quantitative isarithmic* (or: isopleth) map. See Section 7.7.4, pp. 133–139.



This work is based on data provided with the support of the ESRC and JISC and uses boundary material which is copyright of the Crown and the ED-Line Consortium.
Location information is ©Crown Copyright/database right 2011. An Ordnance Survey/EDINA supplied service.

Figure 10.11 A generalized perceptual map of English and Scottish dialects from the point of view of two north of England sites: Brampton and Hexham (Montgomery and Stoeckle 2013, Map 25).

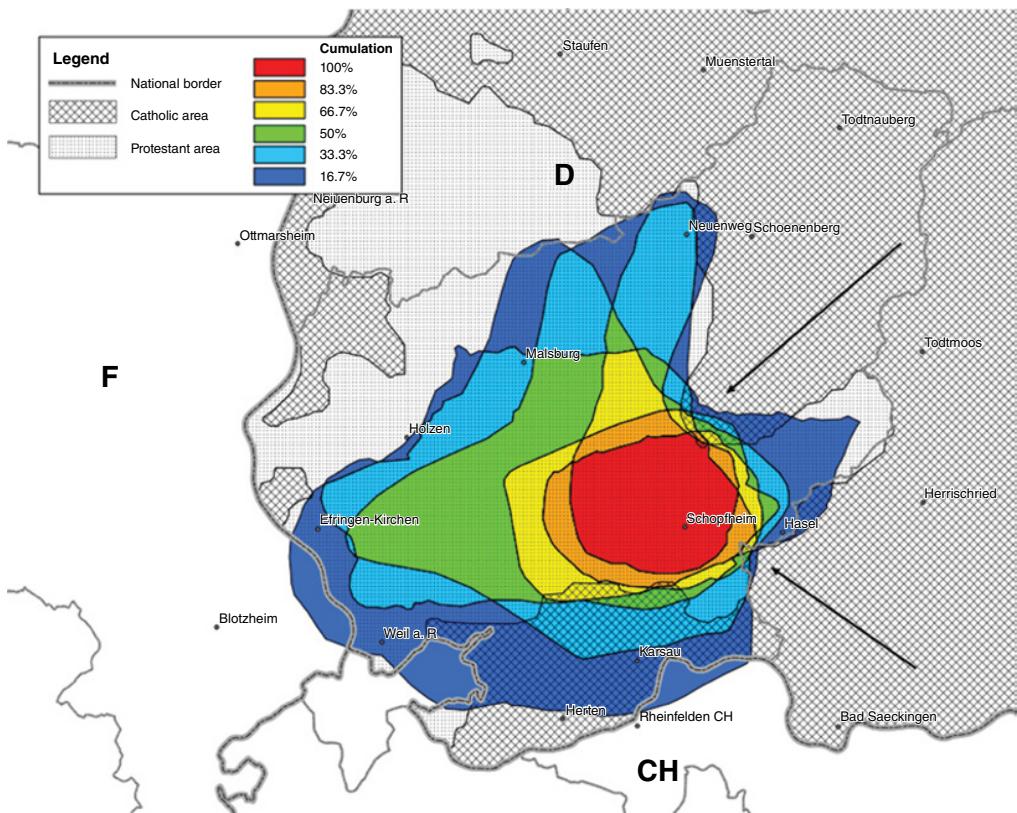


Figure 10.12 A generalized perceptual map of Schopfheim respondent identification of the local dialect area compared to Catholic and Protestant areas in the same region (F=France, D=Germany, CH=Switzerland). Schopfheim is marked with a white arrow (Montgomery and Stoeckle 2013, Map 16).

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley's ebook EULA.