

# (Dia)lects in the 21st century

Selected papers from Methods in  
Dialectology XVII

Edited by

Susanne Wagner

Ulrike Stange-Hundsdörfer

Language Variation 10



## Language Variation

Editors: Martijn Wieling, Alexandra D'Arcy

In this series:

1. Côté, Marie-Hélène, Remco Knooihuizen & John Nerbonne (eds.). *The future of dialects*.
2. Schäfer, Lea. *Sprachliche Imitation: Jiddisch in der deutschsprachigen Literatur (18.–20. Jahrhundert)*.
3. Juskan, Martin. *Sound change, priming, salience: Producing and perceiving variation in Liverpool English*.
4. Dellert, Johannes. *Information-theoretic causal inference of lexical flow*.
5. Zimmer, Christian (ed.). *German(ic) in language contact: Grammatical and sociolinguistic dynamics*.
6. Tahmasebi, Nina, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.). *Computational approaches to semantic change*.
7. Paulsen, Ingrid. *The emergence of American English as a discursive variety: Tracing enregisterment processes in nineteenth-century U.S. newspapers*.
8. van Gijn, Rik, Hanna Ruch, Max Wahlström & Anja Hasse (eds.). *Language contact: Bridging the gap between individual interactions and areal patterns*.
9. Schützler, Ole. *Concessive constructions in varieties of English*.
10. Wagner, Susanne & Ulrike Stange-Hundsdörfer (eds.). *(Dia)lects in the 21st century: Selected papers from Methods in Dialectology XVII*.

# (Dia)lects in the 21st century

Selected papers from Methods in  
Dialectology XVII

Edited by

Susanne Wagner

Ulrike Stange-Hundsdörfer



Susanne Wagner & Ulrike Stange-Hundsdörfer (eds.). 2025. *(Dia)lects in the 21st century: Selected papers from Methods in Dialectology XVII* (Language Variation 10). Berlin: Language Science Press.

This title can be downloaded at:

<http://langsci-press.org/catalog/book/447>

© 2025, the authors

Published under the Creative Commons Attribution 4.0 Licence (CC BY 4.0):

<http://creativecommons.org/licenses/by/4.0/> 

ISBN: 978-3-96110-502-1 (Digital)

978-3-98554-131-7 (Hardcover)

ISSN: 2366-7818

DOI: 10.5281/zenodo.14925847

Source code available from [www.github.com/langsci/447](http://www.github.com/langsci/447)

Errata: [paperhive.org/documents/remote?type=langsci&id=447](http://paperhive.org/documents/remote?type=langsci&id=447)

Cover and concept of design: Ulrike Harbort

Proofreading: Amir Ghorbanpour, Andreas Hödl, Andrew Kato, Brett Reynolds, Christopher Straughn, Elliott Pearl, Erk Deniz Vargez, Georgios Vardakas, Hella Olbertz, Katja Politt, Killian Kiutu, Lea Schäfer, Mary Ann Walter, Jean Nitzke, Rainer Schulze, Raquel Benítez Burraco, Sarah Warchhold

Fonts: Libertinus, Arimo, DejaVu Sans Mono, Source Han Serif ZH

Typesetting software: X<sub>E</sub>La<sub>T</sub>E<sub>X</sub>

Language Science Press

Scharnweberstraße 10

10247 Berlin, Germany

<http://langsci-press.org>

[support@langsci-press.org](mailto:support@langsci-press.org)

Storage and cataloguing done by FU Berlin



# Contents

Acknowledgments	v
Preface	vii
<b>I Geolinguistic methods and big data in dialectology</b>	
1 Extracting “non-standard” data from the Twitter API Kimberley Baxter	3
2 <i>Ain’t</i> + infinitive verb in Black/African American English Kimberley Baxter & Jonathan Stevenson	31
3 The “Atlas of colloquial German in Salzburg” Julian Blaßnigg, Irmtraud Kaiser, Peter Mauser & Konstantin Niehaus	57
4 A cognitive geographic approach to dialectology: Cognitive distance as a predictor for perceptual dialect distance Hedwig G. Sekeres, Martijn Wieling & Remco Knooihuizen	77
<b>II Corpus-based studies and dialect change</b>	
5 A directional shift in linguistic change: A longitudinal study on English-speaking expatriates in Japan Keiko Hirano	107
6 Minimal minimal pairs: Vocalic length in Unterland and Polish Central Yiddish Chaya R. Nove & Benjamin Sadock	123
7 Tracking language change in real time: Challenges for community-based research in the 21 <sup>st</sup> century Katharina Pabst, Sam Brunet, Alison L. Chasteen & Sali A. Tagliamonte	155

## Contents

8	Corpus-based Low Saxon dialectometry Janine Siewert, Yves Scherrer & Martijn Wieling	173
9	Leaner, cleaner, and full of attitude Allison Burkette & Lamont Antieau	201
<b>III Dialectology, linguistic identity, and social factors</b>		
10	“Das ist dann schon total cool zu sagen, <i>Machanot</i> ”: Revealing speakers’ justifications for linguistic choices Esther Jahns	219
11	Regional prosodic variation in the speech of young urban Russians: Quantitative vowel reduction in Moscow and Perm Margje Post	241
12	How important is information about grandparents when selecting a dialect speaker? Akiko Takemura	269
<b>IV Theoretical approaches and innovations in dialectology</b>		
13	Dialectology as “language making”: Hegemonic disciplinary discourse and the One Standard German Axiom (OSGA) Stefan Dollinger	287
14	„Es werden im wesentlichen [sic!] nur Wörter aufgenommen, welche deutlich unterschiedlich zum Hochdeutschen sind.“: On the verticality of lay dialect collections and the attempt to measure it Yvonne Kathrein	319
15	Applying the state-of-the-art tonal distance metrics to a large dialectal dataset Ho Wang Matthew Sung, Jelena Prokić & Yiya Chen	361
16	Convergence and divergence of tone paradigms across Tai dialects in the 21st century Chingduang Yurayong, Saknarin Pimvunkum & Yuttaporn Naksuk	403

*Contents*

<b>Index</b>	<b>435</b>
--------------	------------



# Acknowledgments

The editors would like to acknowledge the support of the following:

- German Research Foundation (DFG grant WA 2432/4-1)
- Johannes-Gutenberg University Mainz (Stufe I inneruniversitäre Förderung)
- Publishers: de Gruyter Mouton, Edinburgh University Press, John Benjamins Publishing Company, Universitätsverlag Winter (Heidelberg)

We are also tremendously grateful to our full-time student helpers, without whom the conference wouldn't have been half as well-organised as it was. Major thanks goes to **Natascha Kessler** and **Johannes Schindler**, who spent too many hours in unairconditioned offices during the hottest time of the year with their heads pretty much in their laptops to figure out how to accommodate all scheduling requests while still coming up with a half-way logical programme (you should have seen the Excel sheet!).

A big thank-you is also due to our other student helpers, **Victoria Fox** and **Lena Jubelius**, who made bag-packing a science and worked overtime willingly because they liked the conference atmosphere so much.

**Christiane Dinger** probably dreamt of money for a couple of months while she was figuring out the finances with our administration – thank you, Christiane! Susanne would also like to thank Ulrike, who took over most of the editorial work while Susanne had to handle some unforeseen commitments (including becoming vice dean) – thank you, Ulrike! Heartfelt thanks also go to **Emelie Lambertz** and **Daniel Kastenholz** for their invaluable assistance with the finer details of the editorial work on this volume, which significantly accelerated the publication process.

Last but not least, the biggest thank-you goes out to our plenary speakers, presenters and participants – obviously, without you, Methods XVII wouldn't have been possible. While we had to wait two years for the Covid pandemic to pass (the conference was originally scheduled for 2020), it didn't take away any of the “Methods magic”. We are looking forward to many more successful conferences in the series.



# Preface

This book offers an in-depth exploration of contemporary issues and methodologies in the fields of dialectology and sociolinguistics. Readers will find a diverse collection of studies that examine how language varies and changes across different regions, communities, and social contexts. The book covers a wide range of languages, including German, English, Yiddish, Russian, and Japanese, providing a global perspective on linguistic diversity. Key themes include the use of modern data sources, such as social media, to study language patterns and the impact of digital communication on regional dialects. The book also addresses the dynamics of language contact in expatriate communities, revealing how speakers adapt and merge linguistic features from different dialects. Several chapters focus on the evolution of dialectological research, offering critiques and new approaches to studying regional language variations. Readers will also encounter innovative methods, such as cognitive geography, which uses mental representations of space to understand dialect variation, and tone distance measures, which are crucial for studying tonal languages. Additionally, the book presents case studies on how non-experts perceive and categorize dialects, providing insights into the public's understanding of linguistic diversity. It also tackles challenges in selecting dialect speakers for research, especially in urban environments, where traditional criteria may no longer apply.

Part I, *Geolinguistic methods and big data in dialectology*, focusses on the application of geospatial analysis, big data, and innovative methods in the study of dialects and language variation. Baxter examines methodology in the use of Twitter in the corpus-based analysis of AAE syntax, focusing on geospatial mapping and data extraction methods. Baxter & Stevenson focus on the creation of an atlas of AAE syntax using Twitter data, with a geospatial mapping of linguistic features across different U.S. regions. Blaßnigg et al. discuss the application of geolinguistic methods in dialectology, using data from the Atlas of Colloquial German in Salzburg, and highlights cluster analysis of regional dialects. Sekeres et al. explore the use of cognitive geography in explaining dialect variation, comparing cognitive and geographic distances in the perception of dialect differences.

Part II, *Corpus-based studies and dialect change*, brings together studies that use corpora to analyze language change, dialect contact, and linguistic variation

over time. Hirano analyses linguistic change in an English-speaking expatriate community in Japan, focusing on dialect contact and long-term accommodation. Nove & Sadock use archival data to study the acoustic correlates of vowel length contrasts in Central Yiddish, highlighting dialect change in the Transcarpathian region. Pabst et al. discuss longitudinal research on language change, focusing on methodological challenges in tracking linguistic variation over time. Siewert et al. investigate changes in Low Saxon dialects over time, using corpus-based methods to explore dialect similarities and divergence across the Dutch-German border. Burkette & Antieau review the evolution of the Linguistic Atlas Project's fieldwork methods from mapping dialect boundaries to recording variation and sociolinguistic attitudes, highlighting changes in interview methods and goals.

Part III, *Dialectology, linguistic identity, and social factors*, explores how linguistic variation relates to social identity, language ideologies, and the intersection of language with social factors. Jahns introduces the linguistic-positioning task to study linguistic choices in the Jewish community in Berlin, emphasizing the impact of language ideologies on speaker choices. Post investigates regional prosodic differences in modern urban Russian speech, focusing on how these differences persist despite standardization. Takemura re-evaluates the criterion for selecting local dialect speakers in Japanese dialectology, emphasizing the role of parental origin in linguistic identity.

Part IV, *Theoretical approaches and innovations in dialectology*, encompasses theoretical discussions and methodological innovations that challenge traditional dialectological frameworks. Dollinger critiques the anti-pluricentric perspectives in German dialectology, proposing a new framework for understanding linguistic standards and dialects. Kathrein evaluates dialect collections from the Tyrol region, comparing entries with standard language to understand how laypersons conceptualize their dialects. Sung et al. compare tone distance measures in Sinitic dialects, proposing improvements to dialectometry methods, especially for tonal languages. Yurayong et al. examine tone paradigms in Tai dialects, applying a historical-comparative approach to dialectal classification and language contact.

Overall, this book is a valuable resource for linguists, researchers, and anyone interested in the complex and ever-changing landscape of human language. It highlights the importance of adapting research methods to keep pace with the evolving nature of language and offers fresh perspectives on how we study and understand dialects and language variation.

## Part I

# Geolinguistic methods and big data in dialectology



# Chapter 1

## Extracting “non-standard” data from the Twitter API

Kimberley Baxter

New York University

The present paper examines methodology in the use of Twitter in the corpus-based analysis of African American English (AAE) syntax. I discuss the extraction and geospatial mapping of indices of use of the perfective marker *done* (hereafter *perfective done*), which, alongside a simple past form, indicates that the action described in the past form has been completed. Widespread use of AAE on archived social media posts creates a living database of timed, dated, and geotagged utterances from which corpora may be built. The Academic Twitter API (ACTW) allows access to their full database of tweets, which is a much larger and more accessible dataset than its large social media contemporaries. I discuss two methods of extracting perfective *done* from the ACTW: a *front-end* approach which aims to isolate uses of perfective *done* by eliminating non-perfective uses of *done* from the search prior to running the query, and a *back-end* approach which first extracts a set of all uses of *done* from 2012–2015 and aims to isolate uses of perfective *done* afterwards. I discuss the results of this method, as well as the implications therein and directions for future research. I conclude that while both methods are effective at extracting perfective *done* from the ACTW, the *back-end* approach is better suited to geospatial mapping.

## 1 Introduction

### 1.1 General remarks

Despite longstanding myths of African American English (AAE) as a linguistic monolith (Wolfram 2007), numerous sociolinguistic studies indicate that AAE



shows regional variation in both phonology and lexicon (see Wolfram 2007, Lanehart 2015, etc. for phonology; Jones 2015, Grieve 2016, etc. for lexicon). Far fewer studies focus on syntactic variation in AAE across regions, and of those that do, many are restricted to a certain geographic area (e.g. Terry 2010, Moody 2011), with Baxter (2025 [this volume]), Baxter & Stevenson (2025 [this volume]), and Masis et al. (2022) constituting notable exceptions. This study aims to fill this gap by using geotagged Twitter data to investigate syntactic variation in AAE across the contiguous United States.

The goals of this study are to:

- a) Present a method for the targeted search of AAE parts of speech which maximizes the number of desired features while minimizing the number of false positive features, which are orthographically identical to the desired feature, yet still syntactically different;
- b) Calculate indices which reflect the rate of feature use by geographical region;
- c) Correlate these indices with demographic data to discover whether indices of feature use vary across locations with a high density of Black/African American people.

By doing so, this paper seeks to address the myth of supraregionality in AAE (Wolfram 2007), which suggests nationwide grammatical universality in urban centers (see Section 1.2).

This paper focuses on the methodology used to plot the geographic distribution of one such feature canonical to AAE: perfective *done* (Rickford 1999, Green 2002, 2010). I describe two approaches to the collection of perfective *done* tokens from the Academic Twitter API (ACTW): a *front-end* approach, through which grammatical restrictions are applied prior to extracting tweets, resulting in a dataset comprised mostly of tweets using perfective *done*, and a *back-end* approach, through which grammatical restrictions are applied after extracting tweets, resulting in a dataset comprised of all uses of *done*, from which a subset of perfective *done* may be extracted.

## 1.2 African American (Vernacular) English

This paper uses *African American English* (AAE) to describe the language previously known as *African American Vernacular English* (AAVE), *Black Vernacular*

*English* (BEV), etc. In doing so, I aim to distance myself from the “vernacularization” of AAE introduced by previous research which focused mainly on male street youth in the inner city (e.g. Labov 1972). AAE is one of many African American languages which fall under the umbrella of AAL, including, but not limited to, Gullah Geechie, Black American Sign Language, Louisiana Creole, and many more. AAE is a sociolect spoken mostly by Black Americans and people who live in community with Black Americans. In discussing AAE, it is also important to consider race and ethnicity in the United States and how that informs not only what AAE is, but who speaks it, and how African American and other Black people in the US are categorized demographically (see Blake (2014) and King (2020), both of which speak to issues of the racialization and categorization of AAE).

The US Census does not account for differences in ethnicity among Black people in the United States – instead, all Black-identifying people are encompassed within the same singular category, *Black/African American*. This category only seeks specification regarding whether one is “Black/African American alone” or “in combination” with other races/ethnicities. This is an issue which affects all studies of AAE which seek to utilize the US Census as a touchstone for demographic metadata, including the proposed dissertation. I use *Black/African American* (henceforth: BAA) in discussions regarding the racial and ethnic identity of Black people in this paper.

### 1.3 Why investigate variation in AAE?

While research on the systematicity of AAE sparked a groundbreaking shift in Sociolinguistics as a whole, researchers also inadvertently established a series of myths about AAE and its uniformity (Wolfram 2007):

- (1) The Supraregional Myth: primary structural features setting apart the vernacular speech of African Americans from their European American cohorts were shared by African American Communities regardless of regional context.
- (2) The Language Change Myth: a uniform path of change for African American English, based on the uniformity of AAE
- (3) The Social Stratification Myth: the prevailing assumption that African American English is most commonly used by working-class speakers, especially those who have had little to no contact with other varieties of English. (Wolfram 2007: 295–306)

These myths persist among the definitions of AAE in previous literature:

[AAE is] the uniform grammar used by African Americans who have minimal contact with other [varieties] in contexts where only speakers of that vernacular are present. (Baugh 1983, as cited by Labov 1998: 6)

Due to sociopsychological barriers between Blacks and Whites, AAE is a marker of identity. This causes uniformity among Blacks from all over the US. (Rickford 1999, as cited by Johnson 2008: 20)

The impression of AAE as a “uniform” sociolect is widespread in previous literature on AAE. While these definitions were published decades ago, this myth of uniformity across regions, also known as *supraregionality*, still persists in more recent literature. Wolfram (2007) goes into great detail about all three of these myths, but the most salient of these for the purpose of this paper is the Supraregional Myth. While the papers cited in Wolfram (2007) were from decades earlier, this myth is still very much present in relatively recent literature, as seen below:

[AAE] is, in contrast to other North American [varieties], not geographically restricted. Although variation in AAE does exist, AAE in urban settings has been established as a uniform system with suprasegmental norms... (Jørgensen et al. 2015: 10)

The use of Twitter as a source of data allows researchers to collect far more data than would be possible via traditional methods such as surveys and sociolinguistic interviews. Twitter’s Academic Developer License (ADL) allows access to Twitter’s entire archive of tweets, from which a maximum of 10,000,000 tweets may be mined per ADL, per month. Because the archived tweets were produced voluntarily by Twitter users, this method may also circumvent the Observer’s Paradox (Labov 1972) and similar issues which often arise during the process of face-to-face data collection. The use of Twitter data also presents a boon for the collection of parts of speech which are more difficult to elicit via interviews because of the presence of linguistic alternatives, through which similar meanings and ideas may be conveyed. For example, where my previous attempts at eliciting perfective *done* from interviews yielded very few results, Twitter allows me to pinpoint parts of speech associated with AAE (Green 2002, Rickford 1999) specifically. This data can then be mapped geospatially via the metadata included with each tweet and provide insight on the geographical distribution of perfective *done*.

## 2 Analyzing language variation in social media

There are well-documented challenges analyzing non-standard varieties of English commonly used in social media (Plank et al. 2016), and AAE is no different. These challenges fit into two broad categories: grammatical challenges, and demographic challenges. I discuss each of these challenges below.

### 2.1 Grammatical challenges

Standard methods of extracting data from Twitter’s APIs often lack the specifications necessary to isolate parts of speech exclusive to AAE, and differentiate them from similar lexical items in Mainstream American English (MAE) (Jørgensen et al. 2015). This is in part because AAE has a robust verbal system which allows its verbal lexicon to be used as aspect and mood/modality markers in addition to their use as tense markers, as seen below in examples (4), (5), and (6). (4) is an example of perfective *done*), indicating that the action of going to the store has been completed. Example (5) shows perfective *done* being used in conjunction with the simple past form of the verb *do*. Example (6) shows perfective *done* being used in conjunction with *do* in its participle form.

- (4) He done went to the store already.  
He PERF went to the store already.  
'He has gone to the store already.'
- (5) He done did his work.  
He PERF did his work.  
'He has done his work.'
- (6) Now look what you done done.  
Now look what you PERF PART.  
'Now look what you've done.'

All three of these examples are grammatically correct in AAE, yet the perfect and participle forms of *done* in (6) are orthographically identical. For the purpose of this paper, an alternate form of *done* which is grammatically different, yet orthographically identical to perfective *done* is called a *false positive* (FP). Conducting a simple search for the word *done* in the ACTW results in a high number of FPs and data which renders the manual elimination of FPs unfeasible due to the sheer size of the dataset, which numbers in the millions of tweets (see Section 4).

This paper ultimately aims to produce an alternative method which allows the user to eliminate the aforementioned FPs by coding the grammatical constraints of perfective *done*, which would otherwise be inaccessible due to the lack of highly accurate part of speech taggers designed for this task. While the focus of this research is on Twitter, I raise further questions about language variation in AAE in Section 7.

## 2.2 Demographic challenges

Tracking variation in a sociolect such as AAE is difficult on social media sites which grant anonymity to its millions of users. In the case of Twitter, with the exception of verified profiles (indicated with a blue check), confirmation of one's race, ethnicity, gender, or other facets of one's identity are completely optional. As a result, a reliable mass verification of users' demographic data is not currently feasible.

Multiple surveys have been conducted in an effort to tease apart the demographics of Twitter, with varying results. This study refers to the 2018 Pew Research Study "Sizing Up Twitter" to describe the broad demographics therein. "Sizing Up Twitter" is a representative survey of 2791 adult Twitter users surveyed via Ipsos KnowledgePanel, a probability-based online panel of US adults. The survey found that at the time of the study, approximately 80% of all tweets were made by 10% of Twitter users. Twitter users tended to be younger, and more likely to be Democrat or otherwise left-leaning, more likely to be women, and more likely to tweet about politics. Eleven percent of Twitter users identify as Black\* (not including Black people who also identify as Hispanic). Of these, approximately 30% are college graduates, 40% have some college education, and 30% have a high school diploma or have not completed their high school education. The survey does not include statistics on political affiliation, age, or use *by race*.

However, the study does present a representative Twitter demographic which shows a percentage of Black tweeters which is comparable to the general population of the United States in the same year (approximately 13% of the total population of the United States; U. S. Census Bureau 2018). The US Census Bureau recorded 22.5% of the Black population in the United States 25 years and older as having a bachelor's degree in 2015 (Ryan & Bauman 2016). While there are some gaps between the "Sizing Up Twitter" dataset and the US Census Bureau dataset, the lower percentage of college graduates in the US Census Bureau dataset appears to support the finding that Black Twitter users are also more likely to be college educated than the general population.

With the recent sale and transformation of Twitter (now X) under the ownership of Elon Musk, the demographics therein, especially along racial and political lines, may now be very different. In a Washington Post article entitled “Fleeing Elon Musk’s X, the quest to re-create ‘Black Twitter’”, Dwoskin (2023) describes a political shift in Twitter’s formerly left-leaning environment:

Hate speech has surged on the platform. Researchers with the Network Contagion Research Institute (NCRI), a group that analyzes hundreds of millions of messages across social media, discovered an account that included a Nazi swastika in its profile picture tweeting antisemitic memes. Use of the n-word soared by nearly 500 percent, and the slur popped up in the handle of an account authorized by Musk’s subscription service, Twitter Blue. And thus the exodus of Black users began. (Dwoskin 2023)

The data used in both this paper and Baxter & Stevenson (2025 [this volume]) was produced by its users and collected from the ACTW prior to many of these changes, and are reflective of the Twitter captured in Wojcik & Hughes (2019). In addition, the talk from which this paper is derived was given before Twitter became X. For the sake of consistency, this paper uses Twitter to refer to the social media platform now known as X. In addition, there does not appear to be a new term for posts made on the platform; as a result, this paper also maintains *tweet(s)* as the label for posts made therein.

Since this talk was given, the ACTW has been discontinued, or at the very least paused, preventing the collection of further data from this source. However, it is my belief that these methods will still prove useful to researchers who wish to extract “non-standard” data from similar social media APIs, including X’s new Enterprise API product.

### 2.3 Linguistic grouping approach

Because of the relative anonymity of Twitter profiles, this study uses a “linguistic grouping” approach (Horvath & Sankoff 1987), through which linguistic features of AAE are collected and analyzed *before* the categorization and analysis of sociological factors such as race and ethnicity. This method is preferred to the traditional “sociological grouping” approach used in most traditional sociolinguistic studies which, because of the aforementioned anonymity of Twitter profiles, makes it difficult to verify the ethnicities of the users behind each account.

The terms social and linguistic grouping do not mean that sociological consideration predominate in one approach and linguistic concerns in the other,

but only refer to the temporal order in which they enter into the statistical analysis. (Horvath & Sankoff 1987: 180)

The present study starts by choosing a linguistic variable, in this case perfective *done*, and calculating indices of use, mapped across the contiguous United States via the geolocation metadata attached to each tweet. These indices are then compared to indices of BAA population in US Census tracts across the contiguous United States. By doing so, I aim to provide a broad yet comprehensive view of perfective *done* usage rates in high-density BAA communities.

I plan to conduct future research (see Section 7) to properly investigate language differences among first- and second-generation Black immigrants' AAE and the AAE spoken by Black people who are ethnically African American. However, this is beyond the scope of this paper, as well as the limits of what Twitter data or US Census data can currently offer.

### 3 Perfective *done*

Similarly to Blodgett et al. (2016), Stevenson (2016), and Willis (2020), the present study aims to examine syntactic variables in AAE. This should not be confused with studies on the spread and use of lexical items in AAE such as *fleek* (Grieve 2016), *eem* (Jones 2015), or others. This distinction is very important because AAE has a robust verbal system which allows the use of verbal items as tense, mood/modality, and aspect markers. It is not enough to simply search the ACTW for all uses of *done*, *be*, *BIN* (spelled *been* by AAE speakers), or any other part of AAE speech. One must be familiar with the syntactic properties of these items and devise alternate strategies to separate them from their verbal, adjectival, or other counterparts when extracting data from the ACTW. It is with that in mind that I discuss perfective *done*.

Green (2002) describes perfective *done* below. Numbers have been changed for continuity within the present paper:

- (7) a. I told him you dən changed. (Bm, 30's)

'I told him that you have changed.'

A: You through with Michael Jordan I bought you?

(Literally: Have you finished reading the magazine that I bought you with Michael Jordan on the cover?)

B: I dən already finished that. (Bm, 9)

'I have already finished that.'

- b. I dən done all you told me to do. I dən visited the sick. (Bm, 60s, 70s)  
‘I have done all you told me to do. I have visited the sick.’
- c. A: Push your seat.  
B: I dən pushed it.  
‘I have (already) pushed it.’  
A: Push it again. (elderly Bfs on Amtrak)  
(Green 2002: 60)
- (8) a. My homework is *done*.  
b. Have you *done* your taxes?

Perfective *done* is distinct from *done* in its verbal or adjectival form as seen above in example (8) and is most often compared to the MAE perfect markers *have*, *has*, and *had*.

Perfective *done* typically occurs with both stative and eventive verbs, as well as time adverbs and negation (Martin 2018). The present study uses the descriptor *perfective* rather than *completive* because *perfective* is inclusive of uses of *done* which indicate that an action, event, or status has been completed, as well as uses of *done* which indicate that an action, event, or status is still ongoing.

Due to its ability to describe completed actions, events, or statuses, perfective *done* often appears with eventive verbs, which have a natural endpoint. However, in certain contexts, perfective *done* can occur with stative verbs, which do not have natural endpoints. For example, the following sentence includes perfective *done* with the stative verb *know*:

- (9) Long as I done known you (you’ve been chasing after women)  
TMP-long as I PERF known you (you’ve been chasing after women)  
‘For as long as I have known you (you’ve been chasing after women).’  
(Wilson (1986), cited in Martin (2018))

Terry (2010) argues that *done* is perfective because it fits the “four perfect constructions” as outlined in Comrie (1976):

- (10) Perfect of persistent situation: this perfect behaves similarly to stative BIN, in that it indicates that an event began in the past, and is still occurring.  
“Mary *done* lived in Chapel Hill for 6 years.”  
The above example indicates that Mary has been living in Chapel Hill since the distant past, and still does.

- (11) Experiential perfect: indicates that a situation took place or held at some time in the past.  
“They *done* took my car.”
- (12) Perfect of result: the present state is referred to as being the result of the past.  
“And now you *done* messed everything up.”
- (13) Perfect of recent past: temporal closeness to the present is the focus, rather than the completion.  
“John *done* just got here.”

Terry (2010) notes that perfective *done* expresses the “continuing relevance of a previous situation” (Comrie 1976: 56) and, as a result, appears to fit the criteria of a perfect marker.

However, it is important to note that while Terry (2010) makes a convincing argument as to why *done* should be a perfective, the use of the “four perfect” structure in Comrie (1976) does not adequately describe perfective *done* in AAE. Certain uses of perfective *done* do not fit neatly into these categories and can even occupy multiple categories at once. For example, when asking informally about the categorization of the utterance listed in example (9), *Mary done lived in Chapel Hill for 6 years*, several AAE speakers agreed with the judgment that it meant *Mary* still lived there while others disagreed, stating that *Mary* lived there already and now lived somewhere else. This difference in interpretation may be the result of several factors including context taken from voice intonation, context from previous portions of the conversation, or in some cases a lack of context as a result of the utterance being written via text or survey.

In addition, while perfective *done* can be compared to use of the perfect in MAE (*have, has, had*), it is not grammatically *identical* to the perfect in MAE (Martin 2018).

Perfective *done* also appears with adverbs that refer to a time in the past, such as *yesterday* in (14a) below. This is in contrast to the standard English perfect, which cannot appear with such adverbs, as illustrated in (14b). Numbers have been changed for continuity with the present paper.

- (14) a. John *done* baked a cake *yesterday*. (AAE)
- b. \* John *has* baked a cake *yesterday*. (standard English)

This may be surprising given that the standard English perfect is otherwise quite similar in meaning to perfective *done* (Martin 2018).

To my knowledge, there is no research that specifically examines perfective *done* in Northern cities such as New York City, but there is a wealth of knowledge on perfective *done* in other regions, particularly the Southern United States (e.g. Terry 2010, Green & Roeper 2007).

I discuss two methods of extracting perfective *done* from the ACTW: a *front-end* approach which aims to isolate uses of perfective *done* by using syntactic parameters to eliminate non-perfective uses of *done* from the search prior to extracting data from the ACTW, and a *back-end* approach which first extracts a set of all uses of *done* from the ACTW and isolates uses of perfective *done* from the resulting dataset. I discuss the results of each method, as well as the implications therein and directions for future research. I conclude that while both methods are effective at extracting perfective *done* from the ACTW, the *back-end* approach is better suited to geospatial mapping.

## 4 Methods

### 4.1 The front-end approach

To extract data from the ACTW, one must first tell the ACTW exactly what items to extract in the first place. However, as mentioned above, if the ACTW is told simply to extract uses of the word *done*, it will extract a sample of all uses of *done* within the requested time frame (2012–2015) with no distinction between tense, aspect, or other grammatical forms. The resulting dataset would be a total of some 8.5 million tweets taken from approximately 1.6 million individual users. Where Willis (2020) was able to eliminate irrelevant users and their respective tweets manually from a relatively small dataset, this becomes unfeasible with such a relatively high number of tweets. In addition, existing part of speech taggers often do not tag AAE parts of speech accurately (Jørgensen et al. 2015, Blodgett et al. 2016). This issue is further complicated by the fact that the ACTW also does not adhere to punctuation in the separation of multiple utterances within the same tweet. Add to this the fact that perfective *done* is orthographically identical to other *done* forms, and it becomes clear that alternate methods must be used to extract perfective *done* in a way that includes as few “don’t count cases” (Blake 1997) as possible.

For the *front-end* approach, which aims to extract a perfective *done* dataset directly from the ACTW, I first establish the syntactic parameters within which perfective *done* occurs and, most importantly, the parameters through which perfective *done* does not occur. I include a summary of “count” and “don’t-count”

cases below to illustrate which orthographical instances of *done* were included and which ones were not. See Table 1.

Table 1: A list of *done* uses, their properties, and examples of each alongside their status as included (yes) or not included (no)

Syntactic Parameter	Example	Perfective or Verbal	Included?
a. <i>done</i> + participle	“He <i>done</i> gone already.”	Perfective	Yes
b. <i>done</i> + simple past	“He <i>done</i> went already.”	Perfective	Yes
c. <i>be</i> + <i>done</i>	“By the time they finish he(‘ll) be <i>done</i> left already.”	Perfective (future)	No
d. conjugated <i>be</i> + <i>done</i>	<i>is done</i> , <i>are done</i> , <i>am done</i> , <i>was done</i> , <i>etc.</i>	Verbal* (see: contracted copula)	No
e. <i>done</i> + progressive	“She’s <i>done</i> working.”	Adjective	No
f. <i>done</i> + preposition	“Wait till the work is <i>done</i> to start packing everything up.”	Adjective	No
g. phrase-final <i>done</i>	“The laundry is done. Sam will take care of the rest.”	Adjective	No
h. <i>done</i> + contracted copula	“I’m done lost my mind!”	Perfective	No

#### 4.1.1 Simple past and participle

As mentioned above, most cases of perfective *done* are *done* + simple past, with *-ed* endings or *done* + participles (Green 2002, Martin 2018). Most importantly, when plugged into the ACTW, these forms yield the greatest number of positive matches to the desired syntactic variable.

#### 4.1.2 Be done

While *be done* is a form of perfective *done*, preliminary searches revealed a high number of false positives among the results, as seen below in Example (15):

- (15) a. I'll be done in a minute.  
 b. By then the job will be done. Left the door open for you.

Example (15a) is an example of a false positive garnered by searching *be done* only. Example (15b) is an example of a false positive garnered by searching *be done left*. While (15a) may be resolved by adding a past tense verb to the *be done* construction, (15b) shows an example of a false positive which appears even with the addition of a past tense verb. This is because the ACTW ignores punctuation in its searches. Because of the high numbers of false positives among the *be done* dataset, I exclude it from this study.

#### 4.1.3 Conjugated copula and auxiliary *be*

Perfective *done* does not typically occur with full forms of conjugated copula and auxiliary *be*, as seen in (16a). This is not to be confused with the contracted copula, as seen in (16c).

- (16) a. \* I am done left already.  
 b. ✓ I done left already.  
 c. ✓ I'm done left already.

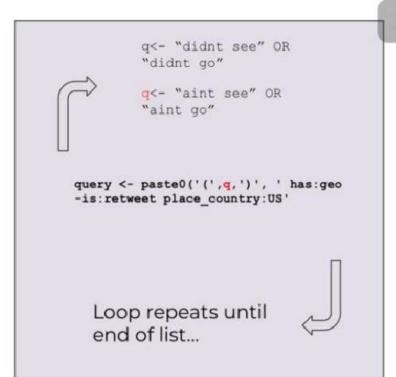
The other cases listed above in Table 1 are representative of verbal or adjectival forms of *done*. These items were eliminated because they do not fit the syntactic parameters of perfective *done*. The search is conducted in a very similar way to the way in which the search for *ain't-for-didn't* (Baxter & Stevenson 2025 [this volume]) is conducted, as shown below in Figure 1.

Figure 1 is an image of the for loop used to collect data from the ACTW for *ain't-for-didn't*. In the case of perfective *done*, I search for *done* + a list of approximately 150 commonly used verbs in their simple past and participle forms, while excluding the aforementioned “don't count” (Blake 1997) cases. The resulting dataset extracted directly from ACTW was approximately 526,000 tweets.

I then examined the resulting data to check for FPs. I then arranged the tweets by “place name,” according to the metadata within each tweet. The place name is represented as “City, State” or “City, Territory” within the metadata, and represents the geographical location where the user was when they published the tweet. I first examine the output of a large city with between 5 and 10 thousand tweets (such as Charlotte) and then I document the “don't count” and other cases therein which show high numbers of FPs. Some examples of constructions with a with high numbers of FPs are listed below. Queries are italicized, FPs are marked with a star\*, and examples of perfective *done* are marked with a check ✓.

## Methods:

- 1) First, we compiled ~150 commonly used English verbs
- 2) With these verbs, we generate query strings with *ain't* + *infinitive* and *didn't* + *infinitive* formations. Each string is stored in a list.
  - "aint see", "aint do", "aint go", etc.
  - "didnt see", "didnt do", "didnt go", etc.
- 3) We then searched the Academic Twitter API sequentially through the list of generated strings, via a 'for loop', specifying 'no retweets' and only tweets with geo-metadata.



15

Figure 1: Baxter & Stevenson (2025 [this volume]), slide 15; screenshot of loop for collecting tweets from the ACTW

- (17) *I've done drunk*
  - a. \* I climbed onto a roof last night. Probably the #1 unsafe thing I've done drunk
  - b. ✓ I've done drunk my problems away before
- (18) *done done*
  - a. \* Done Donee done done done done done done.
  - b. ✓ I done done a lot in my 21.9 years of my life
- (19) *done did*
  - a. \* Done done done. Did I mention I was done?
  - b. ✓ And I done did everything but trust these hoes

The most common FPs included *be done*, *done done*, *have done*, contracted-copula *done*, and *done*-adverb constructions, including temporal adverbs such as *just*, *already*, and *yesterday*. All of these can be used in perfective *done* constructions (see Table 1); however, as mentioned above, due to the large number of FPs within the query results and/or the relative rarity of these cases among the data, they were eliminated.

After eliminating strings which were most likely to lead to FPs, 426,000 tweets of perfective *done* remained. This set of tweets will hereafter be called the perfective *done* dataset.

I calculate an index of perfective *done* use in a similar way to the “straight” model outlined in (Rickford et al. 1991). As a result, it is necessary to find something to which perfective *done* may be compared. As stated above, while perfective *done* can be compared to the MAE perfect (*has/had/have*), it is not identical. In addition, because it is not identical to the MAE perfect and because of the comparative robustness of the verbal TMA system in AAE, there is no single one-to-one yardstick to which perfective *done* may be compared. In an attempt to solve this problem, I choose a calculation in which the *perfective done dataset* is divided by all uses of *done* taken from the same time period (2012–2015); this sample of 8.5 million tweets will hereafter be called the *all done dataset*.

Both the *perfective done dataset* and the *all done dataset* are cleaned in several ways. First, all duplicate tweets are removed from each file to eliminate mass-duplicate tweets containing song lyrics and popular turns of phrase which may not be used in everyday speech. In addition, all utterances with three or more consecutive instances of *done* are removed to eliminate utterances like the one in Example (18a), which are not examples of perfective *done*. Each dataset is then reduced to one tweet per user. This helps to eliminate the misrepresentation of tweets by individual users who tweet more heavily than others. I choose this method rather than the usual method of gathering proportions for each individual because this study frames perfective *done* use in terms of how many twitter users in each location use perfective *done*, rather than how many times each user uses perfective *done*, or what proportion of use each user produces. Limiting tweets to one per user is a clear and straight-forward way to figure out how many twitter users produced the token within a given area.

Once each dataset is reduced to one-per-author, I extract the latitude, longitude, place name, and user id, and calculate a per-city count based on the number of tweets occurring within each location. For the *perfective done dataset*, this count will be called the *perfective done count*, and for the *all done dataset*, this column will be called the *all done count*.

I divide the perfective *done* count by the *all done* count, and the result is a series of indices reflective of the number of users who have used perfective *done* with the selected 150 words in their tweets.

These indices (marked “newindex” on the right side of Figure 2) are mapped according to their geotag and assigned a color gradient, from dark blue for low indices and yellow for high indices. The resulting map is pictured below.

While the *front-end* approach is effective for extracting perfective *done* directly from the ACTW, it is not suitable for calculating indices of use for the following reasons:

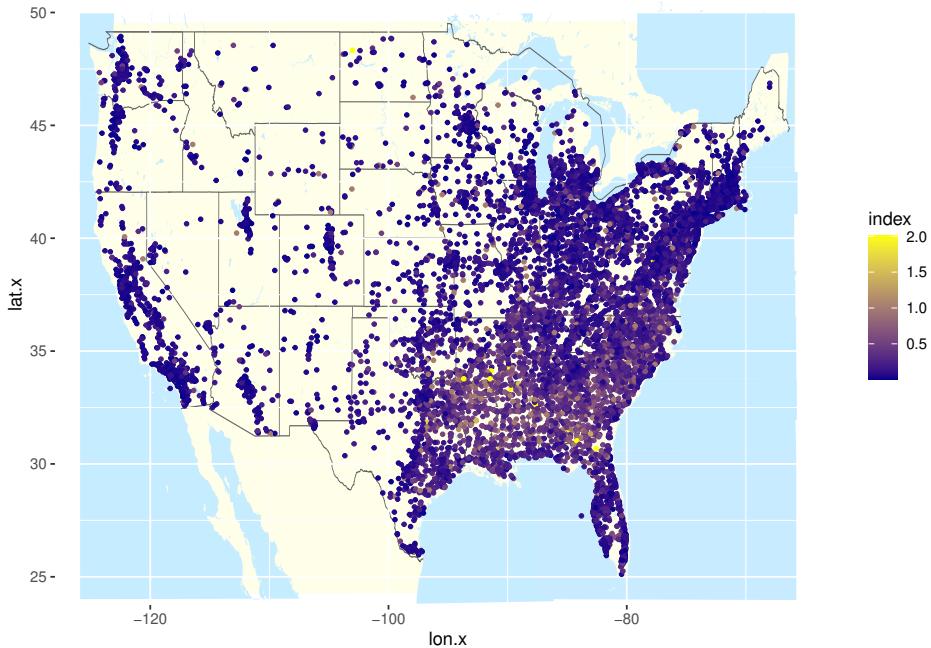


Figure 2: Front-End perfective *done* map

- I. The data extracted from the ACTW is not an exhaustive account of all tweets from that time period (<https://developer.twitter.com>). Rather, it is a collection of data from a given time period which, while still a good sample of Twitter use, may still include or exclude tweets from the *all done dataset*, which was pulled from the ACTW separately, and is thus a separate sample. As a result, indices of certain towns and cities, especially those with a total of 30 user IDs or less, exhibit indices larger than 1. This should not be possible, since all uses of perfective *done* should occur within the *all done dataset*.
- II. In restricting the *all done dataset* to one tweet per user ID, I did not account for the fact that the *all done dataset* included *perfective done dataset* tweets among its 8.5 million tweets. If a user appearing in the *all done dataset* used both perfective *done* and verbal or other forms of *done*, there is no way of knowing which tweet R would use to represent that user. As a result, the resulting 1.6 million tweets, restricted to one per user may not be representative of the *unrestricted all done dataset*.

III. When comparing the *perfective done dataset* to the *all done dataset*, it quickly became apparent that the *perfective done dataset* was missing a great many instances of perfective *done* that still appeared in the *all done dataset*. This is because this method specifically searched for *done* + ~150 commonly used verbs. Those instances of *done* were coded as perfective *done*, while all instances of *done* + other verbs were not coded and therefore not included in the *perfective done dataset*, or coded as perfective *done* within the *all done dataset*.

As a result of these and other problems, I employ a different strategy to formulate the *perfective done dataset* and calculate indices of perfective *done* use.

## 4.2 The back-end approach

Having already formulated the *all done dataset* by extracting all uses of *done* from the ACTW, I extract a new *perfective done dataset* from the existing *all done dataset*, rather than extracting the *perfective done dataset* directly from the ACTW. This immediately remedies the issue of mismatching Twitter samples and by proxy eliminates the occurrence of indices greater than 1.

To do so, I create a list of all lemmas following *done* among the 8.5 million tweets in the *all done dataset*.

While reading all 8.5 million tweets is unfeasible within the given timespan of this study, there are only about 53,000 lemmas following *done* in all 8.5 million tweets, including all alternate spellings of each word (e.g.: *left*, *leff*, *lef*, etc.). I code all ~53,000 lemmas manually and mark simple past forms, participle forms, and any alternate spellings therein with an *x*. This takes roughly 18 hours, but is an incredibly important step in the creation of a database of alternate spellings of past-tense verbs, many of which may not have been predicted by simply guessing what ways users can and cannot spell things.

In addition to the “don’t count” (Blake 1997) cases above, I also do not mark *un*-verbs, and other lemmas which can also be interpreted as adjectives because these are also likely to be FPs (e.g.: adjectives such as *undone*, *unloved*, etc.) I then code the resulting ~3,900 marked tokens as perfective *done* and all unmarked tokens as *other*. I then extract all perfective *done* coded tweets to create the new *perfective done dataset*, and extract all *other* coded tweets to a new dataset called the *other dataset*.

Similarly to the *front-end* method above, I restrict each file to one per user, and count the number of instances by city. I then calculate indices with the equation in example (20).

12344	34128	porterhouse	4 other
12345	34140	portraits	4 other
12346	34148	pose	4 other
12347	34149	posed	4 Pdone
12348	34162	posse	4 other
12349	34165	posessed	4 Pdone

Figure 3: Sample image of perfective *done* documentation

(20)

$$\text{perfective } done \text{ index} = \frac{\text{perfective } done \text{ count}}{\text{perfective } done \text{ count} + \text{other count}}$$

This new calculation of the perfective *done* index addresses the user representation problem in the *front-end* method, as speakers who use both perfective *done* and other forms of *done* now have both uses adequately represented in the denominator. Once the indices are calculated, I use geospatial mapping tools to map these indices across the contiguous United States as shown below in Section 5.

## 5 Results and analysis

Figure 4 is the map resulting from the *back-end* approach mentioned above in Section 4.2. Bright (yellow) dots are representative of high perfective *done* use, while dark (blue) dots are representative of low perfective *done* use. A cluster of bright dots can be seen in the southeastern United States, which indicates higher indices of use in this region than in northern, or midwestern regions of the United States. Based on this, the data shows that perfective *done* is more commonly used among Twitter users in the southeastern United States than in other regions.

To confirm these results, I cross-reference these perfective *done* indices with BAA population rates in US Census Tracts in a selection of states representative of each region: Pennsylvania, New York, Illinois, Alabama, and Georgia. Each scatterplot is cropped down to their maximum perfective *done* and BAA population indices. Where most states have some BAA populations that approach 100%, thereby necessitating the need for the full 1.0 on the y-axis, the maximum indices on the x-axis vary widely. For example in Figure 7 (Georgia), the maximum perfective *done* index is approximately 0.7, whereas in Figure 6 (New York), the maximum perfective *done* index is 0.3. This is done for the sake of visibility in

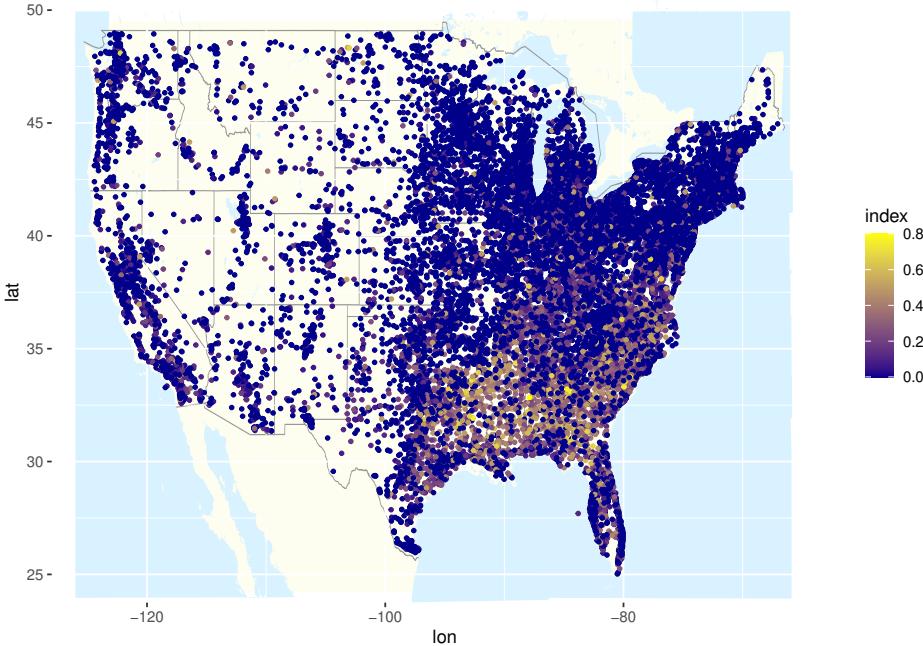


Figure 4: Perfective *done* use in the contiguous United States

states where perfective *done* indices are relatively low, so that the distribution of datapoints therein are still visible and interpretable.

Figure 5 is a scatter plot depicting a comparison of the BAA index: the percentage of BAA-identifying people living in a given area according to the US Census, to the perfective *done* index: the percentage of users with perfective *done* in their tweets. In Pennsylvania, among areas with high populations of BAA-identifying people, indices are relatively low, falling entirely below 0.2.

Figure 6 shows a similar trend to the one shown in Figure 5, with the maximum perfective *done* index at just over 0.3 for any recorded locations in the state. Similarly to Pennsylvania, locations in New York with high populations of BAA-identifying people mostly show perfective *done* indices of 0.2 or less, although very few are just above 0.2. Most notably, New York appears to show many locations with very low perfective *done* indices. Many of these are flatly at the 0.00 mark.

Figure 7 is a scatter plot from Illinois, which shows similar features to previous northern states. Similarly to New York, the maximum perfective *done* index is 0.4. However, unlike New York, Illinois does not appear to show many locations with

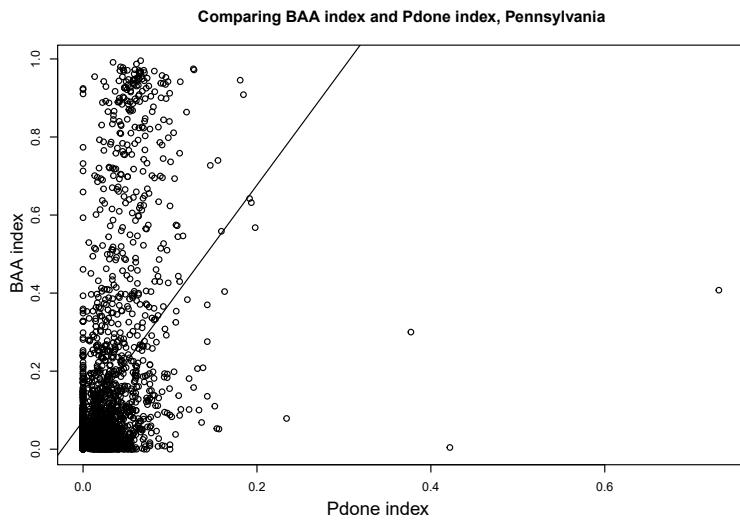


Figure 5: A comparison of the BAA index (vertical; BAA Population index) and the perfective *done* index (horizontal; Index of Perfective *done* use (Pennsylvania; PDONE MAX VALUE = approx. 0.8).

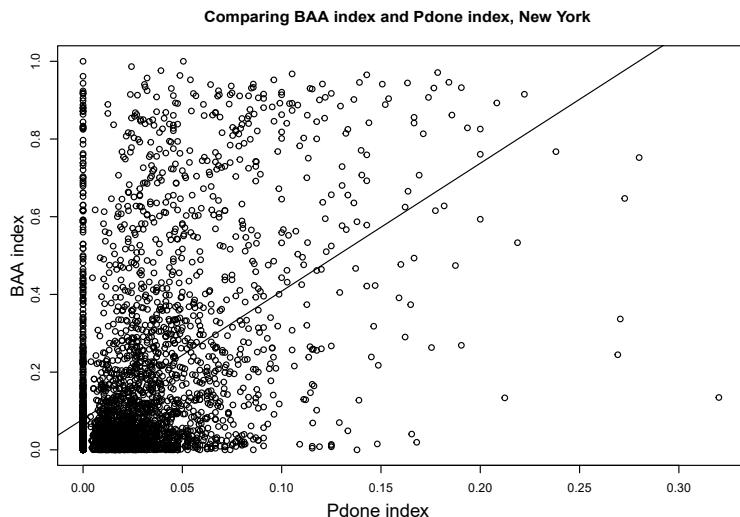


Figure 6: A comparison of the BAA index (vertical; BAA Population index) and the perfective *done* index (horizontal; Index of Perfective *done* use. (New York; PDONE MAX VALUE = approx. 0.35)

zero ratings. Rather, similarly to Pennsylvania, many locations with high BAA populations appear to be spread between the 0.0 and 0.2 mark, with several more between the 0.2–0.3 marks. While Illinois is a northern state, it is also located in the Midwest. The aforementioned divergence from northeastern states like New York whereby relatively few high-density BAA census tracts in Illinois show a 0.0 perfective *done* index may be indicative of a broader trend in the Midwest. Conversely, it is also possible that New York is unique in its apparent rejection of perfective *done* among Twitter users.

Figure 8 is a scatter plot of data taken from North Carolina. Where many locations with high BAA populations in Northern states appear to max out at between 0.2 or 0.3 perfective *done* index, high BAA populations in North Carolina are mostly between 0.1 and 0.4, with none at the 0.00 mark.

Figure 9 is a scatter plot of data taken from Georgia. Similarly to North Carolina, while there is one location with a high BAA population with a perfective *done* index approaching 0.00, most others fall above 0.1, and show much wider distribution than Northern states, with indices ranging up to 0.5, which is higher than the maximum perfective *done* index for all Northern states mentioned above except Pennsylvania, which has one outlier above 0.6.

Figure 10 is a scatter plot of data taken from Alabama. Similarly to Georgia and North Carolina, while there are two high BAA population locations which fall below the 0.1 mark, most others are diffused along a wider range from 0.1 to 0.5.

Scatterplots from northern states in the eastern and midwestern United States show a greater number of areas with high density populations of BAA people with lower indices of perfective *done*. This indicates a lower rate of use, on average, of perfective *done* among Twitter users in these regions. In contrast, scatterplots from Southeastern states show very few locations with high-density populations of BAA people who do not produce perfective *done*. As indicated in Figures 8–10, maximum indices are higher, and the top left corner is empty or mostly empty, which indicates a lack of locations with high-density BAA populations and low perfective *done* indices. Scatterplots of data taken from southeastern states also shows a wider distribution of perfective *done* than northern cities.

These results confirm the hypothesis that regional differences in indices of perfective *done* use will be reflected in the geospatial data, as shown in maps produced by both the *front-end* and *back-end* methods, as well as the scatterplots above. This also confirms the hypothesis that use of the perfect marker *done* is more concentrated in the southeastern states than in northern states.

Additionally, variation in AAE can be mapped geospatially via the geolocation data included in tweets mined from the ACTW.

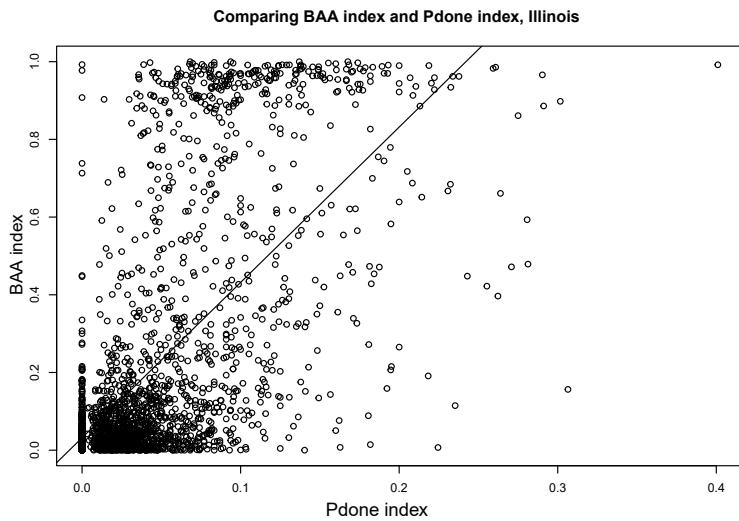


Figure 7: A comparison of the BAA index (vertical; BAA Population index) and the perfective *done* index (horizontal; Index of perfective *done* use). (Illinois; PDONE MAX = approx. 0.4)

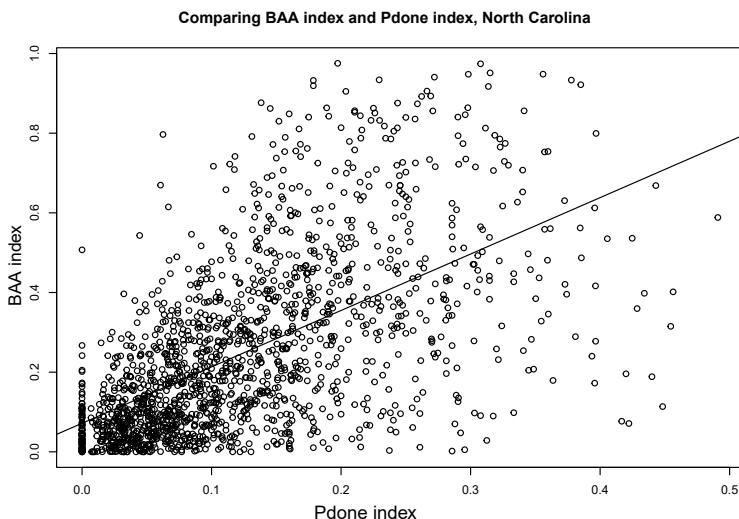


Figure 8: A comparison of the BAA index (vertical; BAA Population index) and the perfective *done* index (horizontal; Index of Perfective *done* use). (North Carolina, PDONE MAX VALUE = approx. 0.5)

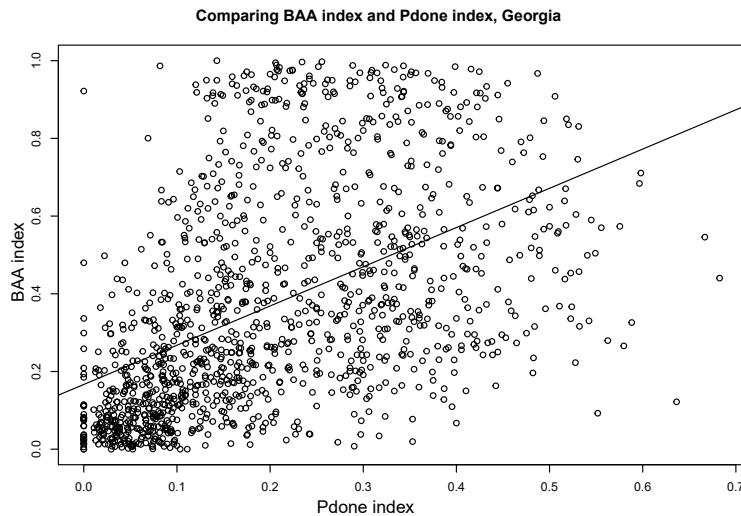


Figure 9: A comparison of the BAA index (vertical; BAA Population index) and the perfective *done* index (horizontal; Index of perfective *done* use). (Georgia; PDONE MAX VALUE = approx. 0.7)

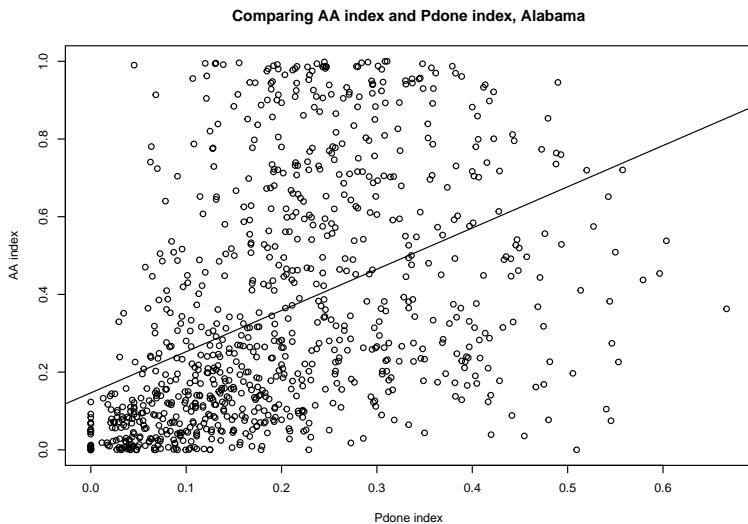


Figure 10: A comparison of the BAA index (vertical; BAA Population index) and the perfective *done* index (horizontal; Index of Perfective *done* use). (Alabama; PDONE MAX = approx. 0.7)

## 6 Conclusion

This paper discusses the extraction of perfective *done* and the geospatial mapping of indices of perfective *done* use via the location data attached to each tweet. I discuss two methods of extracting perfective *done* from the ACTW: a *front-end* approach which aims to isolate uses of perfective *done* by eliminating non-perfective uses of *done* from the search prior to running the query, and a *back-end* approach which first extracts a set of all uses of *done* from 2012–2015 and aims to isolate uses of perfective *done* afterwards. I conclude that while both methods are effective at extracting perfective *done* from the ACTW, the *back-end* approach is better suited to geospatial mapping.

The resulting map shows a cluster of high indices of use in the southeastern United States, which indicates a higher concentration of perfective *done* use among Twitter users therein. Indices of use of perfective *done* were then cross-referenced with demographic data from the US Census. The resulting scatter plots showed that overall, Census tracts with high-density populations of BAA people in New York and Pennsylvania had much lower indices of perfective *done* use than similar populations in Alabama, Georgia, and North Carolina. Not only does this challenge previous assumptions of a “uniform” AAE in urban centers across the United States, it indicates a likely correlation between region and perfective *done* use, with northeastern states showing lower indices overall, and southeastern states showing higher indices overall. Illinois presents an interesting case wherein indices of perfective *done* use in high-density BAA populations are slightly higher than northeastern states, but still lower than southeastern states. It may be that this middle-ground result is indicative of data taken from midwestern states – however, further study must be conducted regarding indices of use in other midwestern states.

Finally, I have shown that a language grouping approach is an appropriate method in the case of anonymized social media data which presents a variety of challenges in identifying and classifying the ethnicities of users. By using this approach and using demographic data taken from the US Census, I provide a broad yet comprehensive view of perfective *done* usage rates in high-density BAA communities across the United States.

## 7 Future research

In the future, I plan to add additional methods to this study to further legitimize the use of social media data and Census data to describe patterns in AAE on such a wide scale. I outline the next steps in this research below:

1. Investigate the new demographics of X (formerly Twitter), and conduct a deeper examination of how recent leadership, moderation, and other administrative changes have affected Black Twitter. Has the hostile environment outlined in Dwoskin (2023) changed the way language is used among those who continue to use X?
2. Revisit the distribution of perfective *done* in northern states and cities. Do BAA populations in New York, Pennsylvania, and other northeastern or mid-Atlantic states all show low indices of perfective *done* use? Are there pockets of high use in any of these states?
3. Further examine perfective *done* use in midwestern states.
4. Expand methodology to include sociolinguistic surveys and interviews, which are better suited toward topics around sociological grouping, such as racial and ethnic variation in AAE use among L1 speakers.
5. Investigate semantic and pragmatic variation in how perfective *done* is used by the people who use it. While the categories in Comrie (1976) fell short of encapsulating all potential uses of perfective *done* in AAE, investigating distinctions between different types of perfective *done* use is still valuable to our knowledge of how AAE works, how AAE is used, and by whom.
6. Investigate other parts of AAE speech to further deepen our understanding of language variation and change therein.

## Abbreviations

ADL	Academic Developer License
ACTW	Academic Twitter API
AAE	African American English
BAA	Black/African American
FPs	False Positives
MAE	Mainstream American English
PDONE	Perfective <i>done</i>
PERF	Perfect marker

## References

- Baugh, John. 1983. *Black street speech: Its history, structure, and survival*. Austin: University of Texas Press.
- Baxter, Kimberley. 2025. Extracting “non-standard” data from the Twitter API. In Susanne Wagner & Ulrike Stange-Hundsdörfer (eds.), *(Dia)lects in the 21st century: Selected papers from Methods in Dialectology XVII*, 3–30. Berlin: Language Science Press. DOI: 10.5281/zenodo.15006593.
- Baxter, Kimberley & Jonathan Stevenson. 2025. *Ain’t* + infinitive verb in Black/African American English. In Susanne Wagner & Ulrike Stange-Hundsdörfer (eds.), *(Dia)lects in the 21st century: Selected papers from Methods in Dialectology XVII*, 31–55. Berlin: Language Science Press. DOI: 10.5281/zenodo.15006595.
- Blake, Renée. 1997. Defining the envelope of linguistic variation: The case of “don’t count” forms in the copula analysis of African American vernacular English. *Language Variation and Change* 9(1). 57–79.
- Blake, Renée. 2014. African American and black as demographic codes. *Language and Linguistics Compass* 8(11). 548–563.
- Blodgett, Sue Lin, Lisa J. Green & Brendan O’Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In Jian Su, Kevin Duh & Xavier Carreras (eds.), *Proceedings of the 2016 conference on empirical methods in Natural Language Processing*, 1119–1130. Austin, TX: Association for Computational Linguistics. DOI: 10.18653/v1/D16-1120.
- Comrie, Bernard. 1976. *Aspect: An introduction to the study of verbal aspect and related problems*, vol. 2. Cambridge: Cambridge University Press.
- Dwoskin, Elizabeth. 2023. Fleeing Elon Musk’s “X”: The quest to recreate “Black Twitter”. *The Washington Post* August 6, 2023. <https://www.washingtonpost.com/technology/2023/08/06/musk-black-twitter-spill/>.
- Green, Lisa J. 2002. *African American English: A linguistic introduction*. Cambridge: Cambridge University Press.
- Green, Lisa J. 2010. *Language and the African American child*. Cambridge: Cambridge University Press.
- Green, Lisa J. & Thomas Roeper. 2007. The acquisition path for tense-aspect. Remote past and habitual in child African American English. *Language Acquisition* 14(3). 269–313.
- Grieve, Jack. 2016. *Regional variation in written American English*. Cambridge: Cambridge University Press.
- Horvath, Barbara & David Sankoff. 1987. Delimiting the Sydney speech community. *Language in Society* 16(2). 179–204. DOI: 10.1017/S0047404500012252.

- Johnson, Sasha Rosena. 2008. *Acknowledging the voices of families: Metadiscourse and linguistic identity of African American speakers of AAE*. University of Georgia. (Doctoral dissertation).
- Jones, Taylor. 2015. Toward a description of African American vernacular English dialect regions using “Black Twitter”. *American Speech* 90. 403–440. DOI: 10.1215/00031283-3442117.
- Jørgensen, Anna, Dirk Hovy & Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In Wei Xu, Bo Han & Alan Ritter (eds.), *Proceedings of the workshop on noisy user-generated text*, 9–18. Beijing, China: Association for Computational Linguistics. DOI: 10.18653/v1/W15-4302.
- King, Sharese. 2020. From African American vernacular English to African American language: Rethinking the study of race and language in African Americans’ speech. *Annual Review of Linguistics* 6. 285–300.
- Labov, William. 1972. Some principles of linguistic methodology. *Language in society* 1(1). 97–120.
- Labov, William. 1998. Coexistent systems in African-American vernacular English. In Salikoko S. Mufwene, John R. Rickford, Guy Bailey & John Baugh (eds.), *African-American English: Structure, history and use*, 154–200. New York: Routledge.
- Lanehart, Sonja (ed.). 2015. *The Oxford handbook of African American language*. New York: Oxford University Press.
- Martin, Katie. 2018. Perfective *done*. In Raffaella Zanuttini, Laurence Horn & Jim Wood (eds.), *Yale grammatical diversity project: English in North America*. New Haven, CT: Yale University. <https://ygdp.yale.edu/phenomena/perfective-done>.
- Masis, Tessa, Anissa Neal, Lisa J. Green & Brendan O’Connor. 2022. Corpus-guided contrast sets for morphosyntactic feature detection in low-resource English varieties. In Oleg Serikov, Ekaterina Voloshina, Anna Postnikova, Elena Klyachko, Ekaterina Neminova, Ekaterina Vylomova, Tatiana Shavrina, Eric Le Ferrand, Valentin Malykh, Francis Tyers, Timofey Arkhangelskiy, Vladislav Mikhailov & Alena Fenogenova (eds.), *Proceedings of the first workshop on NLP applications to field linguistics*, 11–25. Gyeongju, Republic of Korea: International Conference on Computational Linguistics. <https://aclanthology.org/2022.fieldmatters-1.2>.
- Moody, Simanique Davette. 2011. *Language contact and regional variation in African American English: A study of Southeast Georgia*. New York: New York University. (Doctoral dissertation).

- Plank, Barbara, Anders Søgaard & Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th annual meeting of the Association for Computational Linguistics*, 412–418. Berlin, Germany: Association for Computational Linguistics.
- Rickford, John R. 1999. Phonological and grammatical features of African American Vernacular (AAVE). In John Rickford (ed.), *African American vernacular English: Features, evolution, educational implications*, 3–14. Malden, MA: Blackwell.
- Rickford, John R., Arnetha Ball, Renee Blake, Raina Jackson & Nomi Martin. 1991. Rappin on the copula coffin: Theoretical and methodological issues in the analysis of copula variation in African-American Vernacular English. *Language Variation and Change* 3(1). 103–132. DOI: 10.1017/S0954394500000466.
- Ryan, Camille L. & Kurt Bauman. 2016. *Educational attainment in the United States: 2015. Population characteristics. Current population reports* (Report number: P20-578). Suitland, MD: US Census Bureau. <https://www.census.gov/library/publications/2016/demo/p20-578.html>.
- Stevenson, Jonathan. 2016. *Dialect in digitally mediated written interaction: A survey of the geohistorical distribution of the ditransitive in British English using Twitter*. University of York: University of York. (MA thesis).
- Terry, J. Michael. 2010. Variation in the interpretation and use of the African American English preverbal done construction. *American Speech* 85(1). 3–32.
- U. S. Census Bureau. 2018. *2015 annual social and economic supplement: Current population survey*. <https://www.census.gov/data/datasets/2015/demo/cps/cps-asec-2015.html>. Accessed: 18.09.2023.
- Willis, David. 2020. Using social-media data to investigate morphosyntactic variation and dialect syntax in a lesser-used language: Two case studies from Welsh. *Glossa: A journal of general linguistics* 5(1). 103. DOI: 10.5334/gjgl.1073.
- Wilson, August. 1986. *Fences*. New York: Plume.
- Wojcik, Stefan & Adam Hughes. 2019. *Sizing up Twitter users*. Washington, D.C.: PEW Research Center. <https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/>.
- Wolfram, Walt. 2007. Sociolinguistic folklore in the study of African American English. *Language and Linguistic Compass* 1(4). 292–313. DOI: 10.1111/j.1749-818X.2007.00016.x.

# Chapter 2

## ***Ain't + infinitive verb in Black/African American English***

Kimberley Baxter<sup>a</sup> & Jonathan Stevenson<sup>b</sup>

<sup>a</sup>New York University <sup>b</sup>University of York

This study documents the first stage in the creation of an atlas of African American English (AAE) syntax, charting the relative use of *ain't* against *didn't* where they occur with infinitival verbs (*ain't see/ask/buy* etc. vs *didn't see/ask/buy* etc.). The source data is a large, geo-tagged Twitter corpus spanning three years from 2012–2015. The data are plotted for the contiguous United States, but the focus of this paper is on California, Illinois and Georgia. Previous literature (Fisher 2022, Kautzsch 2012, Weldon 1994, Howe 2005, Wolfram & Thomas 2002) notes that the use of *ain't+infinitive* (*ain't+inf*) structures is strongly linked to urban AAE speech communities and that the increased use of *ain't+inf* in rural AAE speech communities is likely due to interaction with urban AAE speech communities. The present paper re-examines the link between *ain't+inf* and AAE speech communities as they appear on social media via the use of Twitter data in a corpus-based analysis of AAE.

We present a method through which *ain't+inf* structures are isolated from other uses of *ain't* (e.g. *ain't-for-isn't*, *ain't+perfective*, etc.) and subsequently compared to *didn't+infinitive* (*didn't+inf*) structures in the same grammatical settings. These results are then compared to demographic information from the US Census and linked to the geographical metadata contained within each tweet.

We find that the strong link between *ain't+inf* and AAE speech communities is mirrored in language use on Twitter for Illinois, confirming results garnered using traditional methods. We find that *ain't+inf* is now prevalent in Georgia, indicating a near complete spread of *ain't+inf* constructions from urban centers in northern areas of the United States to southern areas. Meanwhile, we find that there is little to no correlation between *ain't+inf* use and high-density populations of Black/African American people in California, suggesting that *ain't+inf* has not spread to western coastal areas.



## 1 Introduction

African American English (AAE) is one of the most widely studied varieties of American English. To date, there are numerous sociolinguistic studies that show how AAE syntax varies regionally (Weldon 1994, Moody 2011, Kautzsch 2012, Fisher 2022), thereby challenging earlier myths of AAE as a linguistic monolith (Wolfram 2007). The current paper adds to the growing body of research on syntactic variation in AAE by examining the relative use of *ain't* in conjunction with the infinitival form of a verb (*ain't+inf*) to mainstream *didn't* (*didn't+inf*) (1) on Twitter, across geographical space, focusing on California, Illinois and Georgia.<sup>1</sup>

(1) *ain't+inf*

- a. I ain't see her yet  
'I didn't see her yet.'
- b. She ain't say that  
'She didn't say that.'

The present paper re-examines the link between *ain't+inf* and Black/African-American (BAA) communities as defined by the US Census via the use of geo-tagged Twitter data in a corpus-based analysis of AAE.<sup>2</sup> Widespread use of AAE on archived social media posts creates a living database of timed, dated, and geo-tagged utterances from which corpora may be built.

## 2 On the use of Twitter data

### 2.1 Twitter demographics

As mentioned in Baxter (2025 [this volume]), tracking variation in a sociolect such as AAE is difficult on social media sites which grant anonymity to users. For Twitter, with the exception of verified profiles (indicated with a blue check), confirmation of one's race, ethnicity, gender, or other facets of one's identity are

---

<sup>1</sup>*Ain't+inf* is referred to as *ain't-for-didn't* in (Fisher 2022).

<sup>2</sup>The US Census does not account for differences in ethnicity among Black people in the United States – instead, all Black-identifying people are encompassed within the same singular category, *Black/African-American*. This category only seeks specification regarding whether one is “Black/African-American alone” or “in combination” with other races/ethnicities. This is an issue which affects all studies of AAE which seek to use the US Census as a touchstone for demographic metadata, including the present study. We use *Black/African American* (BAA) when referring to racial and ethnic identity in this chapter.

optional. As a result, a reliable mass verification of users' demographic data is not currently feasible.<sup>3</sup>

Multiple surveys have been conducted in an effort to tease apart the demographics of Twitter users, with varying results. The present study refers to the Pew Research Study "Sizing up Twitter" (Mitchell 2019). This survey is a representative sample of 2791 adult Twitter users surveyed via Ipsos KnowledgePanel, a probability-based online panel of US adults.

The survey found that, at the time of the study, approximately 80% of all tweets were made by about 10% of users. Twitter users tended to be younger, and more likely to be Democrat or otherwise left-leaning, more likely to be women, and more likely to tweet about politics. 11% of Twitter users identified as Black\* (not including Hispanics). Of these, approximately 30% were college graduates, 40% had some college education, and 30% had a high school diploma or had not completed their high school education.

The survey did not include statistics on political affiliation, age, or use by race. However, the study does present a representative Twitter demographic which shows a proportion of Black tweeters which is comparable to the general population of the United States (approximately 13% of the total population of the United States, US Census Bureau 2018). The US Census Bureau recorded 22.5% of the Black population in the United States 25 years and older as having a bachelor's degree in 2015 (Ryan & Bauman 2016). This number is lower than the number in "Sizing Up Twitter" and appears to show that the finding that Twitter users are more likely to be college educated holds among Black users as an isolated demographic.

## 2.2 Twitter location data

A concern frequently raised with the use of Twitter data for linguistic research is that a tweet's geolocation (discussed in more detail in Section 4.3) reflects where the tweeter was located when they sent the tweet rather than where the tweeter is necessarily from.

However, in this research, we do not assume that any individual user is necessarily a long-standing resident of the place from which they tweeted, rather we take the aggregate of a large number of tweets from different users as indicative of the overall picture for that location. This works on the prediction that, for the most part, Twitter users are, in fact, tweeting from the location in which they reside.

---

<sup>3</sup>There have been attempts to estimate user age and gender using crowd-sourcing, see Nguyen et al. (2014), with mixed results.

This prediction is borne out in a number of studies which show that Twitter data tallies closely with data drawn through traditional methods. For example, Stevenson (2016) shows that geolocated tweets for ditransitive verbs with pronominal objects in the UK (*send it me/send me it* etc.) correspond closely to data from the Survey of English Dialects (Orton & Dieth 1962), and Strelluf (2019) shows a similarly close correspondence for positive *anymore* between Twitter data and established distributions.

### 2.3 Summary

With these issues in mind, we would nevertheless expect that if there is an association between *ain't+inf* and communities with high-density BAA populations (Fisher 2022, Labov et al. 1968, Kautzsch 2012, Weldon 1994) then this will also be reflected in Twitter data. Conversely, if there is no such association in the Twitter data for a given location, this would be evidence that either:

- I. AAE language patterns documented in previous research via audio recordings, interviews, and other traditional data collection methods are not readily visible in data taken from Twitter.  
or,  
II. The correlation between *ain't+inf* and high-density BAA communities has diminished.

However, as we will see, it is difficult to explain the systematic distribution of the Twitter data if they are not tallied to actual language use. If there was a fundamental problem with Twitter data, then we would expect to see it across the board, and it would not explain the consistent geographical differences in Twitter use that we find between census tracts and between states.

As will be discussed in the next section, the ethnic identity of a given user is not our first question. Rather, the geographic distribution of a linguistic structure previously associated with the linguistic variety AAE, is the first question. The second question is whether this distribution correlates with the geographic distribution of BAA people in different locations.

The goal here is to both supplement previous research and offer a tool for future research. The hope is that the resulting atlas may be used to find areas that warrant further investigation via traditional, on-the-ground methods.

This, we believe, is a significant advancement. Rather than investigating a given location that may arise through convenience, happenstance or personal

connections, the promise here is for an overview that may reveal previously unknown patterns and locations of particular interest. An analogy might be to that of using aerial photography to scan for areas of archeological interest, rather than relying on serendipitous finds.

### 3 Background

Due the relative anonymity of Twitter profiles, which makes it difficult to verify the ethnicities of the users who own them, this study follows a *linguistic grouping* approach (Horvath & Sankoff 1987), which first groups linguistic features associated with a given variety, then looks for correlations with external, social factors. Accordingly, linguistic features of AAE are collected and coded *before* sociological factors such as race and ethnicity, rather than following the traditional *sociological grouping* approach used in most traditional sociolinguistic studies.

The terms social and linguistic grouping do not mean that sociological consideration predominate in one approach and linguistic concerns in the other, but only refer to the temporal order in which they enter into the statistical analysis. (Horvath & Sankoff 1987: 180)

The present study starts by choosing a linguistic variable, in this case *ain't+inf*, and calculating indices of use against mainstream *didn't+inf*, mapped across the contiguous United States via the geolocation metadata attached to each tweet. These indices are then compared to indices of BAA population in US Census tracts across the same geographical area. By doing so, we aim to provide a broad yet comprehensive view of usage rates in high-density populations of BAA people relative to low-density populations.

While *ain't* itself is common across many varieties of English, it is often considered that *ain't+inf* is a distinctive feature in AAE, where it occurs more frequently than in other varieties of English (Fisher 2022, Labov & Harris 1986, Kautzsch 2012). So, for the present study, following the Language-First model, this means *ain't+inf* tokens, established as being part of AAE, are extracted from the Twitter API prior to testing association with demographic data regarding ethnicity.

*Ain't+inf* is thought to have been recently innovated in northern urban centers and spread as a result of language contact with AAE speakers who moved there during the Great Migration (Fisher 2022). For example, previous research (Jørgensen et al. 2015) suggests that there is an increased use of *ain't+inf* both diachronically and over apparent time by comparing early AAE recordings to later

AAE recordings, or by comparing the speech of Northern (Philadelphia) speakers to speakers who had moved to the Northern United States from the Southern United States.

The scale of the data available via Twitter's API allows the relative frequency of *ain't+inf* to be mapped across the United States and, at the same time, to provide unprecedented resolution at the level of small towns and suburbs. Furthermore, in-line with previous studies that use Twitter data for dialect research (Jones 2015a, Stevenson 2016, Willis 2020, Strelluf 2019, 2020), results in many cases appear to follow established dialect "faultlines" (Eisenstein 2013: 1) while also highlighting particular hotspots of use. If *ain't+inf* is unique to AAE, then we would predict that its distribution would correlate with BAA population distribution.

Principally, then, our main question is:

To what extent is the association between *ain't+inf* and the BAA population reflected in data taken from Twitter?

While the broader study covers the entire US, the present chapter focuses most closely on *ain't+inf* use in Illinois, Georgia and California. These states were chosen because they represent three different regions of the United States, with California being on the west coast, Illinois being in the midwest, and Georgia on the southeastern coast, with all three housing cities which land among the top ten cities most densely populated by BAA people (Tamir et al. 2021). In addition, both Illinois and California were destinations to which many BAAs migrated during the Great Migration, and still have a high population density of BAA residents.

Finally, while our focus is on *ain't+inf*, we acknowledge that the next step in evaluating the extent to which this form of *ain't* is unique to AAE within geotagged Twitter data is to see how it patterns relative to other forms that are well attested in other Englishes. We discuss these next steps in Section 8.

## 4 Method

The semantic near-equivalence of the sentences in (1) allows us to consider the two forms as variants of a single variable (Labov et al. 1968, Wolfram & Schilling-Estes 2016, Fisher 2022) whereby the relative frequency of *ain't+inf* may be measured against *didn't+inf* to provide an index of use across and between places, without needing to know the overall corpus size for a given place.

In this way, *didn't+inf* provides a yardstick against which to measure *ain't+inf* usage via the *straight* model seen in Rickford et al. (1991), through which the use

of a given part of speech is divided by all potential outputs within that grammatical space. Where Rickford et al. (1991) presents the *straight* model with reference to the calculation of copula usage and deletion, we present a similar *straight* model for *ain't+inf* which divides the frequency of *ain't+inf* occurrences by the sum of *ain't+inf* and *didn't+inf*. The resulting number is hereafter referred to as the *ain't+inf index*:

$$\text{ain't+inf index} = \frac{\text{ain't+inf}}{(\text{ain't+inf} + \text{didn't+inf})}$$

The next step, then, is to extract instances of *ain't+inf* and *didn't+inf* from the Twitter API.

#### 4.1 Extracting data using the Academic Twitter API

The Academic Twitter API (ACTW) was used to extract Twitter data. ACTW was made available at the start of 2021 and permits academic access to the entire Twitter archive at no cost.

To extract tweets containing *ain't+inf* and *didn't+inf*, a simple script was written in R to generate a list of strings combining an infinitival verb, from a list of 150 common verbs, with either *ain't* or *didn't* – both with, or without the apostrophe (*aint*, *ain't*, *didnt* and *didn't*).<sup>4</sup> The result was a list of 600 strings (4 × 150):

```
aint see
ain't see
didnt see
didn't see
aint ask
ain't ask
didn't ask
didnt ask
```

The resulting list was then used to generate a search query formatted for the Twitter API that could be used to pull tweets for each item in the list. In the example below, *q* represents a given string combination for that stage of

---

<sup>4</sup>While it would be search for all possible infinitive verbs occurring with *ain't* or *didn't*, using the top 150 most common verbs would capture the vast majority of cases. It is assumed that the inclusion of less frequently occurring verbs would not significantly sway the results, in aggregate.

a *for loop*.<sup>5</sup> The search query also disallowed retweets (-is:retweet), and only tweets that contained geolocation metadata (has.geo). In sum, Twitter is searched for each string in turn until all strings have been searched. Using this method, approximately 3.2 million geolocated tweets were collected, each containing a variation of either *ain't+inf* or *didn't+inf*.

```
query <- paste0('(', q, ')', ' has:geo -is:retweet place_country:US'
```

## 4.2 Data cleaning

The first step in cleaning the data was to remove results where punctuation intervenes between *ain't* and the verb. Twitter's search engine is blind to punctuation, so the raw data will include strings such as in (2).

- (2) But he aint. Ask someone else

Removing these was done using a simple regular expression search in R, which is sensitive to punctuation. A table was generated containing all instances of the strings in the initial search coupled with additional coding data such as verb type and whether it was *ain't* or *didn't*. When the Twitter data was matched to an entry in the table, the additional coding data could also be carried over, resulting in a coded dataset.

Next, the data were spot-checked for further false positives. Problematic cases where the infinitive form is homonymous with another part of speech, which can readily occur in same position in the sentence, were investigated. In some cases – such as: *like* as in *she ain't like me*; *love* as in *that aint love* and *fly* as in *she aint fly* – the non-infinitival form was so dominant and removed entirely. This was only necessary for a handful of verbs, however.

## 4.3 Location data

Each tweet comes packaged together with various forms of associated metadata. Of this metadata, there are three main types directly linked to the user and the tweet location.<sup>6</sup>

---

<sup>5</sup>A *for loop* is a function used in computer programming that performs a set of sub-functions a given number of times. In this case, the *for loop* performs a Twitter search for every construction in the given list of constructions.

<sup>6</sup>There are other ways to garner user and tweet location from the content of a user's tweets, or other information that they provide. For example, a user may state in a previous message that they grew up in Chicago or went to school in South Atlanta. This method of gathering additional location data is particularly useful when data is more scarce, such as when studying

First, the GPS points for the actual location that the tweet was sent from. Evidently, this data is only available for tweets sent from mobile phones. Additionally, GPS data is only available for a relatively small fraction of Twitter messages, around 1–3% by most estimates and data is most prevalent prior to 2015 when a setting in the phone application – to add GPS data to each tweet – was opt-out rather than opt-in. GPS data comes as a set of coordinates:

c(-76.0792, 38.566)

Second, there is user entered location data that is free form, and associated with the user profile itself. This data is information provided by the user when they set up their account, but can be changed later. It is not required and can be any kind of free-entered text. This data can be useful, but may be unreliable and may not correspond to a place at all – it may be used to indicate personality type, for example.

“Trill, Texas”  
“VonteWorld///Beast Coast”  
“Moss Bluff, LA”  
“at the bar”  
“INFP”

Third, tweets also contain a place.id, which is a code corresponding to a rectangle defined by four geographic coordinates, corresponding to a place as can be seen in Figure 1 (page 40).

The place to which the ID is assigned is derived by Twitter through a process of *data enrichment* using a combination of fuzzy matching of user entered location and GPS coordinates.

#### 4.4 Twitter atlas, first iteration

It is possible to generate a useful atlas of tweets using the place.id alone, and indeed, the first iteration of the current atlas used this data. Here, the counts for each variant (*ain't+inf* and *didn't+inf* for each place.id were represented as pie charts on an interactive atlas. Pie charts were placed at the center point of the

---

a lesser used language like Welsh (Willis 2020) or to drill more data from a dataset (Gopal et al. 2021). In addition, taking the location of where a user grew up is closer to the methodology employed in traditional dialectology, and may, in some sense, more authentically reflect the language of a given place, though see Stevenson (forthcoming). This method is, however, less straightforward for the current investigation, in the US, where many places share the name, though there are ways to mitigate this issue. For now, we leave this to future study.

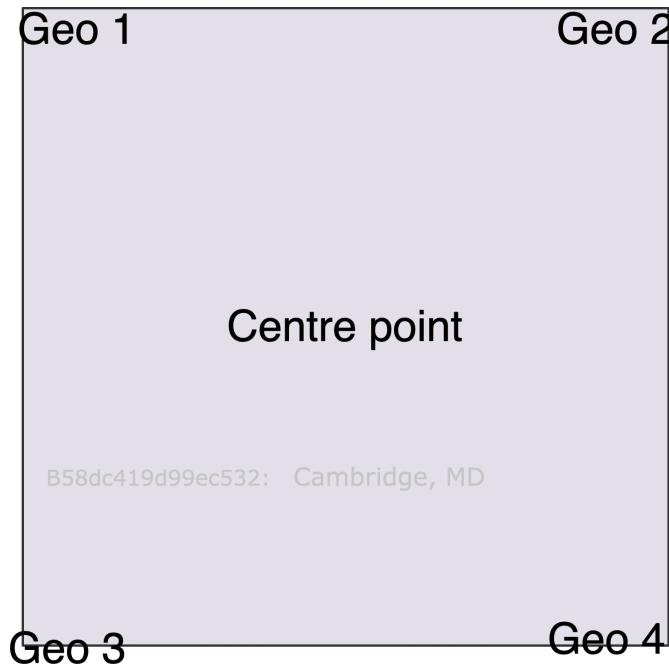


Figure 1: Illustration of Twitter place.id, a square defined by four geocode points, corresponding to a given place.

place.id area (see Figure 1). This was done using the *leafletminicharts* package for R (Bachelier et al. 2021). Using this package it was also possible to add a level of interactivity where the user can click on a pie chart and see a breakdown of the exact counts for that location. The user can then click on a variable and see a sample of the actual data from that place.

This was a valuable tool at this stage in the research process, allowing us to investigate different places, and offering a different way to probe the dataset by place. Figure 2 shows a screenshot of the interactive atlas focussing on the East North Central region of the US, specifically on the Chicago area.

In order to be represented on the map, it was set that a place.id would have to have at least 100 tweets associated with it. This was done to ensure that enough data was present to draw statistically significant conclusions. In addition, this reduced the number of pie charts that had to be rendered on the map, improving legibility and performance (setting to lower than 100 per place meant that scrolling the atlas became too slow).<sup>7</sup>

---

<sup>7</sup>We should note here that the same limited dataset was used for the second iteration of the atlas, which was not necessary for that stage. We do not believe that this will affect the results for the second stage but future versions will have this legacy limit removed.

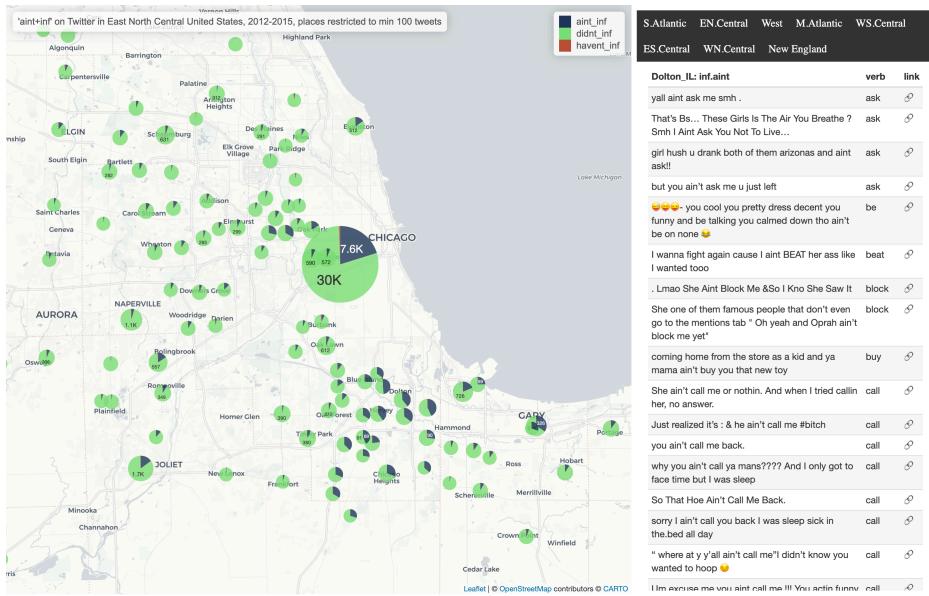


Figure 2: Screenshot of the first iteration of the interactive atlas that uses Twitter place.ids rather than GPS points. Focus here is on the Chicago area. Interactive version available at <http://nwdialectatlas.uk/infaint/>.

It became apparent immediately that there was geographic variation in the use of *ain't+inf* versus *didn't+inf*. Closer examination of places with high *ain't+inf* use showed a pattern: all had high BAA populations. In Figure 2, it is possible to see southern Chicago and Gary with high rates of *ain't+inf*, represented in blue, both areas with high BAA populations, meanwhile northern and western areas have very low rates of *ain't+inf*, which follows if *ain't+inf* is distinctive to AAE.

#### 4.5 Twitter atlas, second iteration: linking to census data

In order to show a systematic link between *ain't+inf* use and BAA population, it is necessary to link Twitter place.id to demographic data. In the US, the most reliable demographic data is provided as part of the census. Census data is organised by tracts, with each defined by geographical polygons, stored as shapefiles. Each tract is associated with demographic information, such as race, about that geographic location. Tracts are defined for the purpose of administering the census, sometimes following political boundaries. They are the smallest geographic areas that are recorded with associated metadata.

However, it is not possible to directly link the Twitter place.id areas to tracts. Sometimes a place.id will encompass numerous tracts and other times, a census tract will contain several place.ids.

So, for the purposes of the present investigation, where the aim is to measure *ain't+inf* against demographic data, it is necessary to have Twitter location data that is compatible with the location information provided for the US census. Associating tweets with census tracts was done using the point-in-polygon technique (similar to Blodgett et al. 2016). This technique, illustrated in Figure 3 is a method for calculating whether a given point falls within the bounds of a polygonal shape. In Figure 3, the red points have been found to be within the polygon.

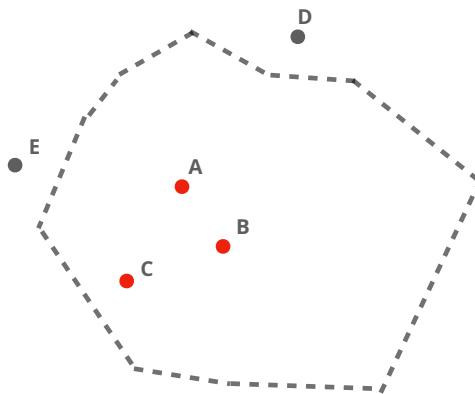


Figure 3: Illustration of point-in-polygon technique. Source: <https://datawanderings.com/2018/09/01/r-point-in-polygon-a-mathematical-cookie-cutter/>

The PinP technique was used to plot from which census tract a given tweet was sent. As we already know the racial demographic data for each tract, then the relative frequency of *didn't+inf* and *ain't+inf* (*ain't+inf index*) could be measured against the relative population of BAA against the total population (*BAA index*).

#### 4.6 Retrieving census data

US census data is freely available from census.gov and accessible directly in R using the *tidycensus* package (Walker & Herman 2023). Tidycensus makes it straightforward to retrieve US Census data that is prepared for use with other R packages in the *tidyverse* suite (Wickham et al. 2019). It also makes it relatively easy to work with shapefiles using the *sf* package. A shapefile is a standard format for storing geographic data as vectors, in terms of points, lines, and polygons.

For our purposes, it is possible to use tract data to calculate an index of the BAA population by dividing BAA population by the total population for each tract.

The index, hereafter referred to as the *BAA index*, is calculated by dividing the BAA population by the total population.

$$\text{BAA index} = \frac{\text{BAA population}}{\text{total population}}$$

This can be seen in the map presented in Figure 4.<sup>8</sup>

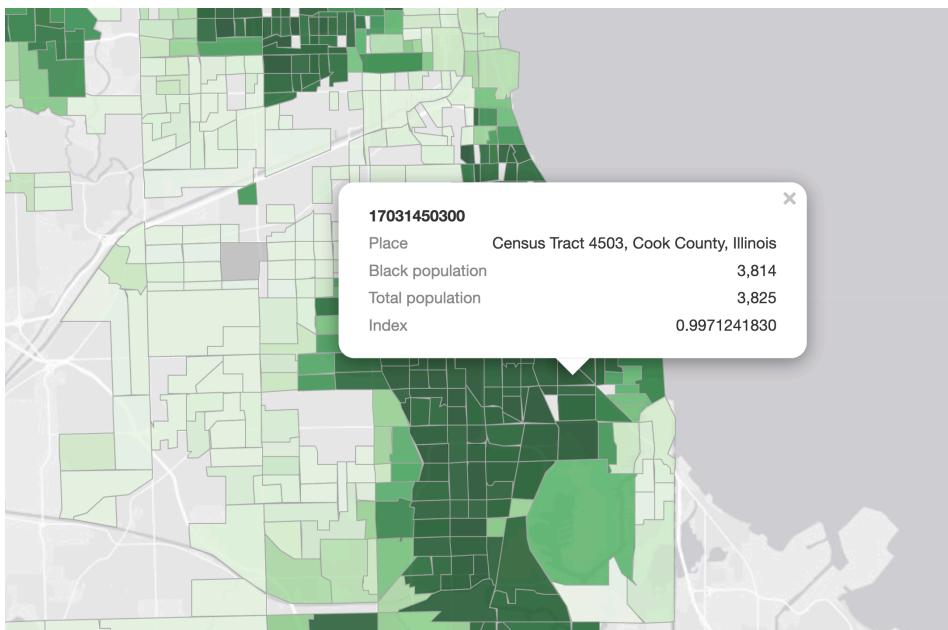


Figure 4: Cook County Census Tracts, based on data from US Census *American Community Survey* 2011–2015 (US Census Bureau 2015)

Figure 4 is a map of Chicago, Illinois. The maps are divided into census tracts, each of which are linked to demographic information taken from the United States Census. The text-box in the center of the map displays information for one of the census tracts within the city: Cook County. This text-box appears as a pop-up when a census tract is clicked by the user.

In the case of Cook County, nearly 99.7% of residents within this census tract are identified as BAA, which is reflected in the darker shade of green in this

<sup>8</sup>An interactive version of the maps presented in this chapter is available at: <http://nwdialectatlas.uk/infaint/>.

and surrounding census tracts. Tracts with low populations of BAA people are represented by lighter shades of green.

#### 4.7 Mapping and plotting Twitter data with census data

First, using the *ain't+inf index* and *BAA index* for each tract, two maps of the US were produced – an *ain't*-usage-census-tract-map, and a racial-demographic-census-tract-map – which could be used for a visual side-by-side comparison. If *ain't+inf* is associated with BAA population groups, then we expect to see a visual similarity in the resulting maps.

Second, scatterplots were produced for each State, plotting *ain't+inf index* against *BAA index* with each point representing a single census tract. If *ain't+inf* is associated with BAA populations, then we expect to see a clear linear correlation between the two indices.

Finally, the data were broken down into high *BAA index* (>90%) and low *BAA index* (<10%) corresponding to high and low *BAA index*. Boxplots could then be produced for each group, with each State represented by one box. Again, if *ain't+inf* is unique to AAE, then we expect to see tracts with a high *BAA index* coalescing (represented by short boxes) at a high rate of *ain't+inf* and, conversely low *BAA index* with low *ain't+inf*.

### 5 Results

Here we present data on the three states under investigation: Illinois, Georgia and California. For each state, we start with heatmaps for Chicago, Atlanta, and Los Angeles and then present scatterplots for each respective state.

We then present a comparison between *ain't+inf* use across 13 states between high BAA population (>90%) and low BAA population (<10%).

#### 5.1 Illinois

The two first maps, presented in Figure 5, compare BAA distribution (left) and *ain't+inf* distribution (right).

As stated, the distribution is represented by the *BAA index*, which is calculated by dividing the BAA population of a given census tract by the total population of that census tract.

From a simple observation of the two maps, it is evident that the prediction that *ain't+inf* is associated with BAA population is borne out. The most concentrated

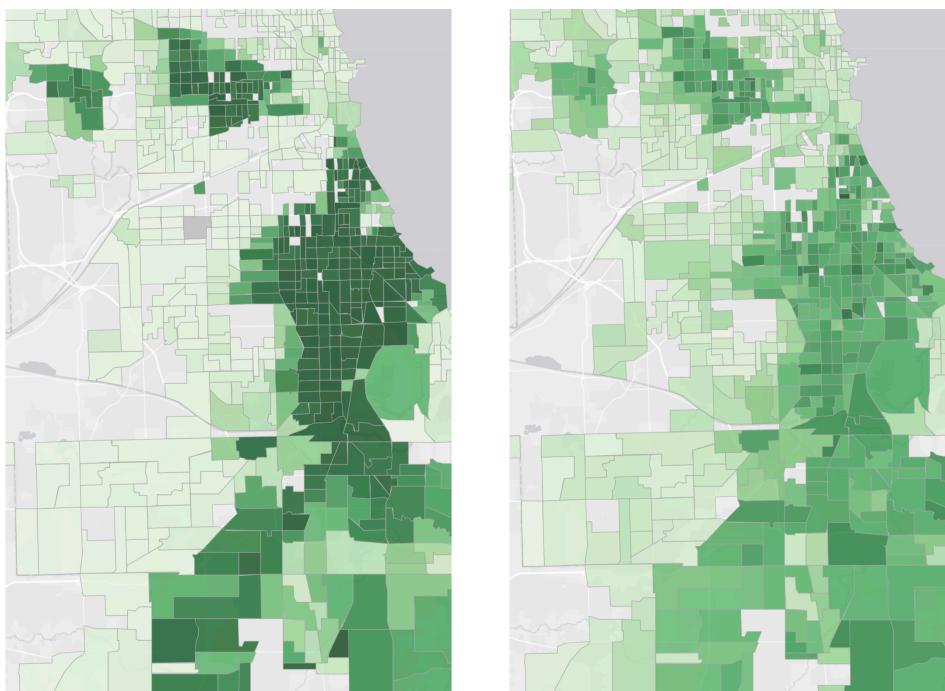


Figure 5: Chicago maps. Left = BAA Population Index, based on data from US Census *American Community Survey 2011–2015* (US Census Bureau 2015); Right = *ain't+inf* index

areas of the *ain't+inf* map are also the most concentrated areas of the *BAA index* map.

This is further clarified in the scatterplot for Illinois, in Figure 6 which shows a clear correlation between *BAA index* and *ain't+inf index*; the higher the BAA population, the higher the *ain't+inf index*. The lower the BAA population, the lower the *ain't+inf index*.

Additionally, the segregation that exists in this part of the country is apparent, with large clusters of census tracts clustering at the top and bottom of the chart corresponding to tracts with nearly 100% of residents identifying as BAA, or nearly 0% BAA, while points in the center of the plot are more sparse, corresponding to there being far fewer tracts with a mixed population.

## 5.2 Georgia

The maps for Atlanta (Figure 7) again show a clear similarity in distribution, comparing *ain't+inf* and BAA populations.

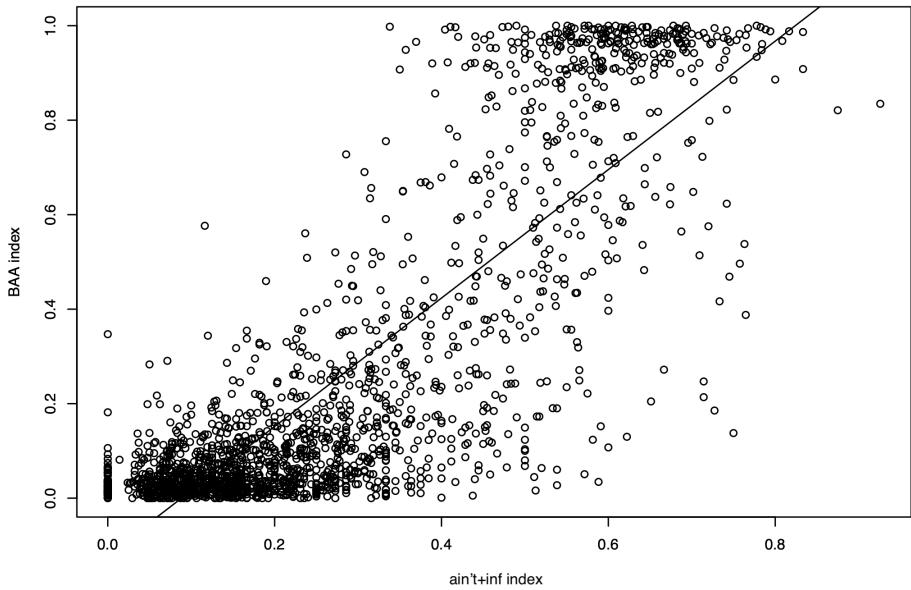


Figure 6: Scatter plot for Illinois: Vertical = BAA Population index; Horizontal =  $ain't+inf$  index

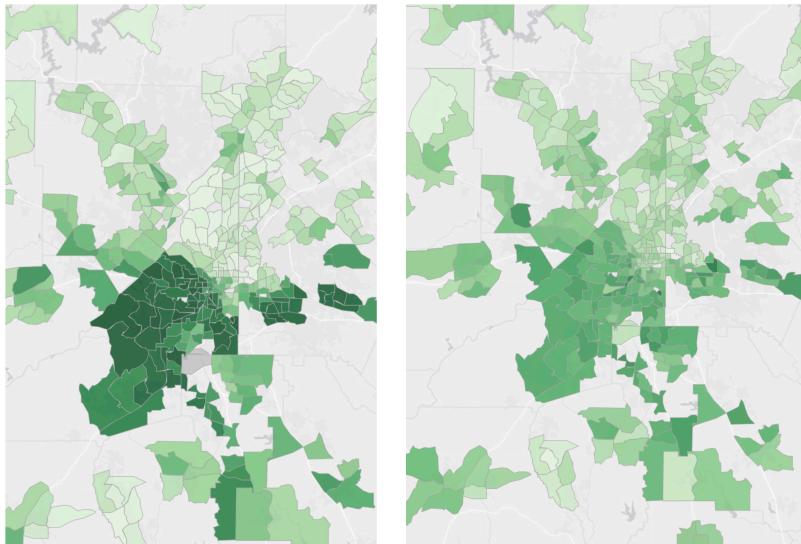


Figure 7: Atlanta maps. Left = BAA Population Index, based on data from US Census *American Community Survey 2011–2015* (US Census Bureau 2015); Right =  $ain't+inf$  index

Meanwhile, the scatterplots for Georgia (Figure 8) contrast with those for Illinois (above) showing a smoother distribution of BAA population, while – crucially – still showing a clear correlation between *BAA index* and *ain't+inf index*.

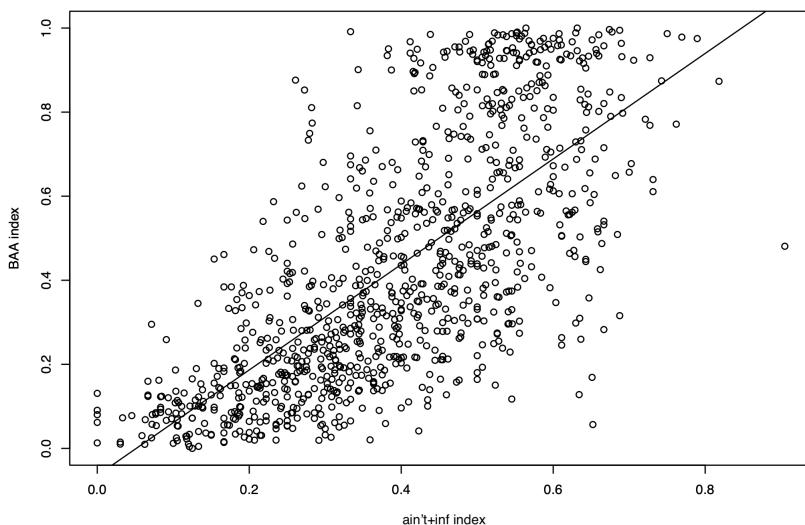


Figure 8: Scatter plot for Georgia: Vertical = BAA Population index; Horizontal = *ain't+inf index*

### 5.3 California

Particularly notable, however, is the absence of this correlation in the maps for Los Angeles, California (Figure 9). While Illinois and Georgia show a very clear correlation between census tracts with high BAA populations and high *ain't+inf* indices, results show that use of *ain't+inf* in California is not only markedly lower, but less concentrated.

The scatter plot in Figure 10 supports this weaker correlation between *BAA index* and the *ain't+inf index* in California. Where the plots of the census tracts for Georgia and Illinois both show clusters of census tracts with a *BAA index* of above 0.8, California shows a much lower number of census tracts which meet the same criteria.

Of the tracts that do show a *BAA index* of above 0.8, none show an *ain't+inf index* above 0.4. This is in contrast to Georgia and Illinois, where most tracts that show a *BAA index* of 0.8 or above also show an *ain't+inf index* of approximately 0.4 and above.

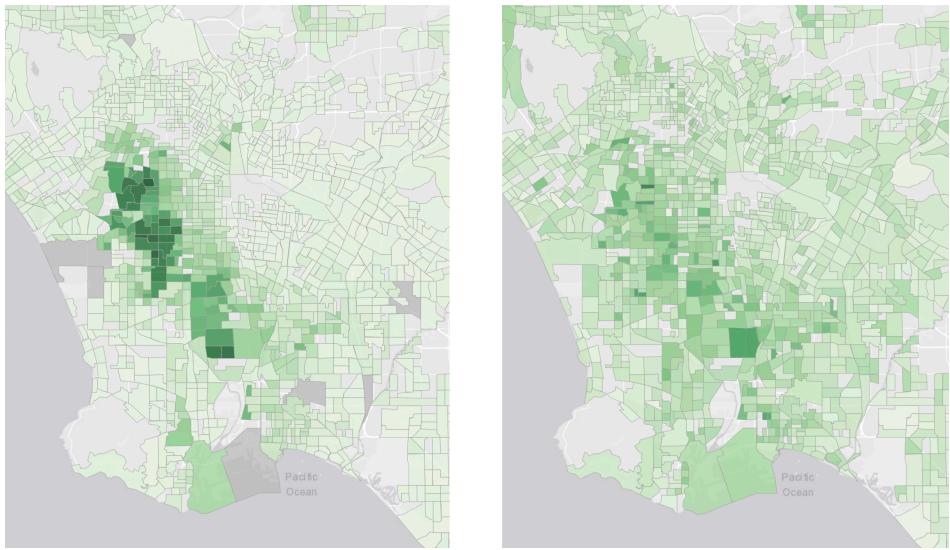


Figure 9: Los Angeles maps. Left = BAA Population Index, based on data from US Census *American Community Survey 2011–2015* (US Census Bureau 2015); Right = *ain't+inf* index

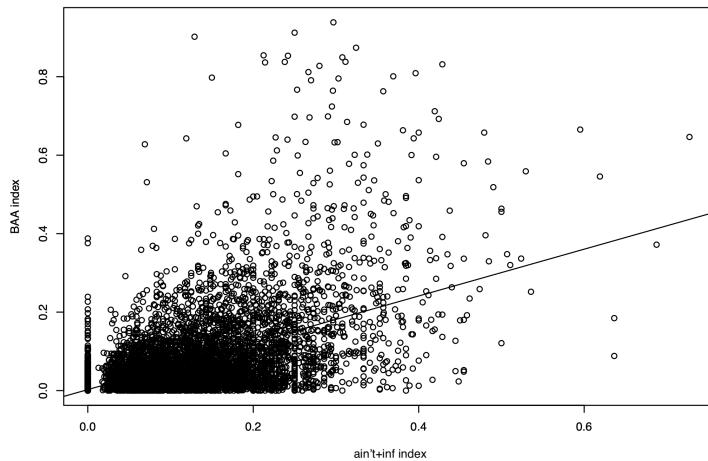


Figure 10: Scatter plot for California: Vertical = BAA Population index; Horizontal = *ain't+inf* index

## 5.4 Comparing states

Figures 11 and 12 are boxplots representing the *ain't+inf* indices across 13 states in census tracts with 10% or less BAA/BAA population and 90% or more BAA/BAA population, respectively, and shows that unlike other states, the *ain't+inf index* in California remains relatively low regardless of *BAA index*.

The boxplots in Figures 11 and 12 further underscore the marked difference in *ain't+inf* use in California, as compared to other states. Both in census tracts with 0.9 *BAA index* or greater and census tracts with 0.1 *BAA index* or less, the average *ain't+inf index* remains relatively low, and does not increase regardless of racial population density. Also notable are *ain't+inf* rates in Alabama (AL), Louisiana (LA), and South Carolina (SC), which all show slightly higher *ain't+inf* indices among census tracts with 0.1 *BAA index*, showing at around 25% usage where other states show >20% usage.

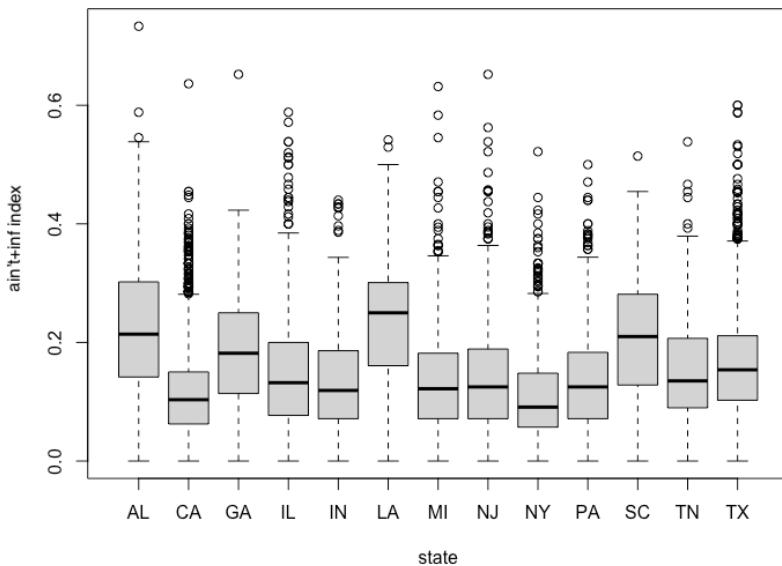


Figure 11: Boxplot showing *ain't+inf index* for tracts with <10% BAA population

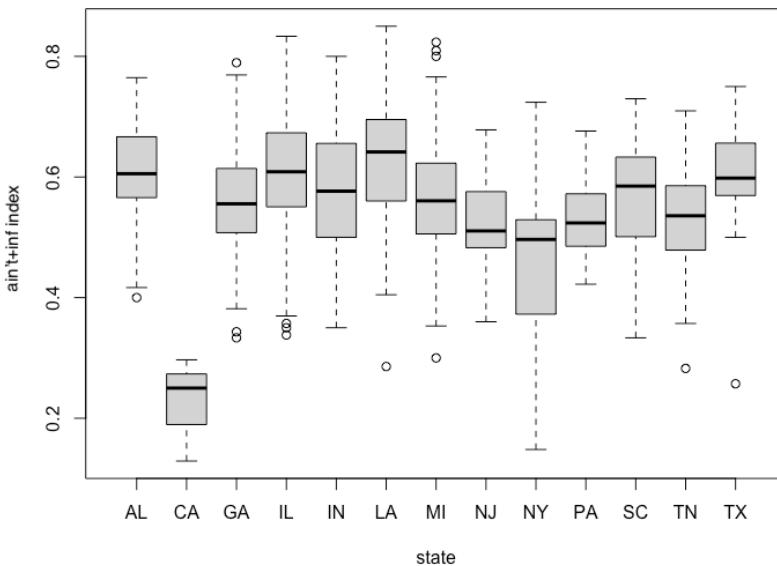


Figure 12: Boxplot showing *ain't+inf* index for tracts with >90% BAA population

## 6 Analysis

The results above confirm the conclusions reached in previous research, which found a correlation between the use of *ain't-for-didn't* constructions and BAA speakers. As a result, they also support previous findings (Eisenstein 2013, Jones 2015b, Willis 2020) that language use on social media platforms such as Twitter patterns with natural language distributions found using traditional methods.

*Ain't+inf* is considered to have been originally relatively rare in early AAE in the Southern US, with the structure having been innovated in the northern cities after the Great Migration. If this is indeed the case, the current distribution shows a continued and near-complete spread of *ain't+inf* within BAA communities in the United States from the northern urban centers to southern urban centers.

This is important to note when we consider Fisher (2022), whose analysis suggested that *ain't+inf* and other forms of *ain't-for-didn't* were used more frequently among speakers who were born and raised in Philadelphia than they were among speakers who had recently migrated from the South. The data also reveal that this spread has not reached as far as California, where *ain't+inf* is shown to be relatively infrequent, and uncorrelated with the distribution of BAA populations.

Why this North-South spread has occurred, and why it has not spread west to California is unclear at this stage, but is a topic ripe for future investigation. A first point of investigation here would be to analyse migration patterns from the northern cities to the south. Cities such as Atlanta have re-emerged in recent years as BAA cultural centers, attracting new waves of BAA migrations. It may also be partly explained by ongoing family and social ties between the north and south that have facilitated the spread of *ain't+inf* south, but not west. It will be interesting to see if *ain't+inf* is part-way through a process of diffusion, and will eventually reach California, or if it will remain a difference between the varieties of AAE in west and east.

## 7 Conclusion

This paper discussed variationist studies in African-American English and aimed to contribute to the study of syntactic variation therein by conducting a study of language data mined from Twitter via the Twitter API.

We presented an analysis which also used Twitter data to confirm analyses reached by more traditional methods in previous literature, thus showing the viability of Twitter data in the study of syntactic variation in AAE. To collect the data, we examined the parameters through which infinitival *ain't* occurred and isolated it from other uses of *ain't* which were more ambiguous.

To map the data, we calculated an index of infinitival *ain't* use and used a point-in-polygon approach to map the geospatial metadata within each tweet within census tracts taken from the US Census. The resulting maps show a correlation between infinitival *ain't* and census tracts with high BAA/African-American populations in Georgia and Illinois, but show a much weaker correlation, as well as a much lower average *ain't+inf index* in California. Data shows that this weaker correlation holds regardless of race.

The results of each map were confirmed by scatter plots and box plots which also showed a correlation between *BAA index* and *ain't+inf index* in communities with high populations of BAA/African-American identifying residents, according to the US Census, with the exception of California.

These results indicate a likely difference in both the frequency and manner of *ain't* usage in the United States, despite previous descriptions of the variety as uniform within urban centers. These results represent a starting point for future researchers to analyse them in more detail via alternate methods such as surveys and/or interviews.

Finally, we have shown how the language-first approach adopted here using Twitter data is compatible with, and reflective of, already-established conclusions gleaned from more traditional identity-first analyses.

## 8 Future directions

As stated at the outset, the work presented in this chapter represents the first step in what is planned to be a much larger atlas of AAE use using both Twitter data and data drawn from traditional methods. With this said, the most pressing next steps are:

1. Investigate the underlying causes for the apparent spread of *ain't+inf* south from the northern cities to those in the south, while not to western coastal areas. Could this be explained by the maintenance of cultural ties (or lack thereof) and migration patterns?
2. Investigate other forms of *ain't*: correlations between other forms of *ain't* and BAA Populations in order to check the hypothesis that Twitter data shows that *ain't+inf* is uniquely tied to AAE. For example, do other uses of *ain't*, particularly *ain't+perfective* (as in *ain't seen*), also show a high correlation with BAA communities?
3. Investigate weak verbs: the extent to which weak verbs (*move, raise*), consonant cluster reduction (etc.) and other phonological properties of AAE influence the orthographical spelling of verbs on Twitter, and could mean some of what appears to be *ain't+inf* is actually perfective *ain't*.
4. Investigate contextual variation: the extent to which contextual variation exists in the use of *ain't+inf* across different regions.
5. Build the Atlas. Begin the process of compiling all of this into an interactive atlas. Combine different visualisation methods (such as pie charts in first map). Research and add other AAE parts of speech.
6. Investigate alternative methods. Use alternative method for getting location data – *I grew up in* rather than GPS point from where a tweet was sent.
7. Investigate outliers among box plots and scatter plots. Identify the locations of outlier census tracts, in order to further investigate the context behind outlier tracts. In the case of the boxplots, this context also addresses the outliers in the BAA minority group, and the disparity in the number of outliers in the BAA majority group.

## Abbreviations

BAA	Black and African American
inf	infinitive verb

## References

- Bachelier, Veronique, Jalal-Edine Zawam, Benoit Thieurmel, Francois Guillem & RTE. 2021. *leaflet.minicharts: Mini charts for interactive maps*. DOI: 10.32614/CRAN.package.leaflet.minicharts.
- Baxter, Kimberley. 2025. Extracting “non-standard” data from the Twitter API. In Susanne Wagner & Ulrike Stange-Hundsdörfer (eds.), *(Dia)lects in the 21st century: Selected papers from Methods in Dialectology XVII*, 3–30. Berlin: Language Science Press. DOI: 10.5281/zenodo.15006593.
- Blodgett, Sue Lin, Lisa J. Green & Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In Jian Su, Kevin Duh & Xavier Carreras (eds.), *Proceedings of the 2016 conference on empirical methods in Natural Language Processing*, 1119–1130. Austin, TX: Association for Computational Linguistics. DOI: 10.18653/v1/D16-1120.
- Eisenstein, Jacob. 2013. Phonological factors in social media writing. In Cristian Danescu-Niculescu-Mizil, Atefeh Farzindar, Michael Gamon, Diana Inkpen & Meena Nagarajan (eds.), *Proceedings of the NAACL/HLT 2013 workshop on Language Analysis in Social Media (LASM 2013)*, 11–19. Atlanta, GA: Association for Computational Linguistics. <https://aclanthology.org/W13-1102/>.
- Fisher, Sabriya. 2022. The status of *ain't* in Philadelphia African American English. *Language Variation and Change* 34(1). 1–28. DOI: 10.1017/S0954394522000060.
- Gopal, Deepthi, Tamsin Blaxter, David Willis & Adrian Leemann. 2021. Testing models of diffusion of morphosyntactic innovations in Twitter data. In Arne Ziegler, Stefanie Edler & Georg Oberdorfer (eds.), *Urban matters: Current approaches in variationist sociolinguistics* (Studies in Language Variation 27), 253–278. Amsterdam: John Benjamins. DOI: 10.1075/silv.27.
- Horvath, Barbara & David Sankoff. 1987. Delimiting the Sydney speech community. *Language in Society* 16(2). 179–204. DOI: 10.1017/S0047404500012252.
- Howe, Darin. 2005. Negation in African American Vernacular English. In Yoko Iyeiri (ed.), *Aspects of English negation*, 173–203. Amsterdam: John Benjamins. DOI: 10.1075/z.132.16how.
- Jones, Mari C. (ed.). 2015a. *Endangered languages and new technologies*. Cambridge, MA: Cambridge University Press.

- Jones, Taylor. 2015b. Toward a description of African American vernacular English dialect regions using “Black Twitter”. *American Speech* 90. 403–440. DOI: 10.1215/00031283-3442117.
- Jørgensen, Anna, Dirk Hovy & Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In Wei Xu, Bo Han & Alan Ritter (eds.), *Proceedings of the workshop on noisy user-generated text*, 9–18. Beijing, China: Association for Computational Linguistics. DOI: 10.18653/v1/W15-4302.
- Kautzsch, Alexander. 2012. *The historical evolution of Earlier African American English: An empirical comparison of early sources*. Berlin: de Gruyter Mouton. DOI: 10.1515/9783110907971.
- Labov, William, Paul Cohen, Clarence Robins & John Lewis. 1968. *A study of the non-standard English of Negro and Puerto Rican speakers in New York City*, vol. 2. Philadelphia: U.S. Regional Survey.
- Labov, William & Wendell A. Harris. 1986. *De facto segregation of black and white vernaculars* (Current Issues in Linguistic Theory). Amsterdam: John Benjamins. 1–24. DOI: 10.1075/cilt.53.04lab.
- Mitchell, Travis. 2019. *Sizing up Twitter users*. <https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/>.
- Moody, Simanique Davette. 2011. *Language contact and regional variation in African American English: A study of Southeast Georgia*. New York: New York University. (Doctoral dissertation).
- Nguyen, Dong, Dolf Trieschnigg, A Seza Dog, Mariet Theune, Theo Meder & Franciska de Jong. 2014. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 1950–1961.
- Orton, Harold & Eugen Dieth. 1962. *Survey of English dialects*. Leeds: E.J. Arnold, University of Leeds.
- Rickford, John R., Arnetha Ball, Renee Blake, Raina Jackson & Nomi Martin. 1991. Rappin on the copula coffin: Theoretical and methodological issues in the analysis of copula variation in African-American Vernacular English. *Language Variation and Change* 3(1). 103–132. DOI: 10.1017/S0954394500000466.
- Ryan, Camille L. & Kurt Bauman. 2016. *Educational attainment in the United States: 2015. Population characteristics. Current population reports* (Report number: P20-578). Suitland, MD: US Census Bureau. <https://www.census.gov/library/publications/2016/demo/p20-578.html>.
- Stevenson, Jonathan. 2016. *Dialect in digitally mediated written interaction: A survey of the geohistorical distribution of the ditransitive in British English using Twitter*. University of York: University of York. (MA thesis).

- Strelluf, Christopher. 2019. Positive-anymore, American regional dialects, and polarity-licensing in tweets. *American Speech* 94(3). 313–351. DOI: 10.1215/00031283-7587883.
- Strelluf, Christopher. 2020. Needs+PAST PARTICIPLE in regional Englishes on Twitter. *World Englishes* 39(1). 119–134. DOI: 10.1111/weng.12451.
- Tamir, Christine, Abby Budiman, Luis Noe-Bustamante & Lauren Mora. 2021. *Facts about the U.S. Black population*. Washington, DC: Pew Research Center's Social & Demographic Trends Project.
- US Census Bureau. 2015. *Data profiles*. <https://www.census.gov/programs-surveys/acs/>.
- US Census Bureau. 2018. *American Community Survey updates: 2018*. <https://www.census.gov/programs-surveys/acs/news/updates/2018.html>.
- Walker, Kyle & Matt Herman. 2023. *tidycensus: Load US census boundary and attribute data as “tidyverse” and “sf”-ready data frames*. DOI: 10.32614/CRAN.package.tidycensus.
- Weldon, Tracey. 1994. Variability in negation in African American Vernacular English. *Language Variation and Change* 6(3). 359–397. DOI: 10.1017/S0954394500001721.
- Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo & Hiroaki Yutani. 2019. Welcome to the Tidyverse. *Journal of Open Source Software* 4(43). DOI: 10.21105/joss.01686.
- Willis, David. 2020. Using social-media data to investigate morphosyntactic variation and dialect syntax in a lesser-used language: Two case studies from Welsh. *Glossa: A journal of general linguistics* 5(1). 103. DOI: 10.5334/gjgl.1073.
- Wolfram, Walt. 2007. Sociolinguistic folklore in the study of African American English. *Language and Linguistic Compass* 1(4). 292–313. DOI: 10.1111/j.1749-818X.2007.00016.x.
- Wolfram, Walt & Natalie Schilling-Estes. 2016. *American English: Dialects and variation*. 3rd edn. (Language in Society 25). Chichester, UK: Wiley Blackwell.
- Wolfram, Walt & Erik R. Thomas. 2002. *The development of African American English* (Language in Society 31). Oxford: Blackwell Publishers.



# Chapter 3

## The “Atlas of colloquial German in Salzburg”

Julian Blaßnigg<sup>a</sup>, Irmtraud Kaiser<sup>a</sup>, Peter Mauser<sup>a</sup> & Konstantin Niehaus<sup>a</sup>

<sup>a</sup>University of Salzburg

This article is concerned with the application of geolinguistic methods onto messy big data in dialectology. It presents data from an on-going project, the “Atlas of Colloquial German in Salzburg” (*Atlas zur Salzburger Alltagssprache*, ASA), which collected data on language use via a large online survey. The questionnaire consisted of 76 items, each of them given in a situational context. The goal of the project is to provide a first detailed description of the everyday German in the federal state of Salzburg, ranging from local dialect to regional dialect (regiolect) to near-standard and standard. The present study provides first results and identifies areas with similar diatopic variation in Salzburg via a cluster analysis of a total of approx. 10,000 data sets (collected via online questionnaires). Regional Spatial Planning Areas (*Planungsregionen*) prove to be a relevant category in order to account for diatopic variation in the state of Salzburg, with clusters of more traditional, dialect-orientated areas in the south, a regiolectal central area, and a northern urban area (encompassing the capital Salzburg City) orientated towards standard German. The article shows that even with anonymous and uncontrolled data collections, the cluster analysis offers a suitable approach towards larger dialect/regiolect areas as well as smaller transition areas. Ultimately, these methods may prove useful not only for the dialectology of (Austrian) German but also for other languages.



## **1 Introduction**

### **1.1 Motivation**

In this contribution, we want to demonstrate from the perspective of German dialectology how recent quantitative geolinguistic methods can be applied to relatively uncontrolled approaches of data collection (compared to traditional survey scenarios in dialectology), in our case crowd sourcing. To this end, we will introduce the project “Atlas of colloquial German in Salzburg” (ASA), which serves as the data basis for our analysis. The ASA arose from the idea of going beyond established dialectological approaches for a central linguistic area of Austria and to present the status quo of current language use in the federal state of Salzburg by collecting data on a broad scale. Compared to earlier dialectological approaches, the focus does not lie on selected prototypical speakers but on the colloquial language use of as many different people as possible. High diversity in terms of age, gender, education, profession etc. should go hand in hand with a dense network of places. We achieved this, as we received answers from nearly all municipalities in Salzburg and by very different people in terms of socio-demographic background. This should allow for statements about small-scale variation as well as it should also render a more representative picture of language use in the region, not least because of the large data source. Before going into more detail about data collection, we start with some general information about Salzburg and the current state of research on diatopic variation in this part of Austria.

### **1.2 General information about Salzburg**

Salzburg is one of the nine federal states of Austria. It is located in the central-western part of Austria. The capital is the city of Salzburg, which is located in the north of the state, right next to the German-Austrian border (cf. Figure 1). Salzburg connects the western part of Austria (Vorarlberg and Tyrol) with the eastern part and borders on four other Austrian federal states (Tyrol, Upper Austria, Styria and Carinthia).

In addition, parts of Salzburg’s border constitute a national border, stretching over 174 km, and mostly shared with the Federal Republic of Germany (districts of Berchtesgadener Land and Traunstein), i.e., 164 km. Only 10 km in the southwest border on the Tauferer Ahrntal in (German-speaking) South Tyrol (Italy) (cf. Landesstatistik Salzburg 2022).



Figure 1: Location of Salzburg in Austria (self created image)

The federal state of Salzburg covers an area of 7,154.6 km<sup>2</sup> and has a population of 562,606 inhabitants (01.01.2022).<sup>1</sup> Salzburg consists of six districts, one of which is the city of Salzburg. Perhaps more interestingly from a geolinguistic perspective, Salzburg is also divided into Außergebirg (literally ‘outside the mountains’) in the north and Innergebirg (‘in the mountains’) in the south of the Northern Limestone Alps (cf. Figure 2). This geographical division is accompanied by differences in terms of transport connections and infrastructure and therefore influences the dialectal situation in Salzburg. From a dialectological view, the South-Central Bavarian dialects cover most of Salzburg, a smaller part in the north belongs to the Central Bavarian group. Salzburg is therefore a central space in more than one respect: on the one hand, it is a purely geographical centre, due to the many geopolitical borders involved, and on the other hand, it also constitutes a centre of (dialect) variation, due to its transitional position between different dialect regions.

---

<sup>1</sup>For comparison: Austria has a population of approx. 9 million (01.01.2023).

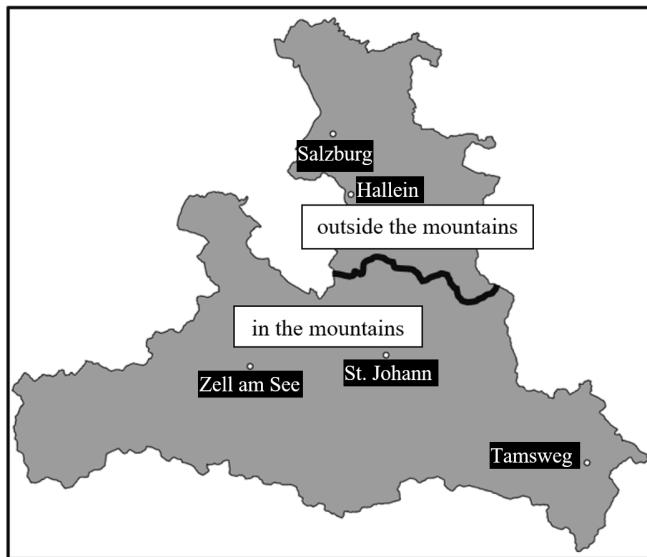


Figure 2: Division between ‘in the mountains’ and ‘outside the mountains’. The border corresponds to the district border between Tennen-gau and Pongau. N.B.: Populous areas in the South are mostly restricted to valleys. (self-created image)

### 1.3 State of research and research gaps

Dialectological and variational research has often been conducted only for selected locations in the state of Salzburg. Most studies in Salzburg dialectology (Reiffenstein 1955, Scheutz 2007, 2022, Scheutz et al. 2019, Mauser 2021, 2022) focus exclusively on the base dialect. There are also non-Salzburg-specific research projects which include data from the region like the Task-Cluster B of SFB *Deutsch in Österreich* (DiÖ) ‘German in Austria’, *Wortatlas der deutschen Umgangssprachen* ‘Lexical atlas of colloquial German varieties’ (Eichhoff 1977, 1978, 1993, 2000) and *Atlas zur deutschen Alltagssprache* (AdA) ‘Atlas of colloquial German’ (Elspaß & Möller 2003ff).<sup>2</sup> However, these research projects often only include data on a few locations and usually cannot capture variation within the individual regions of Salzburg.

There are three main aspects which are missing from previous research: first, data that go beyond the local base dialect; second, up-to-date variational data, not least for variation of regiolectal (regional dialect) and near-standard as well as standard lexis; third, more exhaustive and detailed data not restricted to sub-regions or individual locations in Salzburg. Hence, our goal was to conduct a

<sup>2</sup>For a popular science publication based on AdA data, cf. Leemann et al. (2018).

large online survey which draws from the method of the AdA. Its focus on language use includes any point on the vertical variational spectrum in order to grasp colloquial German in Salzburg regardless of its varietal status. In the next section we will elaborate in more detail on what is understood by *colloquial German* and how exactly the data collection took place.

## 2 Data and methods

### 2.1 Definition of colloquial language

Dialectological research in Salzburg typically concentrates on the base dialect, often archaic variants produced by NORMs/NORFs (the oldest form possible). The focus of this project, however, is the colloquial German language of the majority of speakers in Salzburg. So what is meant by the term *colloquial German*?

The project conforms with the AdA’s socio-pragmatic definition of colloquial language:

[D]iejenige Sprachform, die in informellen alltäglichen Situationen spontan und routiniert gesprochen wird (Elspaß 2010: 419)<sup>3</sup>

This criterion applies regardless of whether the language used would be categorised linguistically as standard language, regiolect, or (base) dialect. As a consequence, the project targets the whole spectrum of varieties, including the regiolectal and near-standard variants neglected so far.

### 2.2 Data collection

For the data collection, the project co-operated with the largest regional newspaper *Salzburger Nachrichten* in order to reach as many speakers as possible. All survey rounds were advertised and published by the *Salzburger Nachrichten*. Items used in the AdA questionnaire were replicated or modified. Some items were adapted to the Salzburg situation, for example in the case of expressions for ‘female child’ several variants in regional dialect were added (round 1, question 2). Additionally, regional variables expected to be more specific to Salzburg were created (e.g. variation of prepositions: *an der Bushaltestelle* vs. *bei der Bushaltestelle* ‘at the bus stop’). Sometimes genuine ASA items were created, i.e., variables that are specifically relevant to the (south-central) Bavarian language area,

---

<sup>3</sup>‘That form of a language which is spoken spontaneously and routinely in informal everyday situations’.

e.g. subordinate clauses introduced by *dass* versus introduced by *um ... zu* (round 4, question 6) (cf. Bayer 2020). The survey participants were asked to select their individual preferred variant from the options given.

Four rounds of questionnaires were distributed from October to December 2019, covering 76 variables of lexis, grammar, phonetics and pragmatics. We also collected multiple types of metadata for a quantitative variational analysis: place of residence, age, gender, highest educational qualification, occupational group, duration of residence and origin of parents. More than 10,000 speakers from Salzburg participated in the survey, of which 9,668 responses have been analysed. Some answers had to be excluded from the analysis due to missing data.

The questionnaire was created using LimeSurvey. We mostly relied on items with single answers but also used multiple answers in some cases, e.g. with variants of greeting and saying farewell. We tried to contextualise the item as much as possible and used visual material to explain meaning and context without additional language input. Each round consisted of approximately 20 thematic items, questions on socio-demographic metadata, and a free-comment section at the end of the questionnaire. The following welcome text provided respondents with an explanatory introduction to the survey:

Bitte geben Sie bei den folgenden Fragen an, was man in Ihrem Ort normalerweise hört – egal, ob eher Mundart, eher Hochdeutsch oder etwas dazwischen. Es gibt keine richtigen und falschen Antworten.

Am Ende des Fragebogens können Sie (unter „Ihre Anmerkungen“) beschreiben, ob und wie einzelne Ausdrücke unterschiedlich gebraucht werden (z.B. wie häufig, mit welcher unterschiedlichen Bedeutung usw.).

Versuchen Sie bitte, sich alle Beispiele in Ihrer ‚normalen‘ Aussprache vorzustellen. (ASA, round 4)<sup>4</sup>

Respondents' average age (40.7 years) roughly corresponds to that of the federal state (43.1 years). There is a slight overrepresentation of participants with higher educational qualifications (A-levels or degree). However, this is not unusual in surveys originating from the university sector – even if they are disseminated via the regional newspaper (for more details see Niehaus et al. 2022:

<sup>4</sup>In the following questions, please indicate what is normally heard in your location – whether it is more dialect, more High German or something in between. There are no right and wrong answers.

At the end of the questionnaire, you can describe (under “Your comments”), whether and how individual expressions are used differently (e.g. how often, with what different meanings, etc.). Please try to imagine all examples in your “normal” pronunciation.

N.B.: “Hochdeutsch” here is used in its lay meaning “Standard German”.

112–114). Moreover, there are significantly more female than male participants in our survey, the proportion of men is only 33.1%.

34.4% of the respondents have always lived in the place for which they provided the information, 18% for more than 30 years, 37.1% for more than 10 years, and only 9.2% have lived there for less than 10 years (no information: 1.3%). Thus, we can conclude that most respondents are sufficiently familiar with their place of residence and the local language.

## 2.3 Data analysis

### 2.3.1 Methods used

Data was collected and cleaned up, i.e. responses with missing data records were deleted. The remaining total of 9,668 responses from all four rounds were then examined more closely and in detail via SPSS. Statistical analyses were conducted for each district, planning region (see next section) and municipality in Salzburg. At the same time, the metadata were utilised in a quantitative variational analysis, e.g. with regard to gender<sup>5</sup> or age. Furthermore, social factors of regional language use were examined, e.g. whether young women from the city chose significantly more standard (or near-standard) variants than old men from rural areas (who chose variants of traditional dialect more often).

In order to determine whether the results are only due to individual phenomena or indeed similarities across all variables, the data were subjected to a factor analysis with GeoLing as well as a cluster analysis with R. To cluster data we used hierarchical cluster analyses with Ward’s algorithm. For variational linguistic applications, the Ward algorithm seems to be relatively useful as it does not overestimate smaller differences and fluctuations in the data but strongly sanctions larger differences. More information on cluster analysis in general can be found in Section 3.3.1.

Each item as well as the results of the factor and cluster analyses are finally visualised using ArcGis-Pro, which has proven to be well suited for the cartographic representation of geolinguistic data.

### 2.3.2 Basic diatopic levels of analyses

Popular belief in Salzburg may refer to several extra-linguistic criteria which index diatopic differences within the federal state. In terms of physical geography, the division into Innergebirg (in the south) and Außergebirg (in the north)

---

<sup>5</sup>In fact, men and women gave very similar answers to almost all questions. This is the reason why our results are not devalued by the somewhat imbalanced gender ratio.

is also of some importance. Innergebirg includes the districts of Zell am See (Pinzgau), St. Johann (Pongau) and Tamsweg (Lungau) while the Außergebirg includes the district of Hallein (Tennengau), Salzburg-Land (Flachgau) and the city of Salzburg (cf. Figure 2).<sup>6</sup>

The areas covered by the individual districts differ considerably, for example the largest district in size – the Pinzgau – occupies a cadastral area of 2,641.1 km<sup>2</sup> and is thus the third largest district in Austria, whereas the city of Salzburg measures only 65.7 km<sup>2</sup> and is thus significantly smaller than many municipalities in the federal state (see Figures 3 and 4). The opposite is true with regard to population size, where the city of Salzburg has the highest number of inhabitants. The data on which the ASA is based was generally analysed at the municipality level. A total of 109 of the 119 municipalities in Salzburg are included in all four rounds of our survey and were thus included in the analyses. Since approx. 92% of the Salzburg municipalities are represented in all survey rounds, it is possible to conduct fine-grained analyses at the municipal level. At the same time, some municipalities are represented by only a few responses, which means that an analysis of the results in larger areas seems to be useful in order to avoid generating a distorted picture of the actual language use by overrating a few individual answers.

Districts, on the other hand, often proved too coarse for our variational as well as geolinguistic analyses. For this reason, recourse was made to a level between municipalities and districts: the so-called *planning regions*. The planning regions are a unit of supra-local spatial planning, they are administratively anchored in the national spatial planning programme and also reflect historically-grown regional identities in many cases, which makes them ideal as a basis for geolinguistic analyses in Salzburg. The planning regions offer a good compromise and therefore lie at the core of our quantitative approach and cartography. In an extract from the *Salzburg Regional Development Programme* (revision 2003, published by the Department of Spatial Planning of the Salzburg Federal State Government) the planning regions are defined as follows:

Nach dem Raumordnungsgesetz ist es eine der ausdrücklich festgelegten Aufgaben des Landesentwicklungsprogramms, das Land in Planungsregionen zu gliedern. [...] Zu Regionen wurden darin nach Anhörung der Gemeinden Gebiete zusammengefasst, die strukturell und funktional zusammengehören und entsprechend den Erfordernissen der Raumordnung als

---

<sup>6</sup>This is reflected, for example, by dialect lexicons compiled by amateurs, which usually attempt to capture the dialect of a district, e.g. *Pinzgauer Mundart Lexikon* (Schwaiger & Höck 2010).

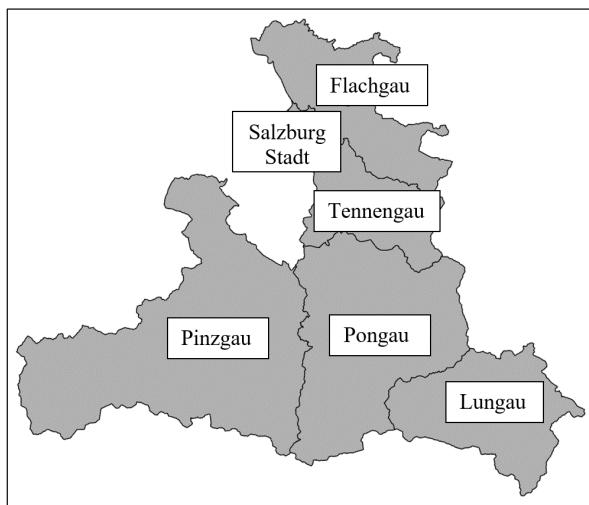


Figure 3: Districts of Salzburg (self created image)

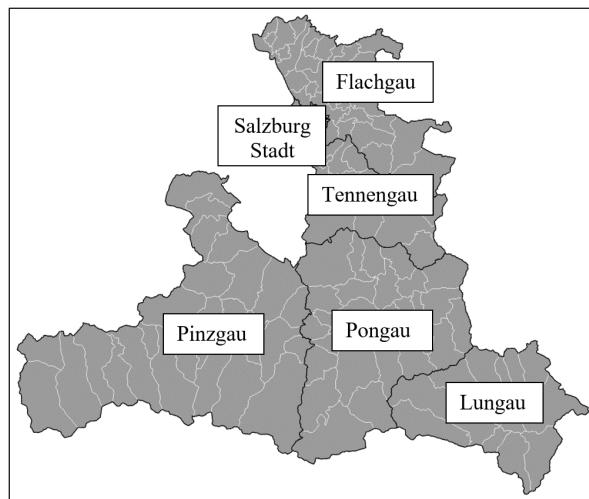


Figure 4: Municipalities in Salzburg (self created image)

Einheit entwickelt werden sollen, wobei auch die „Identifikation“ der Gemeinden mit einer bestimmten Region berücksichtigt wurde. (Mair 2003: 90)<sup>7</sup>

As can be seen, the planning regions are based on national planning as well as regional identities.

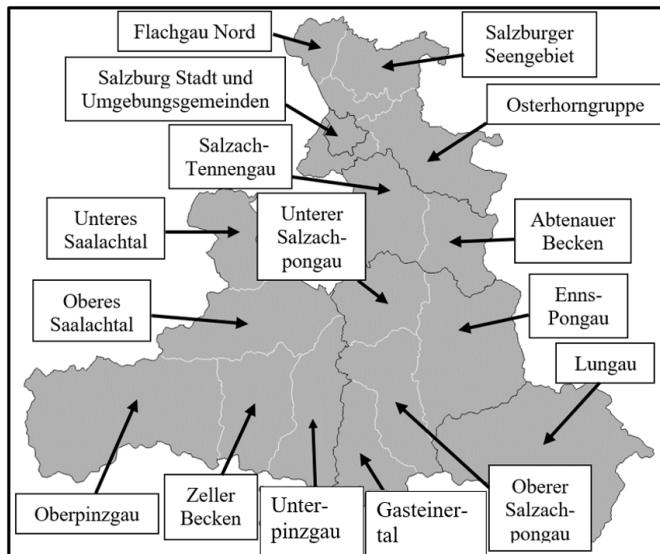


Figure 5: Official regional planning regions (cf. Mair 2003) – (self created image)

Each of the 16 planning regions in the province of Salzburg consists of an average of seven municipalities. As far as the number of respondents per planning region is concerned, numbers range from 102 (round 4, Zeller Becken) to 1003 (round 1, Lungau) per 100,000 inhabitants. These numbers ensure a realistic approximation to inner-regional language use in Salzburg while at the same time much more detailed insights can be gained than if only the district level were considered. Additionally, analyses were carried out at district and municipality level in order to avoid pitfalls during interpretation; most importantly, we applied multidimensional scaling (MDS) on both levels, which allows for a calculation of linguistic similarities.

<sup>7</sup>According to the Spatial Planning Act, one of the expressly defined tasks of the Development Programme is to divide the federal state into Regional Spatial Planning Areas [“Planungsregionen”]. [...] After consultation with the municipalities, areas that belong together structurally and functionally and are to be developed as a unit in accordance with the requirements of spatial planning were grouped together in these regions, whereby the “identification” of the municipalities with a particular region was also taken into account<sup>7</sup>.

## 3 Findings

### 3.1 Mapping: General approach

The following maps depict one variable each, i.e. dominant variants in a combined heat map. For this purpose, each dominant variant is given its own colour, while the relative frequency of the dominant variant is represented by colour intensity. Put simply: the more intense the colour, the stronger the prevalence of a variant. All sub-divisions on the maps are based on the planning regions, with district boundaries and the respective district capitals visualised for better orientation. At the same time, many maps show that the distribution of individual phenomena is often not identical with district boundaries, which increases the additional informative value of analyses at the level of planning regions. To illustrate this, two maps will be showcased below and their results will be presented as examples. The presented maps (Figures 6 and 7, page 68) should be understood such that the intensity of the color – either red or blue – correlates with the prevalence of the variant. The stronger the color shading, the more widespread the variant.

### 3.2 Showcase results

#### 3.2.1 Example ‘female child’

The designations for ‘female child’ are a classic in German-speaking dialectology. The German dialects in general, including those spoken in Austria, boast a multitude of expressions and phonetic variants of these expressions for referring to the category of ‘female child’. Our results for Salzburg show that two expressions are particularly common here: *Mölz* and *Dirndl*.

The south-western region of the Pinzgau is strongly dominated by *Mölz* except for the Unteres Saalachthal (‘Lower Saalach Valley’), an area in the northern Pinzgau. This demonstrates the strength of working with planning regions, for the popular assumption that *Mölz* is the preferred term for ‘female child’ in the entire Pinzgau cannot be confirmed by our data. Our survey would also have yielded this very result if the districts alone had been the basis of our cartographic representation. By using planning regions, however, it becomes apparent on the one hand, that there is a region in the Pinzgau where *Mölz* is not at all the dominant variant and, moreover, that the frequency of this particular variant differs within the district, specifically that *Mölz* is used most frequently in the far west. Similar observations can be made for *Dirndl*: Although it is the most frequently used variant throughout the rest of the federal state, usage is particularly frequent in the

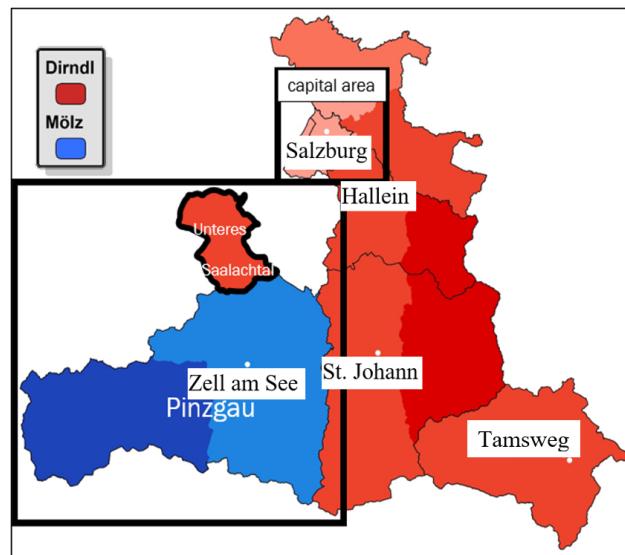


Figure 6: Names for 'female child' (self created image)

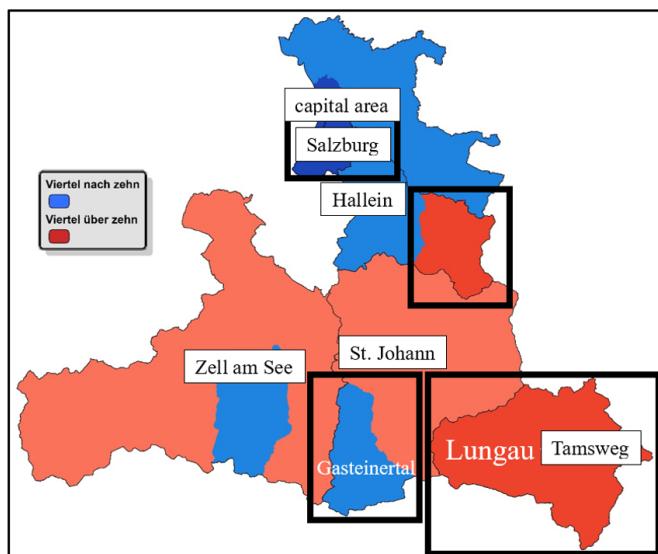


Figure 7: Verbalisation of 10:15 (self created image)

east of the federal state. In the capital area around the city of Salzburg, *Dirndl* is only just the most frequently mentioned designation for ‘female child’: Although it is still dominant, there is a higher percentage of standard variants – and more variation in general (which does not come as a surprise and applies to many variables). In addition to *Dirndl*, the standard German *Mädchen* (approx. 17%) and near-standard *Mädl* (approx. 36%) are also frequently used in and around the city of Salzburg. Hence, *Dirndl* is not a quasi-absolute default variant in most of the federal state but a relatively dominant variant most prevalent in the east.

### 3.2.2 Example 10:15

A particularly suitable item for our demonstration of the data is the verbalisation of a certain time of day, in (standard) English *ten-fifteen* or *quarter past ten*. To avoid verbal stimuli, this question only showed a picture of a clock and asked for the phrase to name the corresponding time. The German variants in the state of Salzburg are *Viertel nach zehn* ‘quarter past ten’ (in blue) and *Viertel über zehn* ‘quarter over ten’ (in red). The map (Figure 7) demonstrates a pattern which surfaces often in our data, i.e. an opposition between the more urban north (capital area) and the more rural southeast (Lungau), notwithstanding a few exceptions such as the northern Abtenauer Becken (‘Abtenau basin’), which often joins the language use of the south – and vice versa the southern Gasteinertal (‘Gastein valley’), which often conforms with the language use of the northern capital area. Within the individual districts clear-cut areas with different degrees of variant use become apparent.

## 3.3 Results of cluster analyses

### 3.3.1 General approach

Our data have a nominal scale form (as is customary in variational linguistics) and can easily be spread in a table: Each row of the table contains the data of exactly one informant. The columns represent the variables of the survey, the individual cells contain the respective variant that the informant has given for the variable. As already mentioned, most of the variables are of a purely linguistic nature, with the exception of the metadata information on the respondents as well as the free comments section. These are also documented individually for each informant.

This was the basis on which we evaluated the most prevalent variants for each planning region. However, our interest is not only in the results for each individual item but we also wanted to identify inter-local and inter-regional similarities,

which only become apparent with a sufficient amount of data combined. Factor and cluster analyses are suitable for such approaches (cf. Goebel 2005: 509, Pröll et al. 2021: 234–235, note 14). In the following, the results of the cluster analysis will be presented.

The basis for these further analytical steps is no longer the original table of nominally scaled, individual raw data, but a table with the relative frequencies of the individual variants per survey location. These calculations were conducted for all our diatopic levels of analysis, i.e. the districts, the planning regions, and single municipalities. In the present paper, we will only take a closer look at two cluster maps which are based on the planning regions. A distance matrix of the planning regions to each other is calculated from the relative frequencies of the linguistic variants per planning region. In order to make this high-dimensional variation less complex and interpretable, cluster analysis was used: The individual planning regions were grouped based on their similarity or dissimilarity. These groups (= clusters) should be as homogeneous as possible within themselves and at the same time as different as possible from each other. Ultimately, therefore, one obtains a concrete allocation of the individual regions to groups, which in turn can be represented and interpreted cartographically. These groupings can be represented at different levels of granularity, e.g. as a 2-cluster, 3-cluster or 4-cluster solution.

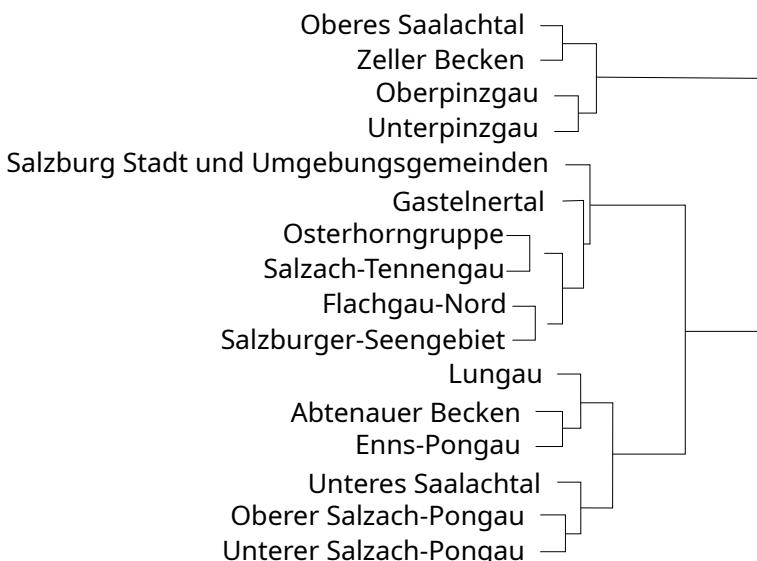


Figure 8: Cluster analysis dendrogram (self created image)

The allocation of the planning regions to individual clusters is best represented in the form of a dendrogram (cf. Figure 8). Depending on where the cut is made to distinguish clusters, solutions with different numbers of clusters are obtained. Since the length of the horizontal lines provides information about the linguistic distance between the individual clusters, the dendrogram can also be used to determine which cluster solutions ultimately represent plausible groupings.

As can be seen from the dendrogram, mainly 2-, 3- and 4-cluster solutions are suitable. We will briefly discuss a 3- and a 4-cluster solution.

### 3.3.2 3-cluster solution

Popular belief as to how the federal state of Salzburg could be divided into dialect areas include the idea that there is a clear linguistic division between the dialects Außergebirg (‘outside the mountains’, i.e. in the north) and Innergebirg (‘in the mountains’, i.e. in the south). In addition, political districts are ascribed their own regional dialects such as *Pinzgauerisch* in the south-western part of the state and the like. The 3-cluster map (cf. Figure 9) shows that the actual linguistic situation is clearly more complex.

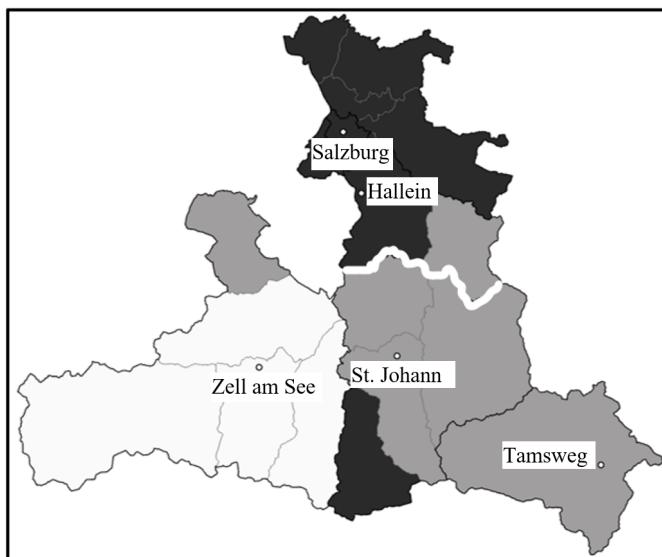


Figure 9: 3-cluster solution (self created image)

In addition to the boundaries of the planning regions and the district capitals, we also drew the border between Innengebirg and Außengebirg (bold white line).

The border runs between the two districts of Tennengau and Pongau. From the perspective of physical geography, the lay-linguistic notion is understandable: Some Salzburgers from the Innengebirg travelling to the capital will immediately notice a change, when the Salzach valley widens outside the Pass Lueg and the rugged mountains give way to gentle hills, and, of course, natural obstacles can be a factor of linguistic variation. However, linguistic difference may subjectively be perceived as much more salient and relevant than the distance measured by linguistic means. For example, the Abtenauer Becken (in grey, north of the white line – a basin in terms of geology) clearly belongs to the south in terms of linguistic distance in this cluster solution. On the other hand, the cluster resembling the Außengebirg area (black) also includes the Gasteinertal in the south, which is geographically clearly a mountain valley of the south. On top of that, the Innengebirg does not add up to a single cluster but instead consists of three clusters, mainly two large clusters in white and in grey in the (south-)west and in the (south-)east respectively. Not even the Pinzgau district (white) consists of a single cluster, as again the northern part, Unteres Saalachtal, belongs to a different cluster – a phenomenon which could already be observed with the lexical variants of ‘female child’ (see Section 3.2.1, Figure 6). The map for ‘10:15’ (see Section 3.2.2, Figure 7) again shows very well the outliers also evident in the 3-cluster solution, namely the Abtenauer Becken and the Gasteinertal, which in many maps stand out with regard to other linguistic variables from their surroundings. In sum, it can be said that neither the linguistic relevance of political districts nor the popular binary distinction of varieties *in* and *outside* the mountains can be confirmed through our cluster analysis. Rather, it seems that other factors play a crucial role, such as the infrastructural proximity to the city of Salzburg. At the same time, the cluster which mainly covers the south-east and centre-most areas seems to form a transition area between the urban north around the city of Salzburg and the rural south(-west). The fact that regional centres such as St. Johann im Pongau are relatively well-connected to the city of Salzburg in terms of infrastructure seems to be at work here. Even from the Unteres Saalachtal – via a short route through German territory – you can reach the city of Salzburg relatively fast, from some places (e.g. Unken) even faster than the district’s own capital, Zell am See.

### 3.3.3 4-cluster solution

The 4-cluster solution (see Figure 10) can refine the 3-cluster picture a little more: Differences to the 3-cluster solution are especially evident in the transition area, which now breaks up into two clusters.

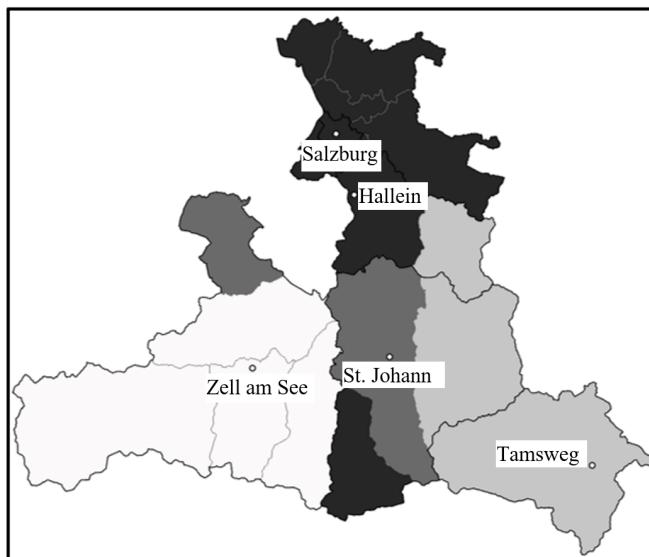


Figure 10: 4-cluster solution (self created image)

Similar to the 3-cluster solution, there is a pseudo-Außengebirg cluster (in black), with a tendency towards either standard variants or Bavaro-Austrian regiolectal variants (that is, variants which are also common in other parts of Austria and in south-eastern Germany/Bavaria). In the Innergebirg, the large eastern cluster of the 3-cluster solution breaks apart into smaller clusters, one transition area more similar to the northern cluster (in dark grey) and the easternmost rural area, which maintains base dialect variants more strongly (in light grey).

The differences between the dark grey and the light grey clusters can be illustrated by numerous examples: the Pongau central area (in the west of the district of St. Johann in dark grey) more often tends to use variants that are common in the city of Salzburg and its surroundings, e.g. the grammatical gender of email, neutr. *das E-Mail* (Enns-Pongau) vs. fem. *die E-Mail* (Pongau central area). The West (in white) is obviously orientated towards dialect variants similar to those of adjacent west Austrian dialects, such as Tyrolean, a relatively stable result regardless of the number of clusters.<sup>8</sup>

The assumption that the infrastructural proximity to the city could play a decisive role is corroborated by this 4-cluster solution. Those regions of the original transition cluster in the 3-cluster solution which are particularly well connected

<sup>8</sup>Cf. data in the *Tiroler Dialektarchiv* (Archive of Tyrolean Dialects): <https://www.tiroler-dialektarchiv.at>. For instance, *Mölz* appears as *Mötz* in Tyrol, cf. [https://wiki.uibk.ac.at/tda/index.php/Bedeutung\\_14](https://wiki.uibk.ac.at/tda/index.php/Bedeutung_14).

to the city of Salzburg now fall into the more urban, regiolectal<sup>9</sup> transition cluster (dark grey), while the eastern and southeastern regions (light grey) are closer to the rural-dialectal forms of the southwest (white). This interpretation is corroborated by an MDS-analysis.

Once again, however, the Gasteinertal in the south (black), which is located in the mountains, joins the northern urban cluster. Infrastructural proximity to the city can hardly account for this. Yet, a long-standing history of (health and spa) tourism and subsequent migration (e.g. German or Viennese retirement residences in the Gasteinertal, going back as far as the late 19th century) may explain the similarities of the Gasteinertal and the northern regions: It entails either the tendency to converge towards wider regional, non-local variants or to use standard variants. Historical migration through jobs in tourism and retirement could have left a linguistically sustainable mark by changing the composition of the population. We will pursue this case further in the future, among other factors that still await analysis.

## 4 Conclusion and outlook

In addition to traditional dialect areas and popular divisions (Innengebirg/Außengebirg), which are only partially confirmed by our data, various factors influence patterns of language variation in the federal state of Salzburg.

Our analysis has demonstrated that infrastructural proximity to the city of Salzburg seems to be a major factor which may have contributed to creating a transitional regiolect. Infrastructure seems to be at least as important as mere geographical distance, sometimes it even trumps the latter. In other cases, this results in groups of diatopic variants which can only be explained by consulting both factors: The Abtenauer Becken, for example, is geographically closer to the city of Salzburg than, for example, St. Johann. However, St. Johann is better connected to the city of Salzburg in terms of infrastructure due to the nearby motorway and the railway connection. The MDS-analysis mentioned above shows that St. Johann is in fact linguistically closer to the city of Salzburg than the Abtenauer Becken.

Smaller regional centers also play a certain role in many cases (in addition to the city of Salzburg as the overall centre of the federal state). These are often also the target of commuter flows and usually have a better infrastructure than many

<sup>9</sup>Common standard linguistic variants would be, for example, *Mädchen* 'female child' as opposed to regiolectal *Dirndl* and dialectal *Mölz/Tättn*, but also pronunciation variants such as [ˈmilç] *Milch* 'milk' as opposed to regiolectal [ˈmy:ç] and dialectal [ˈmi:ç] as well as subjunctive II constructions such as (wir) *bräuchten* 'we would need' as opposed to regiolectal *täten/dadn ... brauchen* and dialectal *brauchatn* (cf. also Niehaus et al. 2022).

rural communities in the mountain regions. At the same time, these regional centres often show linguistic characteristics that bring them closer to the regionally influenced urban language of the capital than to the rural base dialects which surround these regional centres. In sum, the situation is strongly reminiscent if not a prime example of Trudgill’s (1974) gravity model where linguistic change spreads from one (urban) centre to another and only then diffuses into the surrounding areas. Accordingly, there appears to be a preservation of traditional dialect variants outside these smaller centres, i.e. in the westernmost and easternmost rural parts of Salzburg, in contrast to more dynamic areas in the north.

This insight has been gained by a combination of quantitative analyses which are sufficiently fine-grained and could be complemented by qualitative studies to account for the data in future approaches. As the case of the Gasteinertal has shown, even with comprehensive statistics for a large amount of data one must take into account the specific histories of locations which can leave their mark linguistically, irrespective of geographic and/or infrastructural circumstances.

Finally, social factors such as age and profession are expected to play a significant role for the regional language use in Salzburg (e.g. young urban vs. old rural). Analyses on these factors will be presented in forthcoming research.

## References

- Bayer, Josef. 2020. What’s unique in Bavarian syntax? Thoughts on the occasion of a performance of Bach’s St Matthew Passion. In Tomoya Tsutsui Tanaka & Masashi Hashimoto (eds.), *Linguistic research as an interdisciplinary science*, 27–43. Tokyo: Hituzi Publishers.
- Eichhoff, Jürgen. 1977. *Wortatlas der deutschen Umgangssprachen*, vol. 1. Bern: Franke.
- Eichhoff, Jürgen. 1978. *Wortatlas der deutschen Umgangssprachen*, vol. 2. Bern: Franke.
- Eichhoff, Jürgen. 1993. *Wortatlas der deutschen Umgangssprachen*, vol. 3. München: Saur.
- Eichhoff, Jürgen. 2000. *Wortatlas der deutschen Umgangssprachen*, vol. 4. Bern: Saur.
- Elspaß, Stephan & Robert Möller. 2003ff. *Atlas zur deutschen Alltagssprache*. <https://www.atlas-alltagssprache.de>.
- Elspaß, Stephan. 2010. Alltagsdeutsch. In Hans-Jürgen Krumm, Christian Fandrych, Britta Hufeisen & Claudia Riemer (eds.), *Deutsch als Fremd- und Zweitsprache. Ein internationales Handbuch*, 2nd edn., vol. 1, 418–424. Berlin, New York: de Gruyter.

- Goebl, Hans. 2005. Dialektometrie. In Reinhard Köhler, Gabriel Altmann & Rajmund G. Piotrowski (eds.), *Quantitative Linguistik: Ein internationales Handbuch* (Handbücher zur Sprach- und Kommunikationswissenschaft), 498–531. Berlin, New York: de Gruyter.
- Landesstatistik Salzburg. 2022. *Gemeindeportraits*. <http://www.salzburg.gv.at/themen/statistik/gp-statistik-daten-gemeindeportraet>.
- Leemann, Adrian, Stephan Elspaß, Robert Möller & Timo Grossenbacher. 2018. *Grüezi, Moin, Servus*. Reinbek bei Hamburg: Rowohlt.
- Mair, Friedrich (ed.). 2003. *Salzburger Landesentwicklungsprogramm*. Gesamtüberarbeitung. Salzburg: Amt der Salzburger Landesregierung.
- Mauser, Peter. 2021. *Wiarach ba ins ret: Das Lungauer Sprachbuch*. Tamsweg: Pfeifenberger.
- Mauser, Peter. 2022. *Wiarach ba ins ret: Die sprechende Landkarte zum Buch*. <https://www.pfeifenberger.at/wiarach/landkarte/>.
- Niehaus, Konstantin, Irmtraud Kaiser & Peter Mauser. 2022. Der Konjunktiv II in Salzburger Varietäten: Grammatik, Gebrauch, soziale Faktoren. *Linguistik Online* 114(2). 99–128.
- Pröll, Simon, Stephan Elspaß & Simon Pickl. 2021. Areal microvariation in German-speaking urban areas (Ruhr area, Berlin, and Vienna). In Arne Ziegler, Stefanie Edler & Georg Oberdorfer (eds.), *Urban matters: Current approaches of international sociolinguistic research*. (Studies in Language Variation 27), 227–252. Amsterdam, Philadelphia: Benjamins. DOI: 10.1075/silv.27.10pro.
- Reiffenstein, Ingo. 1955. *Salzburgische Dialektographie. Die südmittelbairischen Mundarten zwischen Inn und Enns* (Beiträge zur deutschen Philologie 4). Gießen: Wilhelm Schmitz Verlag.
- Scheutz, Hannes (ed.). 2007. Drent und herent: *Dialekte im salzburgisch-bayerischen Grenzgebiet*. Mit einem sprechenden Dialektatlas auf CD-ROM. Unter Mitarbeit von Sandra Aitzetmüller und Peter Mauser. Salzburg: Salzburg – Berchtesgadener Land – Traunstein EU Regio.
- Scheutz, Hannes. 2022. *Salzburger Sprachatlas*. <https://www.sprachatlas.at/salzburg/index.html>.
- Scheutz, Hannes, Aitzetmüller Sandra & Peter Mauser. 2019. Drent und herent: *Dialekte im salzburgisch-bayerischen Grenzgebiet*. <https://www.sprachatlas.at/drentherent/>.
- Schwaiger, Alois & Leonhard Höck. 2010. *Pinzgauer Mundart Lexikon*. Leogang.
- Trudgill, Peter. 1974. Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. *Language in Society* 2(2). 215–246. DOI: 10.1017/S0047404500004358.

# Chapter 4

## A cognitive geographic approach to dialectology: Cognitive distance as a predictor for perceptual dialect distance

Hedwig G. Sekeres<sup>a</sup>, Martijn Wieling<sup>a</sup> & Remco

Knooihuizen<sup>a</sup>

<sup>a</sup>University of Groningen

In this study, we explain perceptual dialect differences using both geographic distance and a different type of distance that is commonly used in the field of cognitive geography. Cognitive geography is based on the assumption that an individual's mental representation of their environment has a greater effect on their behaviour than the actual environment (Montello 2018). A commonly used metric in cognitive geography is the cognitive distance: the geographic distance between two places as estimated by an individual (Montello 1991). This study also introduces the use of cognitive distances into dialect research and investigates whether these mental representations of space can serve as an explanatory variable in dialectology.

Nearly 800 participants from the north of the Netherlands provided cognitive distances between the place where they grew up and seven other locations in the same region. They also rated the similarity of dialect recordings from these locations to the dialect of the location where they grew up. A linear mixed-effects regression model was built to predict perceptual dialect distance from both cognitive distance and geographic distance. The resulting model indicates that geographic distance is more predictive of perceptual dialect distance than cognitive distance. There was also a significant interaction between cognitive and geographic distance. Cognitive distance is more predictive of perceptual dialect distance when geographic distance is short than when geographic distance is long. Furthermore, an exploratory analysis revealed that gender and proficiency in the participants' local dialect were predictive of perceptual dialect distance. Our findings indicate that cognitive distance can be used to explain dialect variation, especially when the area under investigation is small, and consequently that the framework of cognitive geography can be usefully employed in dialectological research.



## 1 Introduction

In dialectology, space has historically been treated for the most part as a blank canvas on which linguistic (and sometimes social) information is visualised. This treatment of space has been criticised as too one-dimensional and dismissive of findings from different subdisciplines of geography (e.g., Britain 2013), but as of yet only a small amount of research has been done in which a broader consideration of space in dialectology is taken. Perceptual dialectology to some degree forms an exception to this, as methods from the field of cultural geography have been adapted for dialect research by perceptual dialectologists, most notably Preston (e.g., 1981, 1999). However, while cultural geography concerns itself with the spatial distribution of culture (Anderson et al. 2003) and in the case of perceptual dialectology, language and language attitudes, the ways in which space itself can influence attitudes and behaviour are not its main focus. An analysis of this relationship forms an interesting addition to the analysis of dialect variation. One field that does concern itself with this particular relationship is *cognitive* (as opposed to the aforementioned *cultural*) geography, which focuses on the mental representations of space that people have and how these influence their behaviour. This study introduces the cognitive geographic measure of *cognitive distances* into dialect research and investigates whether these mental representations of space can serve as an explanatory variable in dialectology. Specifically, we do this by investigating whether including cognitive distance, in addition to geographic distance, leads to better predictions of perceptual dialect distance in the Gronings dialect area.

## 2 Background

### 2.1 Traditional notions of space in dialectology

In traditional dialectology, space is usually represented by a map on which language variation is presented. The first dialect maps were often representations of the geographical variation of one word or a single linguistic feature, where different variants of that word or the different variants of the feature were written on the map at their respective locations (Rabanus 2017). Although these maps were detailed and precise, they were not very insightful as it can be difficult to discover patterns in the spatial distribution of words. Later dialect maps made use of points with different shapes or colours to signify the different variants that were present and provided a more easily accessible overview of the geographic

patterns of language variation. A different type of dialect map uses isoglosses to signal borders in dialects or languages, usually between different variants of a feature. Maps consisting of bundles of isoglosses made it easier to distinguish between areas in which related variants are used (Rabanus 2017). Although space plays an important role in these visualisation techniques, it is for the most part treated as a blank canvas and a recipient for linguistic (and sometimes extra-linguistic) information. Some conclusions about spatial patterns other than geographic distance can be based on the administrative and geographic features that are sometimes present on these maps, but not in a systematic manner. This is, of course, not to say that social, historical and political patterns that are themselves spatially stratified have not been used to explain (spatial) variation in dialects. In Germany and the Netherlands, in particular, expansiological dialectology has been influential in offering alternatives to explaining spatial patterns that are not purely geographical by including information on communicative patterns (Goossens 1977). Space itself as an explanatory variable had not been taken into account in dialectology until the early 1970s, when different models of the diffusion of language change were proposed. These new models of diffusion were a criticism of the wave model (Bailey 1973), which proposed that new variants moved through space more or less in a straight line, spreading from one location to the next on the basis of proximity. The most prominent example of these new models is the gravity model (Trudgill 1974), which proposed that language change does not spread through space evenly but that new variants take hold in large population centres first and then spread to rural areas. In the gravity metaphor that is used, urban centres exert a certain gravitational pull on new variants, pulling them towards themselves before they reach rural areas. In this model, social aspects of space are granted more importance than in the wave model as it takes into account other (non-linguistic) phenomena that are stratified across space and that influence both space itself and human (linguistic) behaviour. Research on the spread of language change often found conflicting results. One study on Oklahoma dialects found, for example, that change from above followed a different spatial pattern, spreading from urban to rural, than change from below, which spread from rural to urban (Bailey 1973, Labov 1965). Here we find that dialectology really starts to engage more with space and spatial patterns, as the interactions between linguistic and spatial phenomena are researched. In the Netherlands, a dialectometric study investigating the use of the gravity model in the Netherlandic Low Saxon language area found patterns of diffusion that were the opposite of what the gravity model would predict (Nerbonne et al. 2005). However, that study did not differentiate different types of features, so no interaction between spatial and linguistic patterns could be found.

Some special attention regarding the interaction between language and space is merited by the field of perceptual dialectology as it has engaged with methods from different geographic subdisciplines from an early stage. Whereas traditional dialectology concerns itself with dialects as described by linguists, the field of perceptual dialectology investigates the attitudes and beliefs that non-linguists have about dialects and dialect variation (Preston 1999).

The field of perceptual dialectology came into existence more or less simultaneously in Japan and the Netherlands, with accounts differing on the birthplace of the field (Montgomery & Beal 2011). In the Netherlands, the so-called *pijltjes-methode* (or ‘little-arrow method’) was developed. In this approach, participants were asked to indicate in which locations the dialect was the same or very similar to their own dialect. This information was visualised on maps with small arrows that indicated which areas were seen as similar by the participants. These maps were then compared with dialect maps based on known dialect differences and isoglosses in order to assess whether the participants’ perception of dialects matched the actual dialect situation. For many areas, this was the case although interesting differences between production and perception of dialect differences arose as well (Weijnen 1946). In Japan, dialectologists took a similar approach in which they asked participants to rate the similarity of their dialect to other dialects on a continuum. The results of this task were visualised on a map in which lines were drawn to signify the different perceptual areas (Sibata 1999). Results from these studies differed more starkly from the dialect areas that were traditionally distinguished, and gave rise to the investigation of different components of language variation in dialectology, such as pitch accent (Montgomery & Beal 2011).

From the 1980s onwards, the field of perceptual dialectology underwent great development, when Dennis Preston incorporated mental maps from the field of cultural geography in dialectology (Montgomery & Beal 2011, Preston 1981). In cultural geography, mental maps have been used from the 1960s onwards to visualise the mental representations that people have of space (Lynch 1964, Portugali 2018). This is usually done by prompting participants to draw a map of a certain area as they remember or perceive it, sometimes with a specific goal in mind such as to draw important landmarks or their commute to work. Mental maps provide insight into the places people find important and how they believe these places relate to each other (Gould & White 1974). Preston adapted the mental maps to dialectological research and created the so-called *draw-a-map task*. In this task, participants are asked to indicate on a map where they would locate certain speech patterns, and sometimes also which types of people they associate with these patterns. With the introduction of this task, space started to play an

integral role in perceptual dialectology as the visualisation of linguistic information is not only done by researchers after the linguistic data has been collected, but the participants themselves are also required to think about dialects in space. In this way, linguistic data are generated in a manner that is inherently more spatially oriented. However, as the draw-a-map task was borrowed from the field of cultural geography rather than cognitive geography, the mental representations of space itself are not used as an explanatory variable in this type of research. As with the maps made in traditional dialectology, space is rather a blank canvas on which linguistic and – in this case – social data are visualised, and not a factor that is to be taken into consideration on its own. One notable addition to the analysis of these maps is the use of the cultural prominence of locations by Montgomery (2012). He found that areas that were drawn by participants on their dialect maps had almost without exception undergone an increase in cultural prominence based on the number of mentions in two newspapers per head of population. Although the participants themselves in this case did not give indications of the cultural prominence of regions, this approach does acknowledge that properties of a region can affect people's dialect perception.

Next to the draw-a-map task, Preston developed and adapted several other techniques for perceptual dialectological research. One such technique is asking participants to rate the difference between a given dialect and the dialect spoken at the place where the participant grew up. This measure, also called the *perceptual dialect distance*, is also employed in our study. Other methods introduced by Preston are the collection of non-linguists' ratings of the correctness and/or pleasantness of dialects, asking participants to identify a dialect and to link it to a geographic location, and conducting interviews to gather qualitative data on non-linguists' attitudes and beliefs, for example by asking them about the previous tasks or to talk about characteristics of the speakers of certain variants (Preston 1999).

Research in perceptual dialectology actively engages with space in the sense that participants in a draw-a-map task or an interview are asked to think about language spatially. However, the variables used to explain why participants hold certain attitudes or give certain judgments are typically social, not spatial. Treating space itself as an explanatory variable that encompasses more than merely geographic distance is still relatively rare. This omission has been criticised in the field of dialectology in general: "Space has largely been treated as an empty stage on which sociolinguistic processes are enacted. It has been unexamined, untheorised and its role in shaping and being shaped by variation and change untested" (Britain 2013: 471). There are developments in this regard, with several authors pointing out the importance of perceptual space in language variation

(e.g. Britain 2011, Preston 2010), but to our knowledge this has not yet resulted in studies in which perceptual space is quantified and included as a variable.

## 2.2 Alternative notions of space in dialectology

One of the most frequently used alternatives for geographic (as-the-crow-flies) distance in dialectology is travel distance, expressed as the number of kilometres one would need to travel to get from location A to location B. The advantage of using travel distance over geographical distance is that it provides researchers with more insight into possible contact situations. Using travel distance as a way to incorporate this geographic information in dialectology was done, for example, by van Gemert (2002), and later adapted in the form of travel times by Gooskens (2004). Gooskens found that dialect distances in Norway correlated more strongly with historical travel times than with simple Euclidean distances, which was explained by the amount of contact that dialects historically had, due to the amount of time that travel would take (Gooskens 2004). The use of both travel distances and travel times in dialectological research has further successfully been demonstrated in a study on Japanese dialects (Jeszenszky et al. 2019), in which an increase in travel distance and travel time correlated positively with an increase in dialect distance. Although neither travel distance nor travel time outperformed geographic distance in this study (partly because of the difficulty to include travel over water), travel times are still a promising approach to incorporate both space itself and the *experience* of space in dialectological research. Travel times in particular are suitable for this goal, as they relate to the way a journey is experienced: travelling 20 kilometers by car is experienced differently than travelling the same distance by foot.

Next to travel distances and travel times, there are still other alternative notions of space used in dialectology that are more reflective of contact situations than geographic distance. In his study on the small clan-based Sui society in southwest China, Stanford (2012) used the least cost distance between rice paddies instead of Euclidian or travel distances. He postulated that this distance measure forms a better reflection of mobility among the Sui people because it represents a more stable and historical social pattern than roads. The rice paddy distance explained dialect variation marginally but not significantly better than simple geographic distance (Stanford 2012). Although the results of this study are not conclusive regarding the effects of using a different distance measure than geographic or travel distance, the study poses an interesting view on using a distance measure that more accurately reflects mobility patterns of a particular society. In recent years, more importance has been placed on the consideration

of personal mobility in dialect research, although this interest does not necessarily translate into different types of distance measures. Britain (2013), for example, has argued for a more thorough investigation of mobility practices in border regions in order to understand why certain spatial patterns emerge in language use that are not easily explained by geography itself. Furthermore, a recent publication by Jeszenszky et al. (2024) proposes a personal mobility index for use in dialectological research. These developments offer a promising perspective in the future treatment of space in dialectology.

### 2.3 Cognitive geography

In treating space for the most part as a combination of degrees of longitude and latitude, dialectological research does not take into account that people do not simply exist in space. They are also influenced by their spatial environment, and vice versa. One field that recognises this bi-directional relationship between people and their environment is cognitive geography. Like cultural geography, which has previously been employed in perceptual dialectological studies, cognitive geography is a subdiscipline of human geography, which concerns itself primarily with the relationship between humans and their environment. Although there is considerable overlap between cultural geography and cognitive geography, there are also important differences between these disciplines. Cultural geography mainly focuses on the spatial distribution of culture, identity and power dynamics (Anderson et al. 2003). Although researchers recognise that humans are influenced by their environment and vice versa, the cognitive processes that play a role in this influence are not the main object of study in the field. One of the most important characteristics of cognitive geography, on the other hand, is that it maintains that people's behaviour is influenced by their mental representations of their environment more than by the reality of their environment. This means that a person's perception of the distance between two places, for example, has more influence on their willingness to travel between these places than the actual distance. If they falsely believe the distance to be long, this mental representation of distance is more important than the fact that the distance is short. This relationship between mental representations of the environment and behaviour is the central object of study in cognitive geography (Montello 2009, 2018). Other important characteristics of cognitive geography are (1) that the analysis is often disaggregate and focusses on the individual, (2) that the individual and the environment affect each other, and (3) that cognitive geographic research is inter- and multidisciplinary (Montello 2018).

As it is impossible to directly access a person's cognition, cognitive geographers make use of several techniques to access cognition indirectly. One way of doing this is through so-called cognitive distances. Cognitive distances are the geographic distances between two places as estimated by an individual. In the field of cognitive geography, a distinction is made between cognitive and perceptual distances. Perceptual distances can only be estimated when the individual can see the places they have to estimate the distance to, such as a door within the same room or a tree that is visible some distance away. Cognitive distances, on the other hand, are used in situations in which the place that the individual has to estimate the distance towards is obscured from view (Montello 1991), for example when estimating the distance to a building the participant is familiar with but that is not currently visible or the distance between two cities. It is important to note this distinction, as the distance estimates in our study are *perceptual* when dialectology is concerned (participants were able to listen to audio recordings), but *cognitive* when geography is concerned (participants were not able to see the locations). Cognitive distances can be used to quantify whether an individual overestimates or underestimates actual geographical distance. Over- or underestimation of geographic distance is dependent on many factors which are partly environmental, partly individual, and partly a mixture of the two (Qi & Shu 2006). The most important of the individual factors is familiarity, with an increase in familiarity causing a decrease in cognitive distance (Montello 1991). For example, a study in which participants at different levels of familiarity with Sydney were tasked with providing distance estimates for locations within Sydney found that distance is generally overestimated, but that this overestimation is reduced by familiarity to a location (Day 1976). Related but slightly different is the notion that travel times are most predictive of cognitive distance, and that cognitive distance is mostly estimated on the basis of (perceived) travel times (MacEachren 1980). Other factors such as age, gender (Lawton 2018) and individual differences in spatial cognition also play a role (Jenkins & Walmsley 1992). It is important to note that, although there are studies and theories on the underlying factors that influence or stand at the basis of cognitive distance, there is no real way of understanding what happens cognitively when someone makes an estimate of distance as any of the aforementioned (or other) factors could play a role.

Although there is no research known to the authors in which the mental representations of space in the individual are used to explain linguistic phenomena, research in the opposite direction, i.e., using methods and frameworks from linguistics for explaining cognitive geographic phenomena, does exist. Typical examples would be the use of discourse analytic methods to assess wayfinding

strategies and route descriptions (e.g., Hölscher et al. 2011) or the improvement of navigational software by using more human language (e.g., Baltaretu et al. 2015). In these cases, language is not the object of research, but techniques from linguistics are used to access cognition which is then translated into conclusions about spatial cognition (see Tenbrink 2020).

## 2.4 The dialect landscape of Groningen and northern Drenthe

The area under investigation, i.e., the province of Groningen and the northern part of the province of Drenthe, is located in the Low Saxon language area in the Netherlands. Low Saxon forms a dialect continuum that stretches from the East of the Netherlands into Germany and a small part of Denmark (Gooskens & Kürschner 2009). In the Netherlands, Low Saxon is a minority language that is recognised under part two of the European Charter for Regional or Minority Languages (*ECRML* 1998). According to the most recent survey study, Low Saxon is spoken by roughly 75% of the population in the Low Saxon areas in the Netherlands (i.e., the provinces of Groningen, Drenthe and Overijssel, the municipalities of Weststellingwerf and Ooststellingwerf in the province of Fryslân, and the areas of Achterhoek and Veluwe in the province of Gelderland; Bloemhoff 2005). The results from this study are, however, dated (the survey was conducted in 2003) and have been challenged as overestimations (Versloot 2020, Goeman & Jongenburger 2009). See Figure 1 for the area of the Netherlands in which Low Saxon is spoken.

Whereas most dialects spoken in the Netherlands as well as Standard Dutch are descendants of Old Low Franconian (De Schutter 1994), the Low Saxon dialects descend from Old Saxon (Bloemhoff et al. 2008). This historical difference is still reflected in the way the dialects cluster nowadays, with Low Franconian and Low Saxon forming distinct clusters (Nerbonne et al. 1996) and with Low Saxon displaying a relatively large distance towards Standard Dutch (Wieling et al. 2011). Speakers of Low Saxon in the Netherlands generally also speak (Standard) Dutch, as it is necessary to participate in all aspects of society such as education and interaction with more formal organisations. This has also contributed to the stark age-grading in speakers, with a relatively small number of young people speaking Low Saxon dialects (Bloemhoff 2008).

The Low Saxon dialect spoken in most of the area under investigation is Gronings. This label includes the dialects spoken in the northern part of Drenthe, as they are more closely related to dialects spoken in Groningen than the dialects spoken in the rest of Drenthe (Bloemhoff et al. 2020). Gronings is different from other dialects in the Netherlandic Low Saxon language area because of its Frisian



Figure 1: Map of the Low Saxon language area in the Netherlands (from Bloemhoff et al. 2020)

substrate (Reker 2008), which is as of now most visible in diminutive formation with /k/ rather than /t/ after vowels, labials, /s/ and /r/ but without umlaut, as opposed to dialects in Drenthe and Twente, respectively (cf. van Bree 2017), and because it did not (or only to a lesser degree) undergo Westphalian breaking, a sound change involving diphthongisation that resulted in a variety of pronunciations in the Westphalian dialects (Bloemhoff et al. 2020). Some of the most characteristic features of Gronings – although not present in all varieties – are the use of the diphthongs /aɪ/ and /əʊ/ (for example, in the words /laɪf/ ‘sweet’ (of a person) and /bəʊk/ ‘book’) where other Low Saxon variants have monophthongs such as /i/ or /e:/ (/lɪf/ or /le:f/) and /u/ or /o:/ (/buk/ or /bo:k/) respectively (Reker 2008).

## 2.5 Research question and hypothesis

This study investigates whether the framework and methods of cognitive geography can be usefully employed in dialectological research. Concretely, this is done by answering the question of whether including cognitive distance, in addition to geographic distance, leads to better predictions of perceptual dialect distance in the Gronings dialect area. Based on the cognitive geographic theory that mental representations of space are more influential in determining behaviour than space itself, we hypothesise that cognitive distances will add to geographic distance in predicting perceptual dialect distance. More specifically, we expect that an increase in cognitive distance leads to an increase in perceptual dialect distance (regardless of geographic distance), as cognitive distances can also represent the mental distance that a person feels towards a place which in turn might be tied to the distance they feel towards a dialect. Thus, the study sheds more light on the individual experience of space and how this is connected to the experience of language, drawing a parallel between perceptual dialectology and cognitive geography.

## 3 Method

### 3.1 Participants

A total of 1,034 participants from the provinces of Drenthe and Groningen participated in the study through an online survey. Participants who did not provide perceptual dialect distance estimates were removed, as well as participants who either indicated in the control question that they were not able to hear the dialect recordings sufficiently to provide the estimates or who did not answer the control question at all. This resulted in 789 remaining participants. Most of the participants were self-reported dialect speakers, with 96.2% of the participants indicating that they were able to speak their local dialect to at least some degree, and 62.2% of the participants indicating that they were able to effortlessly participate in any kind of conversation in their dialect. All but one of the participants indicated that they could understand their local dialect to at least some degree, and 74.4% indicated that they were able to effortlessly understand anything in the dialect, even when spoken at a fast pace. The age of the participants ranged from 12 to 95, with a mean age of 49 and a standard deviation of 16. A total of 55.1% of the participants identified as women, 43.9% as men and 1% preferred not to disclose their gender identity or identified as a non-binary gender.

### **3.2 Materials**

The survey was implemented in Qualtrics (2005) and consisted of three sections: demographic questions (as discussed under the previous section, in addition to the location in which the participants grew up), cognitive distance estimates, and perceptual dialect distance estimates. The assessment of dialect proficiency (speaking skill and listening skill) was done through self-assessment: participants were asked to rate how well they were able to speak and understand their local dialect on a seven-point Likert scale. For the cognitive distance questions, participants were asked to estimate the distance from the place in which they grew up to seven places in Drenthe and Groningen (namely, Eelde, Finsterwolde, Grijpskerk, Onstwedde, Slochteren, Uithuizen, and Zevenhuizen). The speaker locations as well as the locations in which participants grew up can be found in Figure 2. These seven locations were selected because dialect recordings with identical speech from dialect speakers from these locations were already available from an earlier study on dialect change in the continental part of the Dutch language area (Heeringa & Hinskens 2014). The recordings we used were the first five sentences of a text read aloud by one male speaker over 60 years of age per location, resulting in recordings of between 13 and 25 seconds. These texts were created by two to four of these speakers per location who wrote a consensus translation of a Dutch text, accompanied by movie stills. Because of the nature of the task, the recordings contained substantial lexical and stylistic variation. For each location, a recording was available of an older male and a younger female speaker. We chose to use the recordings from the older male speakers as these were further removed from the Dutch standard language (Heeringa & Hinskens 2014) and we therefore expected them to contain more readily identifiable dialect features. For the perceptual dialect distance estimates, participants were asked to listen to the associated dialect recordings (without knowing that these recordings came from the seven places mentioned earlier) and indicate to what degree they were similar to the dialect that was spoken in the place in which they grew up on a seven-point Likert scale, ranging from 1 (not alike at all) to 7 (very alike). Participants were presented with each recording once, but could replay them as often as necessary and had unlimited time to provide their estimate of similarity. The order of the recordings was the same for every participant, namely alphabetically arranged by speaker location. The survey was conducted in Dutch.

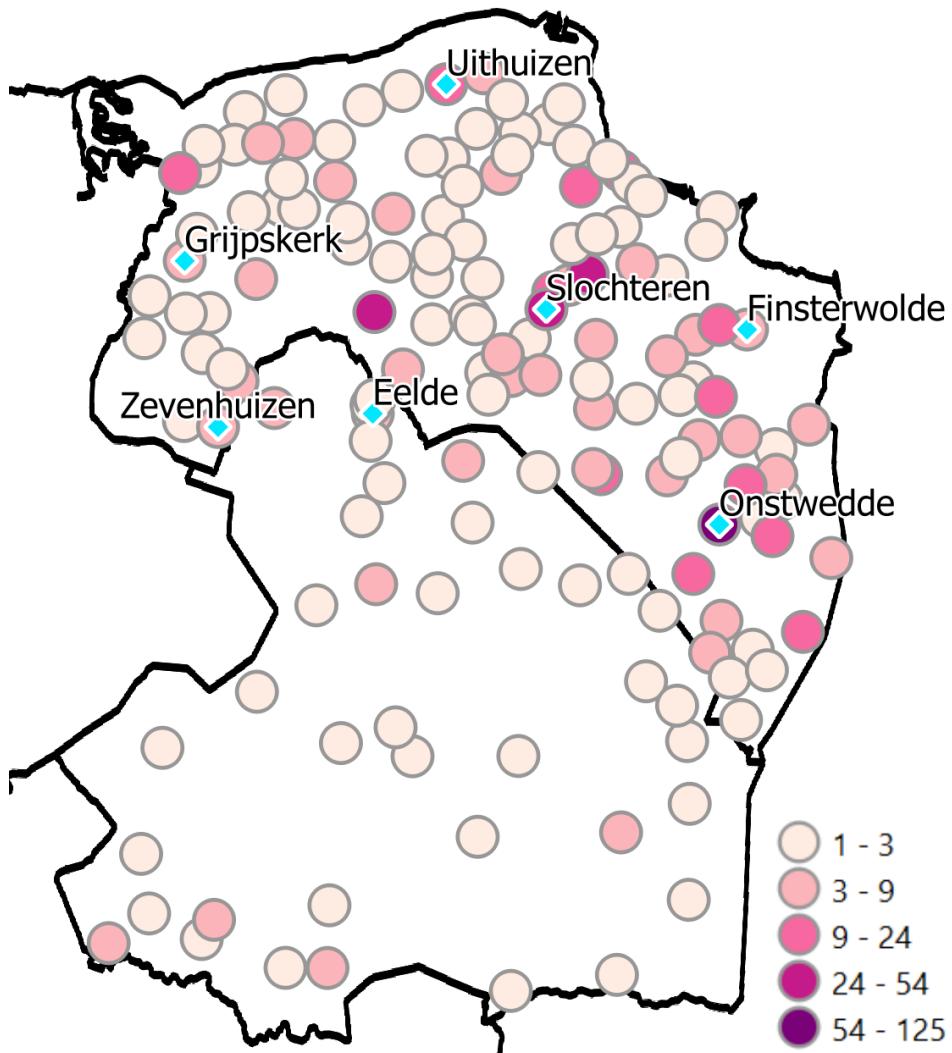


Figure 2: Map of Groningen and Drenthe with the speaker locations (blue diamonds) and the participant locations (pink circles, with darker colours indicating a larger number of participants)

### **3.3 Procedure**

The survey was shared on the online platform of Centrum Groninger Taal & Cultuur, a dialect organisation for Gronings, through social media and through the researchers' own networks. The topic of the study was not fully disclosed to participants. Instead they were informed that the study investigated dialects and regional languages in Drenthe and Groningen and that they would be asked to answer questions about locations in this area and listen to recordings. No compensation was provided for participating in the study. Ethical approval for this study was obtained from the Research Ethics Committee of the Faculty of Arts at the University of Groningen (CETO, reference number 81837041).

### **3.4 Analysis**

A linear mixed-effects regression model was fitted to the data, using the `lme4` package (Bates et al. 2015) in R (R Core Team 2020). Three variables that possibly could account for random variation in the data were included in the model. Subject was included as a random-effect factor as the participants in this study were sampled from a larger possible set of participants. Subject location largely overlapped with subject, but was included as a separate random-effect factor as the set of locations from which a subset was taken for this study was not the same as the population from which the participants were selected. Finally, speaker location was included as a random-effect factor because the dialect recordings were selected from a potentially larger set of locations.<sup>1</sup> As the effect that cognitive distance had on the estimation of dialect distance differed per subject, a by-subject random slope for cognitive distance was added. Similarly, the effect of geographic distance on the estimate of dialect distance differed per speaker location, so a by-speaker location random slope for geographic distance was added. The numeric variables (cognitive distance, geographic distance and birth year) were centered and scaled (i.e., z-transformed) to provide a better model fit and facilitate interpretation as the coefficients then reflect the relative size of the effects. Model comparison (using likelihood ratio tests) was used to identify which fixed and random effects it was necessary to include.

---

<sup>1</sup>Note that including speaker location as a random effect is done here because the nature of this study is methodological. If the goal of the study was to investigate differences between the dialects of these locations, they would be included as a fixed effect.

## 4 Results

### 4.1 Hypothesis testing

The model we used to test our hypothesis consisted of the fixed effects cognitive distance and geographic distance.<sup>2</sup> As indicated before, we took the structural variability of our data into account by including a random intercept for subject, subject location and speaker location, a by-subject random slope for cognitive distance and a by-speaker location random slope for geographic distance. All of these random effects were necessary, as they improved the model according to the likelihood ratio tests. This model thus tested the hypothesis that cognitive distance can partly (i.e., in addition to geographical distance) predict perceptual dialect distance. The fixed effect coefficients for this model can be found in Table 1, and the random effects structure can be found in Table 2. Following Nerbonne (2010), we investigated whether the logarithm of the geographic distance was a better predictor of perceptual dialect distances than assuming a linear relationship, but this did not appear to be the case. The residuals of the final model followed a normal distribution, and autocorrelation and heteroskedasticity fell within normal bounds. A trimmed version of the model in which outliers (i.e., residuals that were more than 2.5 standard deviations away from the mean) were removed showed that the results were robust, i.e., not caused by outliers.

Table 1: Fixed effect coefficients of a minimal model for predicting perceptual dialect distance

	Estimate	Std. Error	t	p
Intercept	3.58283	0.28158	12.724	$8.04 \times 10^{-6}$ ***
Geographic distance	0.52906	0.07392	7.157	$1.49 \times 10^{-2}$ *
Cognitive distance	0.09792	0.04016	2.438	$5.85 \times 10^{-5}$ ***

The hypothesis-testing model indicates that the predictor with the strongest effect on perceptual dialect distance (as estimated by the participants listening to dialect recordings) is geographic distance, as it has the highest estimate out of both  $z$ -transformed predictors. Larger geographical distances between places

<sup>2</sup>We also ran two versions of this model containing only cognitive distance and only geographic distance in order to test whether it was really the combination of the two factors that was effective and not just one of them. These models did not outperform the regular hypothesis model.

Table 2: Random effect structure of a minimal model for predicting perceptual dialect distance

Random-effect factor	Intercept/Slopes	Variance	Std. Dev.	Corr.
Subject	Intercept	0.56921	0.7545	0.38
	Cognitive distance	0.06672	0.2583	
Sbj. location	Intercept	0.17954	0.4237	
Spk. location	Intercept	0.53213	0.7295	0.90
	Geographic distance	0.02684	0.1638	
Residual		1.73765	1.3182	

are associated with larger perceptual dialect distances. The effect of cognitive distance is less strong, but goes in the same direction: the further away participants believe a place is, the larger the perceptual dialect distance.

## 4.2 Exploratory analysis

For the exploratory model, predictors were added step by step and were only included in consecutive models if they significantly improved the model, which was evaluated using model comparison. The set of potential predictors that were considered were cognitive distance, geographic distance, participant listening skill, participant speaking skill, participant gender, participant birth year, interactions between cognitive distance and geographic distance, between cognitive distance and listening skill, between cognitive distance and speaking skill, between cognitive distance and gender, between gender and listening skill, between gender and speaking skill as well as the random effects described above. Main effects and interactions were added to the model in this order, based on how likely they were estimated to have an effect following previous literature. In the final model, predictors with low  $t$ -values (i.e., lower than 2) were re-evaluated and removed from the model if model comparison indicated that including each of these predictors did not significantly improve the model compared to a model without that predictor. The fixed effect coefficients for this model can be found in Table 3, and the random effects structure can be found in Table 4. In the final model, the residuals followed a normal distribution and both autocorrelation and heteroskedasticity fell within normal bounds. A trimmed version of the model in which outliers that were more than 2.5 standard deviations away from the mean were removed was run as well in order to assess whether the model was carried by outliers, which was not the case.

Table 3: Fixed effect coefficients of an exploratory model for predicting perceptual dialect distance

	Est.	SE	t	p	
Intercept	5.13280	0.37622	13.643	$2.25 \times 10^{-11}$	***
Cognitive distance	0.10632	0.03998	2.660	0.007907	**
Geographic distance	0.49770	0.07559	6.584	0.000123	***
Listening skill	-0.17891	0.06446	-2.775	0.005647	**
Speaking skill	-0.14201	0.03716	-3.822	0.000144	***
Gender (female vs. male)	-0.18418	0.06643	-2.772	0.005705	**
Cognitive distance	-0.06082	0.02158	-2.818	0.005036	**
Geographic distance					

Table 4: Random effect structure of an exploratory model for predicting perceptual dialect distance

Random-effect factor	Intercept/Slope	Variance	Std. Dev.	Corr.
Subject	Intercept	0.52090	0.7217	
	Cognitive distance	0.05919	0.2433	0.48
Sbj. location	Intercept	0.12852	0.3585	
Spk. location	Intercept	0.54786	0.7402	
	Geographic distance	0.02820	0.1679	0.94
Residual		1.74861	1.3224	

For the main effects, we now find an interaction between geographic distance and cognitive distance. The effect of cognitive distance is positive, but it is strongest when geographic distance is short, i.e., when the speaker location and the participant location are closer together. Figure 3 visualizes this pattern.

In addition, gender contributes to the model, with female participants providing lower perceptual dialect distances. Furthermore, an increase in speaking skill and listening skill (as gauged through self-assessment by the participants) corresponds to a decrease in perceived dialect distances.

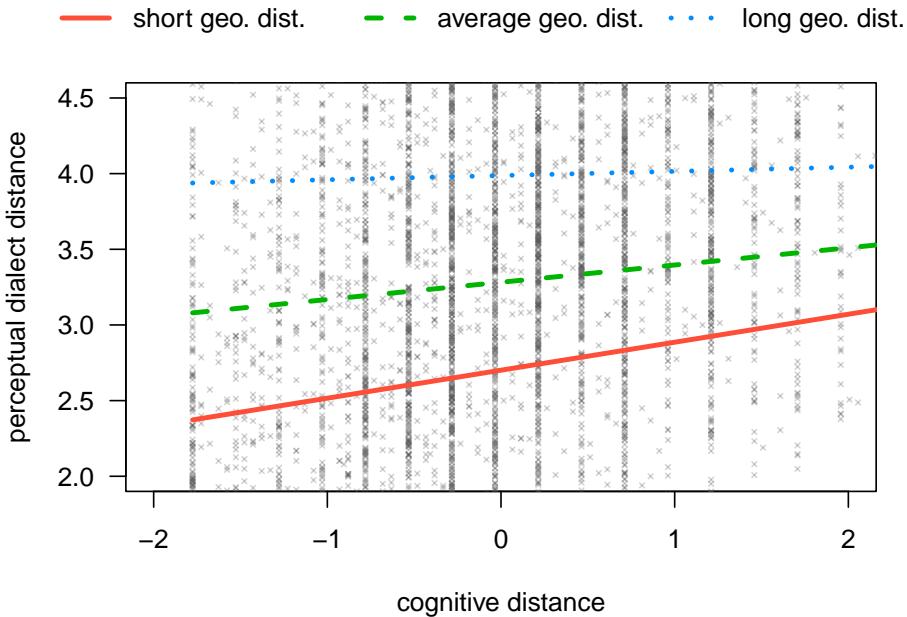


Figure 3: The interaction between cognitive distance and geographic distance in an exploratory model for predicting perceptual dialect distance. The crosses indicate the partial residuals.

## 5 Discussion

### 5.1 General

The research question of this study was whether cognitive distances could be used, in addition to geographic distance, to explain perceptual dialect distances. In our analysis, we first built a model to test our basic hypothesis and consecutively built an exploratory model to investigate other factors that might contribute to predicting perceptual dialect distances. The hypothesis-testing model revealed that cognitive distance does contribute to predicting perceptual dialect distance, as an increase in cognitive distance was related to an increase in perceptual dialect distance. Our initial hypothesis was thus confirmed. However, the effect of geographic distance was more predictive of differences in perceptual dialect distance than cognitive distance was. In contrast to earlier research assessing the relationship between geographic and dialect distances in large datasets (Heeringa et al. 2002), the logarithm of geographic distance did not provide a

better fitting model than simple linear geographic distance did. This may have been caused by the relatively small geographic area that was under investigation (Heeringa et al. 2007).

Our exploratory analysis revealed an interaction between cognitive and geographic distance, in which the predictive value of cognitive distances was greater for smaller geographic distances. As people are better at predicting distances towards locations they are familiar with (Day 1976, Montello 1991), it is not surprising that cognitive distance is more predictive for locations that participants are closer to geographically. Geographic distance, however, does not necessarily have a direct relation to familiarity, so it could be beneficial for future studies to separately measure the degree of familiarity between participants and the locations. The relationship between cognitive and geographic distance indicates that the use of cognitive distance in dialect research might be especially useful for studying small geographical areas. In this case, the provinces of Groningen and Drenthe, which span a combined 5,640 km<sup>2</sup>, already represent a relatively large geographical area for a study on cognitive distances.

Three additional predictors for perceptual dialect distance were significant in the exploratory model. First, both an increase in speaking skill and listening skill were connected to a lower perceptual dialect distance. An effect of dialect proficiency could unfold in two ways. Either more familiarity with the dialect could lead to lower perceptual distances in general, or more familiarity could lead to a better ability to distinguish the different dialects which in turn leads to higher perceptual distances. In this case, it appears that higher proficiency in each skill leads to lower perceived dialect distances. This could be caused by the composition of our sample which was relatively biased toward highly proficient users of dialect: 62.2% of participants indicated that they were able to effortlessly participate in any kind of conversation in their dialect and 74.4% indicated that they were able to effortlessly understand anything in the dialect, even when spoken at a fast pace. It could be possible that familiarity with the dialect leads to lower perceptual distances up to a certain degree and that only people who are very familiar with the dialects can use this familiarity to better distinguish among them. A more balanced study design with more people who are not proficient or have only limited proficiency in their local dialect could provide more insight in this finding.

Second, women in general provided lower perceptual dialect distances than men. This result might seem unexpected, as men, or more specifically non-mobile, older, rural males or NORMs, are usually seen as more proficient dialect speakers (Chambers & Trudgill 1980), and familiarity might lower the perceptual ratings. Welch two sample *t*-tests revealed that the men in the sample did

indeed report significantly higher speaking and listening skills than the women. However, no significant interaction between gender and either listening skill or speaking skill was found in the exploratory model, making this explanation unlikely. Studies on the effect of listener gender (rather than speaker gender) on dialect classification are sparse, but one study on regional variants in Turkey found that men distinguished more different dialect areas than women did (Demirci 2002). However, as the author mentions, this could also be an effect that is specific to a context in which women have less access to education and social institutions. Future research would have to determine whether robust differences in listener gender exist in estimating perceptual dialect distances.

## **5.2 Limitations**

As the recordings used in this study came from a pre-existing dataset, the choice of locations was rather limited. In the area under investigation, only seven locations were available out of which five were very small, having fewer than 3,000 inhabitants. Therefore, not all participants may have been familiar with every location they were asked to provide a distance estimate for. Although the random intercept for speaker location partly compensates for this, it would be better for future studies to only incorporate locations that are likely to be known to all or most of the participants, for example by conducting a pre-test. Furthermore, an assessment of cultural prominence (Montgomery 2012) could provide additional information on how these locations might be represented in the minds of participants.

Additionally, some participants indicated in the survey that they found it difficult to specify a single place where they grew up, for example because they spent their youth in several locations, or because a large part of their social life took place in a different location than where they lived. Other aspects of mobility that could influence dialect perception or the estimates of cognitive distance are locations where participants did not grow up but where they lived for a significant part of their adult life, locations where they worked, the location(s) where their parents grew up or locations they were otherwise familiar with. These relations to other locations were not captured by the survey used in this study, but could provide interesting additional insight for future studies. Gathering more information on mobility patterns and the familiarity that participants have with different locations (see Jeszenszky et al. 2024) could improve both our understanding of the cognitive distance estimates that participants provided and the perceptual dialect distances they reported.

### 5.3 Broader implications and future research

Within the field of dialectology, space is usually treated as a relatively static variable, rather than as an environment in which people move around and which they experience. We have attempted to incorporate this experience of space in dialectology, in the same way that the experience of language has been incorporated in the field through perceptual dialectology. Earlier endeavours to incorporate techniques from (cultural) geography in dialectological research and perceptual dialectology in particular have greatly improved our understanding of the relationship between language, space and culture. The use of cognitive geographic techniques expands the toolbox of dialectologists by offering new explanations for the perception of language variation in a manner that is cognitively informed.

In our study, cognitive distance estimates served as the quantification of the experience of space. These distances contributed to the prediction of perceptual dialect distance, especially when geographic distance was short. The effect of cognitive distance on perceptual dialect distance in our model displayed a large amount of variation per subject, as cognitive distance is a highly individual measure. Nonetheless, it seems that the aggregate analysis that was used was suitable for our study, as individual and demographic differences between participants could be taken into account. Although there was a clear effect of cognitive distance, it does appear that this type of analysis is especially useful when smaller areas are considered than was done in this study. However, the area under investigation cannot be too small either as there would still need to be a sufficient amount of linguistic diversity in the geographic sense. Perhaps a collection of villages on the border of two language areas or neighbourhoods within a large city would be suitable places for a similar study, as they allow for a relatively high amount of language variation within a geographically small area. In these cases, and especially for a study in an urban environment, the linguistic variants under investigation could also be socially stratified (in addition to their spatial stratification) in order to further assess to what degree the cognitive distances are different from pure geographic distances. Furthermore, an analysis that is more focused on individual linguistic and geographic behaviour could provide more insight into the exact relation between the experience of space and the experience of language.

The results of this study show that the framework and methods of cognitive geography can be usefully employed in the field of (especially perceptual) dialectology. Although the aggregate nature of our study makes it difficult to assess what the relationship between language and space in the mind entails exactly,

we have provided a first glimpse into the possible use of cognitive geography in the field of dialectology. As the current study is methodological in nature, the dialect areas under investigation served as a test case. In future research, this approach can now be used to conduct studies in which the dialects themselves take a more central focus.

## 6 Conclusion

In this study, we have attempted to answer the question of whether the framework and methods from the field of cognitive geography could be usefully employed in dialectological research. This was done through a study that investigated the effect of cognitive distances on perceptual dialect distances in Groningen and Drenthe. The results of this study indicate that it is indeed the case that the framework and methods from cognitive geography can be used in dialectological research. This opens up new directions of research in which the human experience of space is used to explain linguistic phenomena.

## Contributions

Hedwig Sekeres contributed to conceptualisation, methodology, formal analysis, investigation, and writing – original draft. Martijn Wieling contributed to conceptualisation, writing – review & editing, and supervision. Remco Knooihuizen contributed to writing – review & editing.

## References

- Anderson, Kay, Mona Domosh, Steve Pile & Nigel Thrift. 2003. *Handbook of cultural geography*. London: Sage. DOI: 10.4135/9781848608252.
- Bailey, Charles-James N. 1973. *Variation and linguistic theory*. Arlington, VA: Center for Applied Linguistics.
- Baltaretu, Adriana, Emiel Krahmer & Alfons Maes. 2015. Improving route directions: The role of intersection type and visual clutter for spatial reference. *Applied Cognitive Psychology* 29(5). 647–660. DOI: 10.1002/acp.3145.
- Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. DOI: 10.18637/jss.v067.i01.

- Bloemhoff, Henk. 2005. *Taaltelling Nedersaksisch: Een enquête naar het gebruik en de beheersing van het Nedersaksisch in Nederland*. Groningen: Stichting Sasland.
- Bloemhoff, Henk. 2008. Taalsociologische aspecten. In Henk Bloemhoff, Jurjen van der Kooi, Hermann Niebaum & Siemon Reker (eds.), *Handboek Nedersaksische taal- en letterkunde*, 294–323. Assen: Van Gorcum.
- Bloemhoff, Henk, Philomène Bloemhoff-de Bruijn, Jan Nijen Twilhaar, Henk Nijkeuter & Harrie Scholtmeijer. 2020. *Introduction to Low Saxon language and literature*. Assen: Van Gorcum.
- Bloemhoff, Henk, Jurjen van der Kooi, Hermann Niebaum & Siemon Reker (eds.). 2008. *Handboek Nedersaksische taal- en letterkunde*. Assen: Van Gorcum.
- Britain, David. 2011. Conceptualizations of geographic space in linguistics. In Roland Kehrein, Alfred Lameli & Stefan Rabanus (eds.), *Language and space: An international handbook of linguistic variation*, vol. 2: Language mapping (Handbooks of Linguistics and Communication Science (HSK) 30/2), 69–102. Berlin: de Gruyter Mouton.
- Britain, David. 2013. Space, diffusion and mobility. In Jack K. Chambers, Peter Trudgill & Natalie Schilling (eds.), *The handbook of language variation and change*, 469–500. Hoboken: Blackwell. DOI: 10.1002/9781118335598.ch22.
- Chambers, Jack K. & Peter Trudgill. 1980. *Dialectology*. Cambridge: Cambridge University Press.
- Day, R. A. 1976. Urban distance cognition: Review and contribution. *Australian Geographer* 13(3). 193–200. DOI: 10.1080/00049187608702693.
- De Schutter, Georges. 1994. Dutch. In Ekkehard König & Johan van der Auwera (eds.), *The Germanic languages*, 439–477. London: Routledge.
- Demirci, Mahide. 2002. Gender differences in the perception of Turkish regional dialects. In Dennis R. Preston & Daniel Long (eds.), *Handbook of perceptual dialectology*, vol. 2, 41–50. Amsterdam: John Benjamins.
- ECRML. 1998. *Europees Handvest voor Regionale Talen of Talen van Minderheden, Straatsburg*, 05-11-1992. <https://wetten.overheid.nl/BWBV0001223/1998-03-01>.
- Goeman, Ton & Willy Jongenburger. 2009. Dimensions and determinants of dialect use in the Netherlands at the individual and regional levels at the end of the twentieth century. *International Journal of the Sociology of Language* 2009(196-197). 31–72. DOI: 10.1515/ijsl.2009.016.
- Gooskens, Charlotte. 2004. Norwegian dialect distances geographically explained. In Britt-Louise Gunnarson, Lena Bergström, Gerd Eklund, Staffan Fridell, Lise H. Hansen, Angela Karstadt, Bengt Nordberg, Eva Sundgren & Mats Thelander (eds.), *Language variation in Europe: Papers from the Second*

- International Conference on Language Variation in Europe, ICLAVE 2 Uppsala University, Sweden, June 12-14, 2003, 195–206.* Uppsala.
- Gooskens, Charlotte & Sebastian Kürschner. 2009. Cross-border intelligibility: On the intelligibility of Low German among speakers of Danish and Dutch. In Alexandra N. Lenz, Charlotte Gooskens & Siemon Reker (eds.), *Low Saxon dialects across borders: Niedersächsische Dialekte über Grenzen hinweg*, 273–295. Stuttgart: Steiner.
- Goossens, Jan. 1977. *Geschiedenis van de Nederlandse dialectstudie*. In Dirk Miente Bakker & Gerardus Rutgerus Wilhelmus Dibbets (eds.), *Geschiedenis van de Nederlandse taalkunde*, 285–311. Den Bosch: Malmberg.
- Gould, Peter & Rodney White. 1974. *Mental maps*. Harmondsworth: Penguin.
- Heeringa, Wilbert & Frans Hinskens. 2014. Convergence between dialect varieties and dialect groups in the Dutch language area. In Benedikt Szemrecsanyi & Bernhard Wälchli (eds.), *Aggregating dialectology, typology, and register analysis: Linguistic variation in text and speech*, 26–52. Berlin: de Gruyter.
- Heeringa, Wilbert, John Nerbonne & Peter Kleiweg. 2002. Validating dialect comparison methods. In Wolfgang Gaul & Gunter Ritter (eds.), *Classification, automation, and new media* (Studies in Classification, Data Analysis, and Knowledge Organization), 445–452. Berlin: Springer Verlag. DOI: [10.1007/978-3-642-55991-4\\_48](https://doi.org/10.1007/978-3-642-55991-4_48).
- Heeringa, Wilbert, John Nerbonne & Peter Kleiweg. 2007. Geographic distributions of linguistic variation reflect dynamics of differentiation. In Sam Featherston & Wolfgang Sternefeld (eds.), *Roots: Linguistics in search of its evidential base*, 267–297. Berlin, New York: de Gruyter Mouton. DOI: [10.1515/9783110198621.267](https://doi.org/10.1515/9783110198621.267).
- Hölscher, Christoph, Thora Tenbrink & Jan M. Wiener. 2011. Would you follow your own route description? Cognitive strategies in urban route planning. *Cognition* 121(2). 228–247. DOI: [10.1016/j.cognition.2011.06.005](https://doi.org/10.1016/j.cognition.2011.06.005).
- Jenkins, John Michael & Dennis James Walmsley. 1992. Cognitive distance: A neglected issue in travel behavior. *Journal of Travel Research* 31(1). 24–29. DOI: [10.1177/004728759203100106](https://doi.org/10.1177/004728759203100106).
- Jeszczyszky, Péter, Yoshinobu Hikosaka, Satoshi Imamura & Keiji Yano. 2019. Japanese lexical variation explained by spatial contact patterns. *ISPRS International Journal of Geo-Information* 8(9). DOI: [10.3390/ijgi8090400](https://doi.org/10.3390/ijgi8090400).
- Jeszczyszky, Péter, Carina Steiner & Adrian Leemann. 2024. Effects of mobility on dialect change: Introducing the Linguistic Mobility Index. *PLoS ONE* 19(4). DOI: [10.1371/journal.pone.0300735](https://doi.org/10.1371/journal.pone.0300735).
- Labov, William. 1965. On the mechanisms of linguistic change. *Georgetown Monographs on Language and Linguistics* 18. 91–114.

- Lawton, Carol A. 2018. Sex and gender in geographic behavior and cognition. In Daniel R. Montello (ed.), *Handbook of behavioral and cognitive geography*, 247–259. Cheltenham, UK: Edward Elgar.
- Lynch, Kevin. 1964. *The image of the city*. Cambridge, MA: MIT Press.
- MacEachren, Alan M. 1980. Travel time as the basis of cognitive distance. *The Professional Geographer* 32(1). 30–36. DOI: 10.1111/j.0033-0124.1980.00030.x.
- Montello, Daniel R. 1991. The measurement of cognitive distance: Methods and construct validity. *Journal of Environmental Psychology* 11(2). 101–122. DOI: 10.1016/S0272-4944(05)80071-4.
- Montello, Daniel R. 2009. Cognitive geography. In Robin Kitchin & Nigel Thrift (eds.), *International encyclopedia of human geography*, 160–166. Amsterdam: Elsevier.
- Montello, Daniel R. 2018. Behavioral and cognitive geography: Introduction and overview. In Daniel R. Montello (ed.), *Handbook of behavioral and cognitive geography*, 3–15. Cheltenham, UK: Edward Elgar.
- Montgomery, Chris. 2012. The effect of proximity in perceptual dialectology. *Journal of Sociolinguistics* 16(5). 638–668. DOI: 10.1111/josl.12003.
- Montgomery, Chris & Joan Beal. 2011. Perceptual dialectology. In Warren Maguire & April McMahon (eds.), *Analysing variation in English*, 121–148. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511976360.007.
- Nerbonne, John. 2010. Measuring the diffusion of linguistic change. *Philosophical Transactions: Biological Sciences* 365(1559). 3821–3828. <https://www.jstor.org/stable/20789197>.
- Nerbonne, John, Wilbert Heeringa, Erik van den Hout, Peter Kooi, Simone Otten & Willem van de Vis. 1996. Phonetic distance between Dutch dialects. In Gert Durieux, Walter Daelemans & Steven Gillis (eds.), *CLIN VI, Papers from the sixth CLIN meeting*, 185–202. Antwerp.
- Nerbonne, John, Ilse van Gemert & Wilbert Heeringa. 2005. A dialectometric view of linguistic “gravity”. *Rijksuniversiteit Groningen*.
- Portugali, Juval. 2018. History and theoretical perspectives of behavioral and cognitive geography. In Daniel R. Montello (ed.), *Handbook of behavioral and cognitive geography*, 16–38. Cheltenham, UK: Edward Elgar.
- Preston, Dennis R. 1981. Perceptual dialectology: Mental maps of United States dialects from a Hawaiian perspective (summary). In Henry Warkentyne (ed.), *Papers from the Fourth International Conference on Methods in Dialectology*, 192–198. Vancouver: Department of Linguistics, University of British Columbia.
- Preston, Dennis R. 1999. *Handbook of perceptual dialectology*, vol. 1. Amsterdam: John Benjamins.

- Preston, Dennis R. 2010. Mapping the geolinguistic spaces of the brain. In Alfred Lameli, Roland Kehrein & Stefan Rabanus (eds.), *Language and space: An international handbook of linguistic variation*, vol. 2: Language mapping (Handbooks of Linguistics and Communication Science (HSK) 30/2), 121–153. Berlin, New York: de Gruyter Mouton.
- Qi, Cuihong & Hong Shu. 2006. Formal properties of cognitive distance in geographical space. In *16th International Conference on Artificial Reality and Telexistence–Workshops (ICAT’06)*, 408–412. DOI: 10.1109/ICAT.2006.66.
- Qualtrics. 2005. Qualtrics. Provo, Utah. [www.qualtrics.com](http://www.qualtrics.com).
- R Core Team. 2020. *R: A language and environment for statistical computing*. <https://www.r-project.org/>.
- Rabanus, Stefan. 2017. Dialect maps. In Charles Boberg, John Nerbonne & Dominic Watt (eds.), *The handbook of dialectology*, 348–367. Hoboken, NJ: John Wiley & Sons. DOI: 10.1002/9781118827628.ch20.
- Reker, Siemon. 2008. Talige beschrijving van de regio's: Groningen. In Henk Bloemhoff, Jurjen van der Kooi, Hermann Niebaum & Siemon Reker (eds.), *Handboek Nedersaksische taal- en letterkunde*, 157–165. Assen: Van Gorcum.
- Sibata, Takesi. 1999. Consciousness of dialect boundaries. In Dennis R. Preston (ed.), *Handbook of perceptual dialectology*, 39–62. Amsterdam: John Benjamins.
- Stanford, James N. 2012. One size fits all? Dialectometry in a small clan-based indigenous society. *Language Variation and Change* 24(2). 247–278. DOI: 10.1017/S0954394512000087.
- Tenbrink, Thora. 2020. *Cognitive discourse analysis: An introduction*. Cambridge: Cambridge University Press. DOI: 10.1017/978108525176.
- Trudgill, Peter. 1974. Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. *Language in Society* 2(2). 215–246. DOI: 10.1017/S0047404500004358.
- van Bree, Cor. 2017. The Frisian substrate beneath the Groningen dialect. *Amsterdamse Beiträge zur älteren Germanistik* 77(1-2). 65–87. DOI: 10.1163/18756719-12340067.
- van Gemert, Ilse. 2002. *Het geografisch verklaren van dialectafstanden met een GIS*. University of Groningen. (MA thesis).
- Versloot, Arjen. 2020. Streekaalood in de Lage Landen. *Taal en Tongval* 72(1). 7–16. DOI: 10.5117/TET2020.1.VERS.
- Weijnen, Antonius Angelus. 1946. De grenzen tussen de Oost-Noordbrabantse dialecten onderling. In Antonius Angelus Weijnen, J. Renders & Jacques van Ginneken (eds.), *Oost-Noordbrabantse dialectproblemen 8* (Bijdragen en Mededelingen der Dialectencommissie van de Koninklijke Nederlandse Akademie

- van Wetenschappen te Amsterdam), 1–15. Amsterdam: Noord Hollandsche Uitgevers Maatschappij.
- Wieling, Martijn, John Nerbonne & R. Harald Baayen. 2011. Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLoS ONE* 6(9). DOI: 10.1371/journal.pone.0023613.



## Part II

# Corpus-based studies and dialect change



# Chapter 5

## A directional shift in linguistic change: A longitudinal study on English-speaking expatriates in Japan

Keiko Hirano

The University of Kitakyushu, Japan

This paper attempts to demonstrate that the direction of linguistic change in a dialect contact environment can shift over time. The analysis reports on the linguistic change in an English-speaking expatriate community in Japan in which dialect contact (Britain 2018, Trudgill 1986, 2004) occurs among different varieties of English. Speakers' choice of possessive verbs (*have got*, *have*, and *got*) and obligatory verbs (*must*, *have got to*, *have to*, and *got to*) is examined. This longitudinal study is based on three sets of linguistic data: (1) a corpus of English-language conversations collected in 2000 from young British and American English speakers who had recently arrived in Japan [Data 1], (2) a corpus collected in 2001 from the same speakers after they had lived in Japan for a year [Data 2], and (3) a corpus collected in and after 2018 from British and Americans who had worked and lived in Japan for an average of 20 years [Data 3]. The analysis of the use of possessive verbs suggests that the British English speakers used more typically "British" grammatical constructions (*have got*) in Data 2 (one year after their arrival in Japan), while the American English speakers maintained their use of more typically "American" constructions (*have* and *got*) in Data 2. However, the analysis of Data 3 (after living in Japan for an average of 20 years) suggests an alteration in the direction of this linguistic change. Both the British and American speakers adopted the use of verbs that had strong associations with the other nationality's English style. A similar tendency was observed in terms of the choice of verbs of obligation. These results indicate that the direction of linguistic change in a dialect contact environment is not always unidirectional but may shift over the long term. These linguistic changes appear to be the outcome of "long-term accommodation" (Trudgill 1986) in which speakers adopt certain linguistic features in another variety of the same language and incorporate these features permanently.



## 1 Introduction

This paper attempts to demonstrate that the direction of linguistic change in a dialect contact environment can shift over time. This provisional analysis reports on linguistic change that occurs in an English-speaking expatriate community in Japan in which dialect contact (Britain 2018, Trudgill 1986, 2004) among English varieties occurs by comparing corpus data from 2000 (Data 1) and 2001 (Data 2) (Hirano 2013, 2016, Hirano & Britain 2016, 2020) with more recent data (Data 3). Speakers' choice of possessive verbs (*have got*, *have*, and *got*) (Hirano 2016, Hirano & Britain 2020), as indicated in example (1), and obligatory verbs (*must*, *have got to*, *have to*, and *got to*) (Hirano & Britain 2016, 2020), as indicated in example (2), are examined. From a total of 26 hours of speech collected from 40 native speakers of English (NSsE) from the United Kingdom (UK) and the United States (US), 820 tokens of verbs of possession and 379 tokens of verbs of obligation were extracted.

- (1)
  - a. I've got an elder sister.
  - b. I *have* an apartment.
  - c. You *got* the West Coast beaches for surfing.
- (2)
  - a. We *must* do it like every ... three or four months.
  - b. I've *got to* go to school.
  - c. You *have to* have a steady hand.
  - d. You *got to* start from somewhere.

The analysis revealed that the directions of linguistic change from Data 1 to Data 2 and those from Data 2 to Data 3 were not identical. This study reveals that a linguistic change which takes place in a long-term dialect contact environment does not necessarily continue to proceed unidirectionally, but may shift in the course of long-term change.

The discussion section attempts to interpret the results of the analyses using the concept of levelling, one of the processes that occurs when a new dialect is created as a result of dialect contact, and the theory of long-term accommodation. Levelling is the eradication of minority or marked variants in a mixed dialect environment (Britain 2018: 149–150, Trudgill 1986: 98–102). Long-term accommodation occurs as a result of repeated short-term accommodation (Coupland 1984, Giles & Powesland 1975) on a long-term basis towards certain features of specific varieties of the same language (Trudgill 1986: 11–21). This paper explores whether one of these theories can explain the process of the change observed in the use

of verbs of possession and obligation induced by dialect contact with speakers of different English varieties in Japan.

## 2 The community of native speakers of English

The NSsE working in Japan range from those who stay in Japan for one to several years as assistant language teachers on the Japan Exchange and Teaching (JET) Programme, or as English teachers in private language schools, to those who stay for several years or several decades as university teachers and researchers or for the purpose of business. The JET Programme employs young university graduates from overseas on fixed-term contracts, and approximately 5,700 people from 57 countries participated in the JET Programme from 2019 to 2020 (Council of Local Authorities for International Relations 2021). As of December 2019, there were approximately three million foreign residents in Japan, of which over 430,000 were from major English-speaking countries (National Statistics Center of Japan 2021). Here, “major English-speaking countries” implies those countries where English is used as the first language or one of the official languages and from where over 1000 people resided in Japan as of December 2019. They include, in the order of the largest number of residents, the Philippines, the U.S., India, the U.K., Australia, Canada, New Zealand, Nigeria, Singapore, Ghana, Ireland, and South Africa. These residents engage in a wide variety of social interactions and communication with speakers of their own English dialect, speakers of different English varieties, and non-native speakers of English including Japanese; they are constantly in dialect and language contact situations on a daily basis over a long period of time.

## 3 Linguistic variables

### 3.1 Verbs of possession

One of the two linguistic variables for this paper includes the verbs of possession: *have got*, *have*, and *got*. The oldest variants of the possessive verb, *have*, are believed to have their origins in the late tenth century, and *got* is believed to have been added in around the sixteenth century. This resulted in the construction *have got*, which had the same function as *have*. *Got* was the last variant to appear. Presently, *have got* is considered informal, and *got* is considered even more informal. *Got* has been frequently used in American English from the mid-nineteenth century onwards (Kroch 1989: 207–208, Lorenz 2016: 489, Tagliamonte 2003: 532–534, 2013a: 146–147, 2013b: 141–142).

Further, *have got* is characteristic of British English and is used more frequently than *have*; *got* is used less frequently. In the UK, younger speakers tend to use *have got* more frequently than older speakers. On the other hand, *have* and *got* are expressions characteristic of North American English. *Have* is the most frequently used form and *got* is particularly strongly associated with North American English. Further, in North American English, younger people tend to use *have* more often than older people (Jankowski 2016: 25–29, Kroch 1989: 207–208, Tagliamonte 2013a: 146–147, 2013b: 141–142, Tagliamonte et al. 2010: 157–160).

### 3.2 Verbs of obligation

The second linguistic variable is the verbs of obligation: *must*, *have got to*, *have to*, and *got to*. *Must*, the oldest variant of obligatory verbs, is inferred to appear in the Old English period. It is regarded to be formal and used more in written language. The first use of *have to* goes back to the sixteenth century or earlier, but *have got to* and *got to* did not make an appearance until the nineteenth century. However, they were categorised as colloquial and vulgar, considered to be informal, and used mostly in spoken form (Tagliamonte 2013a: 135–136, 2013b: 142, Tagliamonte & D’Arcy 2007: 50–52).

According to certain studies, *have to* is the most frequently used form in British English, while other studies say *have got to* is the most frequently used form. Further, the use of *must* has been reduced and the use of *have to* and *have got to* has been on the rise in both British and North American English. *Have to* is most commonly used in North American English. Younger people show a tendency to use the phrase more frequently than older people in North America. However, *got to* is considered to be typical in American English (Collins 2005: 253–256, Tagliamonte 2013a: 136–138, 2013b: 142–145).

## 4 Methodology

### 4.1 Linguistic data

This longitudinal study of grammatical variation is based on three sets of linguistic data: Data 1, Data 2, and Data 3. Data 1 is a corpus of spontaneous conversations in English collected in the year 2000 from young British and American English speakers who had recently arrived in Japan (Hirano & Britain 2020). Data 2 is a corpus of conversations collected in the year 2001 from the same speakers after they had lived and worked in Japan for a year (Hirano & Britain 2020). Data 3

is a corpus of conversations collected in and after 2018 from different British and American speakers who had lived and worked in Japan for over seven years. Since the data collection for Data 3 is still in progress, this paper only uses the data already collected up until this point in time for the analysis.

For all three data sets, natural and spontaneous conversations between two NSsE from the same country were recorded in their homes or private offices in a relaxed atmosphere for 45 minutes for Data 1 and 2, and for 30 minutes for Data 3. The researcher was not present during their conversations. Typically, the speakers discussed their daily life and work, everyday events and activities, and personal experiences in Japan, and also gossiped about their friends and colleagues. A total of 26 hours of speech, which amounted to approximately 370,000 words, were used for the present study. SPSS version 25 was used for statistical tests.

## 4.2 Speakers

The number of speakers in the current study is presented in Table 1. There are 26 speakers in Data 1 and 2, out of which 15 are British (5 males and 10 females) and 11 Americans (7 males and 4 females). These include 24 assistant language teachers for the JET Programme and two English conversation instructors for private language schools. They were aged between 21 and 32 years at the time of the first data collection, and the average age of all speakers was 23 years. All of them had approximately the same level of education – university or college degree or above. The speakers in Data 3 were six British (all males) who have lived in Japan for 11 years or longer, and eight Americans (six males and two females) who have lived in Japan for seven years or longer. They are all university teachers, except for one British IT engineer. They were aged between 36 and 67 years at the time of the data collection, with average age being 50 years. Moreover, their ages at the beginning of their stay in Japan ranged from 19 to 32 years, with an average age of 27 years. The duration of their stay in Japan ranged from 7 to 37 years, with an average stay of 20 years. In addition, the research fields of the university teachers range from linguistics, literature, and art to history and law. They do not necessarily specialise in English teaching. The speakers for Data 1 and 2 were living in Fukuoka, Saga, and Kumamoto prefectures, and the speakers for Data 3 were residents of Fukuoka prefecture, mainly in Fukuoka and Kitakyushu cities.

It is accurate to state that Data 1 and Data 2 consist of panel data obtained from identical speakers with a gap of one year between the collections, while Data 3 was collected from a different set of speakers. Considering that the speakers included in Data 3 had an average age of 27 upon their arrival in Japan and

an average age of 50 at the time of data collection, with an average duration of 20 years of staying in Japan, a typical speaker from Data 3 must have arrived in Japan a few years before the year 2000 at the age of 27. These facts imply that the speakers in Data 3 have a similar age and length of stay in Japan to what speakers in Data 1 and Data 2 would have had if they had remained in Japan. Thus, the three sets of data appear to be reasonably comparable to one another.

Table 1: Linguistic data and speakers (figures in parentheses present the averages)

	Data 1	Data 2	Data 3
Year of collection	2000	2001	2018–present
Speakers			
UK	15	15	6
US	11	11	8
Current age	21–32 (23)	+1	36–67 (50)
Age on arrival	21–32 (23)	+1	19–32 (27)
Duration in Japan	Just arrived	1 year	7–37 (20)
Place of residence	Fukuoka, Saga, & Kumamoto		Fukuoka

### 4.3 Tokens

A total of 820 tokens of verbs of possession – *have got*, *have*, and *got* – and a total of 379 tokens of verbs of obligation – *must*, *have got to*, *have to*, and *got to* – were extracted from the three sets of data, as depicted in Table 2. Tokens found in negative and interrogative sentences were excluded from the analysis and only the tokens found in affirmative sentences were included. All the tokens were identified manually.

Table 2: Number of tokens

Verbtype	Country	Data 1	Data 2	Data 3	Total
Possession	UK	217	219	45	481
	US	107	124	108	339
	Total	324	343	153	820
Obligation	UK	82	116	15	213
	US	63	60	43	166
	Total	145	176	58	379

## 5 Results

### 5.1 Distribution of variants of verbs of possession and obligation for British speakers

This section presents the results of an analysis of the verbs of possession and obligation for British speakers. Table 3 presents the distribution of verbs of possession by British speakers for Data 1, 2, and 3, and Figure 1 is the graphic form of the table. The analysis of Data 1 reveals that *have got* was the most frequently used form (55%), followed by *have* (42%); *got* was rarely used (3%). The analysis of Data 2 reveals that there was an increased use of *have got* among British speakers – from 55% in Data 1 to 62% in Data 2. Moreover, there was a decreased use of *have* among British speakers – from 42% in Data 1 to 35% in Data 2. These changes in the speakers’ choice of possessive verbs appear to indicate that British speakers were using more typically “British” grammatical construction.

Table 3: Verbs of possession used by British speakers

Variant	Data 1		Data 2		Data 3	
	n	%	n	%	n	%
<i>have got</i>	119	55%	136	62%	22	49%
<i>have</i>	91	42%	76	35%	22	49%
<i>got</i>	7	3%	7	3%	1	2%

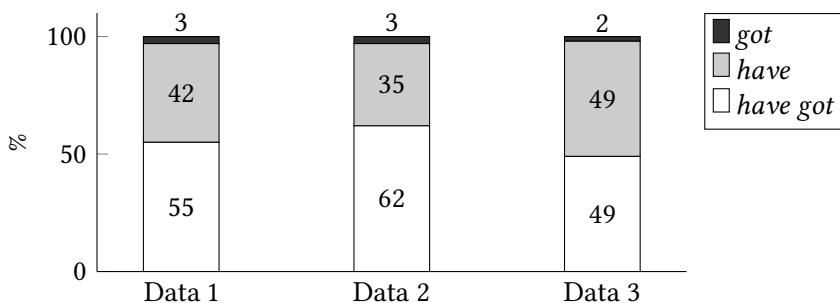


Figure 1: Verbs of possession used by British speakers

However, the analysis of Data 3 suggests a reversal of the direction of this linguistic change. Compared with the results in Data 2, the use of *have got* by the British speakers was lower in Data 3 (62%→49%), while their use of *have* was higher (35%→49%). Further, the use of *have got* in Data 3 is even lower

than that in Data 1 (55%→62%→49%) and the use of *have* is higher than that in Data 1 (42%→35%→49%). Thus, the use of *have got*, which is indicative of British English characteristics, increased temporarily one year after arriving in Japan, but began to decrease in the course of their long-term stay in Japan. In contrast, the use of *have*, which is characteristic of American English, decreased *temporarily* one year after arriving in Japan, but then began to increase.

Table 4: Verbs of obligation used by British speakers (Pearson's Chi-Square (two-sided): \*\* significance at  $p < 0.01$ .)

Variant	Data 1		Data 2		Data 3	
	n	%	n	%	n	%
<i>must</i>	8	10%	2	2%	0	0%
<i>have got to</i>	20	24%	52	45%**	3	20%
<i>have to</i>	51	62%	56	48%	11	73%
<i>got to</i>	3	4%	6	5%	1	7%

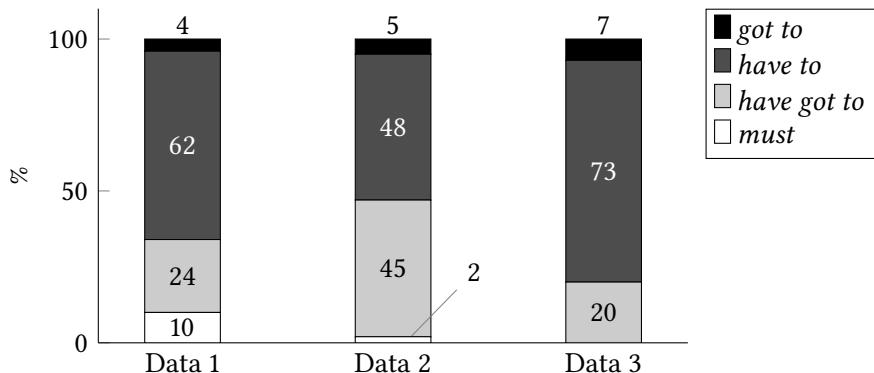


Figure 2: Verbs of obligation used by British speakers

Table 4 presents the distribution of verbs of obligation by British speakers, and Figure 2 is the graphic form of the table. The result indicates the same tendency as verbs of possession. There was an increase in British English tendencies temporarily one year after arriving in Japan, but this decreased during their long-term stay. The analysis of Data 1 reveals that *have to* was the most frequently used form (62%), followed by *have got to* (24%), and *must* (10%), and that *got to* was used only to a small extent (4%). Further, the analysis of Data 2 reveals that

there was a decreased use of *have to* among British speakers from 62% in Data 1 to 48% in Data 2; moreover, there was an increased use of *have got to* from 24% in Data 1 to 45% in Data 2, which is a statistically significant increase. These changes in the speakers' choice of obligatory verbs appear to indicate that the British speakers exhibited more "Britishness" in their linguistic behaviour one year after their arrival in Japan.

However, the analysis of Data 3 indicates a backward movement in the linguistic change. The use of *have got to* by the British speakers was reduced by more than half in Data 3 from that in Data 2 (45%→20%), while their use of *have to* increased by 25%, from 48% to 73%. Compared with the results in Data 1, the use of *have got to* in Data 3 is actually 4% lower (24%→45%→20%) and the use of *have to* is 11% higher (62%→48%→73%). These changes in the speakers' choice of obligatory verbs from Data 1 to Data 2 and from Data 2 to Data 3 appear to indicate again, similar to the directional alteration of the linguistic change in verbs of possession, that the British speakers were inclined to exhibit more British characteristics in their linguistic behaviour after one year; however, the direction of this shift reversed, and they began to adopt the more widely used form in the course of their long-term stay in Japan.

## 5.2 Distribution of variants of verbs of possession and obligation by American speakers

This section presents the distribution of variants of verbs of possession and obligation used by American speakers. Table 5 presents the distribution of verbs of possession used by American speakers for Data 1, 2, and 3, and Figure 3 is the graphic form of the table. The analysis of Data 1 reveals that *have* was the most frequently used form by the American speakers (74%), and there was less frequent usage of *have got* (12%) and *got* (14%). The analysis of Data 2 reveals that there was an increased use of *have* (74%→84%) among the American speakers and a decreased use of *got* (14%→4%) with statistical significance, while their use of *have got* remained unchanged (12%→12%). In other words, Americans neither increased nor decreased their use of the British-English feature, maintaining more typically "American" constructions a year later.

However, when the result of the analysis of Data 3 was compared with that of Data 2, the American speakers' use of *have* was found to be lower (84%→73%) with statistical significance, while their use of *have got* was higher (12%→19%). Thus, the percentage use of *have got*, which is characteristic of British English, did not change from the time they arrived in Japan to one year later but increased

Table 5: Verbs of possession used by American speakers (Pearson's Chi-Square (two-sided): \*\* significance at  $p < 0.01$ ; \* significance at  $p < 0.05$ .)

Variant	Data 1		Data 2		Data 3	
	n	%	n	%	n	%
<i>have got</i>	13	12%	15	12%	20	19%
<i>have</i>	79	74%	104	84%	79	73%*
<i>got</i>	15	14%	5	4%**	9	8%

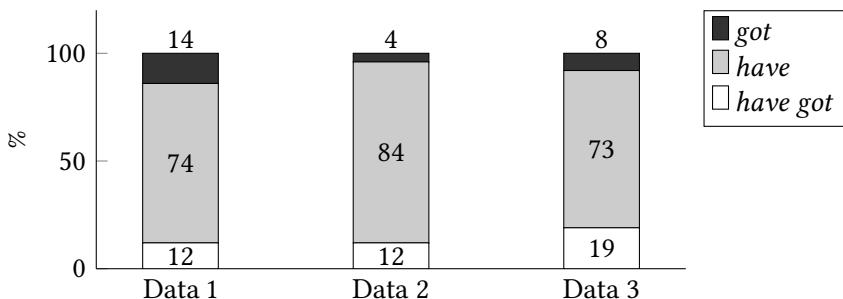


Figure 3: Verbs of possession used by American speakers

during their long-term stay in Japan (12% → 12% → 19%). The combined percentage use of *have* and *got*, which are typical features of American English, also remained unchanged for a year, but subsequently began to decrease (88% → 88% → 81%).

Table 6 presents the distribution of verbs of obligation by American speakers, and Figure 4 is the graphic form of the table. The distribution reveals a slightly different tendency from that of verbs of possession. American speakers adopted British English characteristics only to a small extent one year later, but this increased in the course of their long-term stay in Japan. The analysis of Data 1 indicates that *have to* was the most frequently used form (81%), followed by *got to* (16%) and *must* (3%); however, *have got to* was not used at all. The analysis of Data 2 reveals that there was a slight decrease of the use of *have to* from 81% in Data 1 to 71% in Data 2 among American speakers and an increase in their use of *got to* from 16% in Data 1 to 22% in Data 2. The use of *have got to* increased from null in Data 1 to 5% in Data 2, but the use of *must* slightly reduced from 3% in Data 1 to 2% in Data 2. The combined percentage use of *have to* and *got to*, which are characteristic of American English, slightly decreased from 97% to 93%, while

the combined percentage use of *have got to* and *must*, which are typical features of British English, slightly increased from 3% to 7%.

Table 6: Verbs of obligation used by American speakers

Variant	Data 1		Data 2		Data 3	
	n	%	n	%	n	%
<i>must</i>	2	3%	1	2%	0	0%
<i>have got to</i>	0	0%	3	5%	7	16%
<i>have to</i>	51	81%	43	71%	34	79%
<i>got to</i>	10	16%	13	22%	2	5%

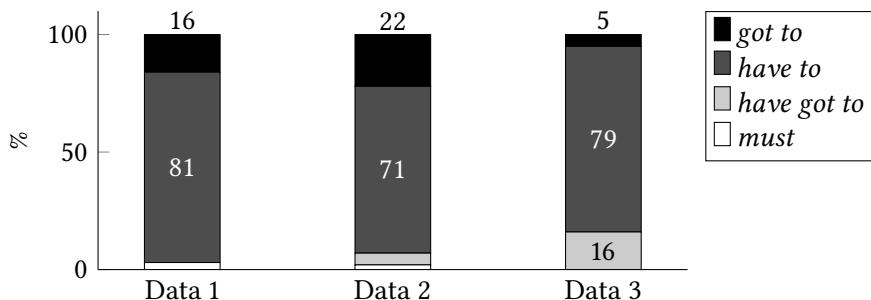


Figure 4: Verbs of obligation used by American speakers

The analysis of Data 3 indicates a further decrease in the use of forms associated with American English and a further increase in the use of forms characteristic of British English. The use of *have to* by the American speakers increased from 71% in Data 2 to 79% in Data 3, but their use of *got to* notably reduced from 22% in Data 2 to 5% in Data 3. The combined percentage use of *have to* and *got to* decreased from 93% in Data 2 to 84% in Data 3. In contrast, the use of *have got to* increased from 5% in Data 2 to 16% in Data 3, although *must* was no longer used. Further, the changes in the speakers' choice of obligatory verbs from Data 1 to Data 2 were somewhat small, while these changes were rather noticeable from Data 2 to Data 3. This indicates that the level of linguistic changes does not remain stable, but may advance further with the progression of time.

## 6 Discussion and conclusion

The analysis and comparison of the three corpora indicate that the linguistic change in the dialect contact environment in Japan does not necessarily take place unidirectionally but may shift over time. One linguistic process that occurs in the formation of new dialects in a dialect contact environment is “levelling” (Britain 2018: 149–150, Trudgill 1986: 98–102) – a phenomenon in which marked or minority variants gradually disappear and unmarked or majority variants tend to survive in the dialect mixture. The nationality of the largest number of NSsE living in Japan is American (National Statistics Center of Japan 2021). If the proportion of American residents among the total number of NSsE in Japan is considered, *have* and *have to* should be the most commonly used form of possessive and obligatory verbs. Indeed, *have* and *have to* are the most frequently used forms among British and American residents in Japan, and the proportions of the use of *have* and *have to* among these speakers have, in fact, slightly increased from Data 1 to Data 3. However, the proportions of the use of *have got* and *have got to* among these speakers in Data 3 did not decrease from that in Data 1. Further, the proportion of the use of *have got* among these speakers in Data 3 was preserved at almost the same level as that in Data 1, and the proportion of the use of *have got to* actually increased in Data 3. Therefore, there was no strong evidence of levelling in terms of the use of verbs of possession and obligation observed in Japan.

One year after arriving in Japan, British speakers showed more British English characteristics in the use of both verbs of possession and obligation. American speakers maintained the same level of American English characteristics in the use of verbs of possession as when they had arrived in Japan, but there was a slight decrease in terms of the use of verbs of obligation. However, there appeared to be a shift in the direction of linguistic change in the course of their long-term stay in Japan. Further, there was an increased use of *have* and *have to* among British speakers – both of which have a strong association with American English – at some point after a year of living in Japan. On the other hand, the American speakers began to increasingly use *have got* and *have got to*, which have a strong association with British English, thereby reducing the effect of the American English characteristic.

These linguistic changes appear to be the outcome of “long-term accommodation” (Trudgill 1986: 11–21) in which speakers adopt certain linguistic features in another variety of the same language and incorporate these features permanently. According to research which investigated long-term speech accommodation in adults who had moved from one dialect region to another, the language of

these adults tended to be modified to eliminate the linguistic features of the old dialect and adopt the linguistic features of the new dialect (Evans & Iverson 2007, Kerswill 1993, Nycz 2011, Omdal 1994, Shockley 1984, Stanford 2008). In Japan, English is not the primary language among the local citizens and, therefore, there is no single target English dialect which the NSsE are expected to assimilate. However, in the Anglophone community in Japan, which is the focus of the current research, both British and American speakers come in contact with different varieties and dialects of the English language while living in Japan on a daily basis. They do not appear to accommodate each other's linguistic features of grammatical aspects immediately after their encounter with other English dialects or during the early stages of the mixing of dialects; however, eventually they appear to begin to adopt each other's linguistic features. As far as the findings of the current study are concerned, the speed of linguistic accommodation differs depending on which variety of English the speakers modify and which variety of English they adopt. The American speakers appeared to begin adapting to the different variety of English earlier than the British speakers in this study.

In conclusion, the present study attempted to address the different stages of long-term accommodation by reporting the outcomes of analyses of three sets of corpora in a real-time study on dialect contact and grammatical variation in a community in which no single dominant variety or regional dialect of English exists. It is important to continue to monitor the change in the linguistic characteristics of individual speakers in order to explore the complexity of the mechanisms of linguistic change in English due to dialect contact in Japan. Once additional linguistic data from a greater variety of speakers are collected, further statistical analyses must be conducted in various possible aspects in order to unfold new findings and validate the outcomes.

## Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 26370494.

## References

- Britain, David. 2018. Dialect contact and new dialect formation. In Charles Boberg, John Nerbonne & Dominic Watt (eds.), *Handbook of dialectology*, 143–158. Oxford: Wiley Blackwell. DOI: 10.1002/9781118827628.

- Collins, Peter C. 2005. The modals and quasi-modals of obligation and necessity in Australian English and other Englishes. *English World-Wide* 26. 249–273. DOI: 10.1075/eww.26.3.02col.
- Council of Local Authorities for International Relations. 2021. *JET programme participant numbers (2019–2020)*. <http://jetprogramme.org/ja/countries/>.
- Coupland, Nikolas. 1984. Accommodation at work: Some phonological data and their implications. *International Journal of the Sociology of Language* 46. 49–70. DOI: 10.1515/ijsl.1984.46.49.
- Evans, Bronwen G. & 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. 3814–3826. DOI: 10.1121/1.2722209.
- Giles, Howard & Peter F. Powesland. 1975. *Speech style and social evaluation*. London: Academic Press.
- Hirano, Keiko. 2013. *Dialect contact and social networks: Language change in an anglophone community in Japan*. Frankfurt: Peter Lang. DOI: 10.3726/978-3-653-02782-2.
- Hirano, Keiko. 2016. *Convergence or divergence? Social network and grammatical variation in a community of expatriate English speakers*. Paper presented at the 21st Sociolinguistics Symposium. University of Murcia, Spain.
- Hirano, Keiko & David Britain. 2016. Accommodation, dialect contact and grammatical variation: Verbs of obligation in the anglophone community in Japan. In Olga Timofeeva, Anne-Christine Gardner, Alpo Honkapolohja & Sarah Chevalier (eds.), *New approaches to English linguistics: Building bridges*, 13–33. Amsterdam: John Benjamins. DOI: 10.1075/slcs.177.
- Hirano, Keiko & David Britain. 2020. Accommodation and social network: Grammatical variation in a community of expatriate English speakers in Japan. In Yoshiyuki Asahi (ed.), *Proceedings of methods XVI: Papers from the sixteenth international conference on methods in dialectology*, 2017, 91–104. Berlin: Peter Lang. DOI: 10.3726/b17102.
- Jankowski, Bridget L. 2016. “*We’ve got our own little ways of doing things here*”: *Cross-variety variation, change and divergence in the English stative possessive*. University of Toronto Linguistics PhD students’ Generals Papers. Toronto. <http://twpl.library.utoronto.ca/index.php/twpl/article/view/19589/19646>.
- 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. DOI: 10.1111/j.1473-4192.1993.tb00042.x.
- Kroch, Anthony S. 1989. Reflexes of grammar in patterns of language change. *Language Variation and Change* 1. 199–244. DOI: 10.1017/S0954394500000168.

- Lorenz, David. 2016. Form does not follow function, but variation does: The origin and early usage of possessive HAVEgot in English. *English Language and Linguistics* 20(3). 487–510. DOI: 10.1017/S1360674316000332.
- National Statistics Center of Japan. 2021. *Statistics for foreign residents: Foreign residents by the administrative divisions of Japan, nationality and area, and qualification (surveyed in December 2019)*. <https://www.e-stat.go.jp/dbview?sid=0003416093%3E>.
- Nycz, Jennifer R. 2011. *Second dialect acquisition: Implications for theories of phonological representation*. New York University. (Doctoral dissertation). <https://api.semanticscholar.org/CorpusID:60375475>.
- Omdal, Helge. 1994. From the valley to the city: Language modification and language attitudes. In Bengt Nordberg (ed.), *The sociolinguistics of urbanization: The case of the Nordic countries*, 116–148. Berlin: de Gruyter. DOI: 10.1515/9783110852622.
- Shockley, Linda. 1984. All in a flap: Long-term accommodation in phonology. *International Journal of the Sociology of Language* 46. 87–95. DOI: 10.1515/ijsl.1984.46.87.
- Stanford, James N. 2008. A sociophonetic analysis of Sui dialect contact. *Language Variation and Change* 20. 409–450. DOI: 10.1017/S0954394508000161.
- Tagliamonte, Sali A. 2003. “Every place has a different toll”: Determinants of grammatical variation in cross-variety perspective. In Günter Rohdenburg & Britta Mondorf (eds.), *Determinants of grammatical variation in English*, 531–554. Berlin: de Gruyter Mouton. DOI: 10.1515/9783110900019.
- Tagliamonte, Sali A. 2013a. *Roots of English: Exploring the history of dialects*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9781139046718.
- Tagliamonte, Sali A. 2013b. The verb phrase in contemporary Canadian English. In Bas Aarts, Joanne Close, Geoffrey Leech & Sean Wallis (eds.), *The verb phrase in English: Investigating recent language change with corpora*, 133–154. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9781139060998.
- Tagliamonte, Sali A. & Alexandra D’Arcy. 2007. The modals of obligation/necessity in Canadian perspective. *English World-Wide* 28(1). 47–87. DOI: 10.1075/eww.28.1.04tag.
- Tagliamonte, Sali A., Alexandra D’Arcy & Bridget Jankowski. 2010. Social work and linguistic systems: Marking possession in Canadian English. *Language Variation and Change* 22. 149–173. DOI: 10.1017/S0954394510000050.
- Trudgill, Peter. 1986. *Dialects in contact*. Oxford: Basil Blackwell.
- Trudgill, Peter. 2004. *New-dialect formation: The inevitability of colonial Englishes*. Oxford: Oxford University Press. <https://www.jstor.org/stable/10.3366/j.ctv2f4vkzd>.



# Chapter 6

## Minimal minimal pairs: Vocalic length in Unterland and Polish Central Yiddish

Chaya R. Nove<sup>a</sup> & Benjamin Sadock<sup>b</sup>

<sup>a</sup>Brown University <sup>b</sup>Independent researcher

Existing descriptions of Central Yiddish, one of three main Eastern Yiddish dialects, are based on Yiddish in the Polish lands where it was spoken. The southern portion of this dialect region, the Transcarpathian area known as the Unterland in folk terminology, has been mostly ignored by Yiddish linguists. This borderland region was settled by Yiddish-speaking Jews relatively late, and was the site of ongoing migration, language and dialect mixing, and geopolitical upheaval in the century prior to World War II. Unterland Yiddish is also the ancestral dialect of the vast majority of contemporary Yiddish speakers, and as a result of the research lacuna, linguists studying contemporary Yiddish have only the literature on Polish Central Yiddish to use as a baseline for tracing innovation and change in present-day spoken Yiddish. This study uses archival data to analyze differences in the acoustic correlates of the length contrast in Central Yiddish peripheral vowels /i/, /u/, and /a/ across genders and regions (Poland vs. Unterland). The results show a diminishing duration distinction between the long and short vowels in two of the three pairs among speakers of the Unterland, with Unterland female speakers exhibiting shorter differences than the males in that region. These results point to a change in progress in the interwar period in the Unterland affecting, and possibly destabilizing, the length contrast. This study provides an important link between the Yiddish of today and that of the prewar era, making it possible to evaluate claims about innovation in the vowel systems of contemporary Yiddish.

### 1 Introduction

In Yiddish dialectology, the primary basis for the classification of major European dialects is their vowel systems, which vary markedly and systematically



across the territory where the language was spoken before World War II. Central Yiddish, which was spoken in a sort of vertical strip from Central Poland down through Galicia and the northeastern sector of the Hungarian portion of the Habsburg Empire, is characterized by, among other features, a length contrast in the peripheral vowels /a/, /i/, and /u/. These length distinctions are also observed in contemporary New York Hasidic Yiddish, which derives from Central Yiddish (Nove 2021). However, when the durations of these vowels are measured in recordings of speakers raised in the Transcarpathian region of interwar Europe known colloquially as the Unterland, the ancestral homeland of most New York Hasidim, a surprising result emerges: while all three vowel pairs exhibit significant differences, the long-short ratios are smaller than those observed in other languages with length-distinguishing vowel systems (Nove 2021: 92). This result suggests that the length feature was in flux and possibly beginning to disappear in Unterland Yiddish but reemerged again as a qualitative (tense-lax) distinction in New York Hasidic Yiddish. Aside from this reinforcement of an historical contrast in a new language environment, which is theoretically interesting but outside the scope of the present study, the observation about the apparent tenuousness of the length contrast in Unterland Yiddish raises questions about the nature of the contrast in other (non-Hungarian) Central Yiddish regions, including: Are the duration differences between the long and short vowels in each pair larger in Polish Central Yiddish than in Unterland Central Yiddish? If yes, what factors influenced the change in the latter? Are there observable changes in vowel quality that correlate with the changes in vowel duration? If not, is a contrast that relies primarily on such small duration differences sustainable?

This study is a first attempt to address some of these questions. Utilizing data from the newly developed *Corpus of Spoken Yiddish in Europe* (Bleaman & Nove in press), we analyze the acoustic correlates (quality and duration) of the vocalic length contrast in the Yiddish peripheral vowels of sixteen speakers, male and female, raised in the historical Unterland region, and eight male speakers from Central Poland. While the main objective is a comparative analysis of Unterland versus Polish Yiddish, it is also the first acoustic analysis of Polish Yiddish vowels. The results, which replicate Nove's (2021) findings, also provide support for the author's hypothesis about a diminishing length contrast in prewar Unterland Yiddish by revealing significantly shorter duration differences in some Unterland Yiddish vowel pairs relative to their Polish Yiddish correlates.

While Yiddish dialects in general, and Central Yiddish in particular, have been relatively well documented, descriptions are based exclusively on Yiddish in the Polish lands (Herzog 1965, Jacobs 1993, 2005); the Yiddish spoken in the Unterland region was essentially ignored by linguists for a variety of reasons including ideological biases (U. Weinreich 1964; see however Sadock & Masor 2018). This neglect has had a larger than anticipated effect on Yiddish scholarship, as the different timelines of Germany's invasion of Poland vs. Hungary during World War II and subsequently higher survival rates for Jews in the latter resulted in Unterland Yiddish becoming the dialect of the vast majority of contemporary speakers. Linguists who study contemporary Yiddish typically resort to descriptions of Central Yiddish as a baseline for comparison, without taking into account the potential systematic differences between Unterland Yiddish and Polish Yiddish that may already have existed in the interwar period. This methodology risks casting all particularities of Hasidic Yiddish that differ from Polish Central Yiddish as innovations, when they may in fact simply be preserved features of Unterland Yiddish. The study reported in this chapter is a first step toward remedying this problem.

The following section of this paper provides some background information about the historical development of Yiddish vowel systems and the ways in which this contrast manifests in Central Yiddish. In Section 3 we briefly lay out the sociohistorical differences between the Jewish communities in Central Poland and the Unterland that help motivate our view of the Yiddish spoken in these regions as distinct subdialects. The data and methods used in this study are described in Section 4, and Section 5 presents the results. A discussion of the findings along with some conclusive remarks is provided in Section 6.

## 2 Yiddish dialects in prewar Europe

Yiddish is divided into Western and Eastern Yiddish (good treatments of Yiddish dialectology can be found in Birnbaum 1923, 1979, Herzog 1965, M. Weinreich 1973, Jacobs 2005, Beider 2015; the discussion that follows draws from these sources). Most of the western varieties, generally spoken coteritorially with German in central Europe, were moribund by the mid-nineteenth century. For this paper, we can disregard those dialects and focus on Eastern Yiddish.

Eastern Yiddish is divided into three main dialects, most commonly referred to by scholars as Northeastern Yiddish (NEY), Southeastern Yiddish (SEY), and Central Yiddish (CY).<sup>1</sup> As noted above, these dialects are distinguished based on their vowel systems. NEY, often called Lithuanian or Litvish Yiddish, was spoken historically in (what is today) Lithuania, Belarus, parts of Latvia, northeastern Poland, and northeastern Ukraine. It is characterized by, among other features, a series of mergers of long and short vowel pairs. CY was spoken in much of the rest of Poland as well as in the region south of the Carpathian Mountains known to Yiddish-speaking residents as the Unterland, which comprises parts of northern Romania, western Ukraine, eastern Slovakia, and northeastern Hungary. This dialect, which is sometimes popularly referred to as Polish Yiddish, Galician Yiddish, or Hungarian Yiddish, depending on the region one is referencing, maintains vocalic length distinctions. SEY, popularly known as Ukrainian Yiddish, was spoken in northern, central, and southern Ukraine as well as Moldova and eastern Romania. Overall, it has much more in common with CY than with NEY, the two together forming a sort of Southern Yiddish in contrast with NEY. SEY preserves some length contrasts but not as many as CY. At this point we will turn to the contrasts in question and examine their status in the three dialects of Eastern Yiddish in more detail.

## 2.1 Length contrast in Yiddish vowels

The consensus among Yiddish linguists is that proto-Yiddish had robust length distinctions in five vowels, reflecting its Middle High German origins (e.g., M. Weinreich 1973). Diphthongization occurred in some cases, preserving phonemic contrast but not as a length distinction. A vowel shift raised /a:/ to /o:/ and then to /u:/ in CY. Modern CY eventually wound up with length contrasts in three vowels: (1) {/a/, /a:/}, the latter reflecting a diachronic process of /a/ monophthongization; (2) {/i/, /i:/} which reflects not just the historical length distinction of /i/ vowel but also that of proto-Yiddish /u/, which in the southern dialects was fronted and later unrounded to merge with /i/; and (3) {/u/, /u:/}, initially a

---

<sup>1</sup>This is the terminology developed by Max Weinreich; Dovid Katz uses two of these terms as well but calls Central Yiddish “Mideastern Yiddish.” In many ways this is a superior name, as it does not suggest a greater affinity between Southeastern and Northeastern Yiddish, reflecting a somewhat idiosyncratic belief of Max Weinreich’s, following Prilutski (1920). It also does not seem to place Central Yiddish as somehow intermediate between Western and Eastern Yiddish, which Weinreich himself believes to be the case. Though Katz’s term may be preferable, it has not been widely adopted by scholars, and so we are using Weinreich’s more conventional terminology here. Note that the tripartite division of Eastern Yiddish and the general boundaries of the three dialects are not controversial.

conditioned contrast that may have been phonologized over time;<sup>2</sup> its phonemic status is unclear and it has no minimal pairs. Additionally, in some CY regions, the diphthong /ou/ was monophthongized to /o:/, contrasting with short /o/.

NEY completely lacks quantitative vocalic length distinctions: long /i/ is merged with short /i/, long /u/ is merged with short /u/, and /o/, which is a reflex of MHG /a:/ and corresponds to CY/SEY /u/, is further merged with the reflex of MHG short /o/. The situation with SEY is somewhat more complex: Some areas of SEY adjacent to CY have the same three length distinctions discussed above; in other areas, /ou/ underwent monophthongization to /u:/. Not coincidentally, those areas do not exhibit the conditioned shortening of /u/ that occurred in CY. Thus, in these areas there is a truly phonemic length distinction for /u/, seen in minimal pairs like /mul/ ‘time’ and /mul/ ‘mouth’. In much of SEY /a/ was monophthongized to short /a/ (unlike CY /a:/). In these regions, the historically short /a/ underwent a partial shift to /o/ so the contrast was maintained. This leaves /i/, which preserves a short-long contrast manifested as a lax-tense distinction (see Glasser 2008, M. Weinreich 1973).<sup>3</sup>

In the following section we provide a brief historical overview of the Jewish communities in Central Poland and the Unterland to underscore the factors (geographical, sociocultural, political) that may have led to dialect divergence.

### 3 Sociohistorical background

#### 3.1 Jewish communities in Central Poland

Jewish migration to Polish lands was initiated around the tenth century, primarily by Jews fleeing persecution in the west, and continued throughout the millennium. These Yiddish-speaking migrants were attracted by laws that afforded Jewish residents comparatively more civil rights and better opportunities than elsewhere in Europe (Weinryb 1973). The first documented evidence of a Jewish presence in the Duchy of Mazovia, where Warsaw is located and where most of the speakers in our sample were raised, is from 1414; however, it is likely that some Jews lived since there since its establishment in the thirteenth century (Polonsky 2010, Weinryb 1973). A series of variably successful expulsions kept the Jewish

<sup>2</sup>This process, sometimes referred to as Birnbaum’s law, shortens /u:/ when it is followed by a labial or velar consonant (Birnbaum 1934, 1979, Jacobs 1990, Katz 1982, M. Weinreich 1973, U. Weinreich 1964).

<sup>3</sup>Note that assertions about the acoustic manifestations of the length contrast (i.e., duration versus quality) are not based on actual analysis of phonological data. As such, the length contrast of /i/ in SEY is an important question for a future study.

population in check during the Middle Ages, but by the eighteenth-century Warsaw's Jewish population had grown dramatically, and with it its importance as a center of Jewish cultural development (Weinryb 1973). The majority remained Yiddish speaking and Orthodox even as the waves of assimilation swept through Western Europe in the nineteenth century. Moreover, Hasidism, a pietistic form of Judaism that began in eighteenth century Medzhybizh (presently in Ukraine) and spread throughout Eastern Europe, had gained a strong foothold in the region. At the eve of World War II, Warsaw Jews numbered approximately 350,000 and constituted about 30 % of the total population – the largest Jewish community in Europe and second largest in the world (Polonsky 2010).

### 3.2 Jewish communities in the Unterland

The western and eastern sectors of the Hungarian portion of the Habsburg Empire followed distinct trajectories of Jewish migration, history, and culture. Accordingly, the area was considered by Jews to comprise of two subregions: The Oyberland (western region), corresponding to most of modern Hungary along with southern Slovakia and the Burgenland region of modern Austria; and the Unterland (eastern region), delineated above (U. Weinreich 1964, Krogh 2012) and shown on a map in Figure 1.<sup>4</sup> The Oyberland was settled by Jews somewhat earlier and was mostly unaffected by the development of Hasidism. In contrast, mass Jewish migration to the Unterland began only in the nineteenth century (Keren-Kratz 2019, Cooper 2019, Jelinek 2007). New settlers came primarily from Galicia, a region that is now in southern Poland and western Ukraine, following the annexation of Galicia by Austria, which transformed this section of the Carpathian Mountains from an international border between Poland and Hungary into an internal border within the Habsburg Empire. Unterlender Jews tended to be Hasidic and thus somewhat more resistant to linguistic assimilation than their Oyberlender neighbors, who had largely abandoned Yiddish for Hungarian and/or German by the twentieth century (Komoróczy 2018). The Yiddish spoken respectively by Oyberlender and Unterlender Jews also reflected their different histories: While Jews in the Unterland, as noted, spoke a variety of Central Yiddish, evidence of their roots in and ties to Galicia, the Yiddish of Oyberlender Jews has traditionally been grouped by scholars with Western Yiddish (U. Weinreich 1964).

While not nearly as linguistically assimilated as the Oyberlender, Unterlender Jews were still somewhat less likely to be Yiddish dominant than their Polish

---

<sup>4</sup>All maps included in this paper were created using the *ggmap* package (Kahle & Wickham 2013) in R software (version 3.5.0, R Core Team 2021).



Figure 1: Present-day map showing the approximate boundaries of the historical Oyberland and Unterland regions based on demarcations by Krogh (2012) and Weinreich (1964).

brethren; and Unterlander women, who typically attended public school as children, were far more prone to linguistic assimilation than their male counterparts, most of whom received a religious education, in Yiddish (see e.g., Rubin 1972: 152, Jelinek 2007: 12).

In summary, if the Jewish communities of Central Poland can be regarded as emblematic of Eastern European Jewish culture by virtue of their longevity, geographical centrality, and size, the Unterland communities should be viewed as peripheral. Established relatively late in European Jewish history, they were somewhat insulated from other Jewish communities to the north and the east by the Carpathian Mountains that encircled them. Although the population grew rapidly in the nineteenth century, the number of Jews, even in the big cities of the region, remained low in comparison to Polish urban centers (Jelinek 2007: 13). Moreover, the Unterland was, in many ways and for many people, a transitional place, a waystation, as Jelinek (2007: 34) describes it:

[...] it appears that during the nineteenth century Subcarpathian Rus' [corresponding to the Zakarpattia Oblast in modern Ukraine] served, on the one hand, as a place of refuge for Jews from neighboring lands and, on the other, as a transit point for those who wanted to leave central Europe altogether.

Furthermore, Jewish communities in the Unterland were embedded within a larger multiethnic society and developed against a backdrop of shifting political borders and conflicting Jewish ideologies (Keren-Kratz 2019, Cooper 2019, Jelinek 2007, Švorc 2020; see also Schäfer 2022 on the influence of political boundaries on Yiddish dialects more generally.). Indeed, Cooper (2019: 200) categorizes Carpathian Ruthenia, which is contained within the Unterland, as a borderland, and refers to it as a metaphorical “catchbasin” containing elements of Jewish ideologies “collected” from the surrounding regions. Linguistically, Unterland communities were in contact with a variety of co-territorial languages, including Hungarian, German, Romanian, Ukrainian, Ruthenian, Czech, and Slovak; and were in close contact with Oyberland communities, whose (Western) Yiddish dialect influenced their own (U. Weinreich 1964). Finally, as noted above, there was a strong tendency among female Unterland speakers toward assimilation to the local majority language, usually Hungarian.

We believe that the circumstances described here make it highly likely that Unterland Central Yiddish diverged from Polish Central Yiddish and developed independently in the century preceding WWII and justify our view of Polish Yiddish as the more stable, conservative dialect and Unterland Yiddish as potentially innovative.

### 3.3 Problems and hypotheses

As noted above, initial research comparing data from interviews with New York-area speakers of Hasidic Yiddish and data from interviews with Holocaust survivors from the Hungarian Unterland revealed a surprisingly short duration distinction on the part of the latter for the vowels /a/ and /i/ – below the 50 milliseconds proposed by Labov & Baranowski (2006) as the threshold for perceiving difference in vowels that differ primarily in length (Nove 2021). Furthermore, the duration differences between the long and short vowels /i/ and /a/ of the female speakers in that study were significantly smaller than those of the males. Given the recurring trend in sociolinguistics for females to be at the forefront of change, and especially in light of the female tendency for linguistic assimilation in the Unterland, this result suggests that the shrinking differences reflect a change in progress.

The aim of this study is thus to further explore vowel length contrasts in Central Yiddish. The first goal was to replicate the findings by Nove (2021) using a slightly expanded dataset of fifteen (versus twelve) speakers, eight of them female. The second and more important objective was to compare the duration differences in the Yiddish spoken in the Central Poland and Unterland regions in

order to get a better idea of the nature of the length contrast in the vowel systems of these subdialects; and to discover whether there is evidence of divergence. Our prediction was that the duration differences in the vowels of Unterland speakers would be significantly smaller than those of the Central Poland speakers, with an effect of gender in the former group similar to the one found in the previous study.

## 4 Data and methods

The data for this project come from the newly developed *Corpus of Spoken Yiddish in Europe* (CSYE), which is funded by the U.S. National Science Foundation under the supervision of Isaac Bleaman of the University of California, Berkeley (Bleaman & Nove in press). The CSYE is an open-access digital language archive based on approximately two hundred interviews with Holocaust survivors, each of which averages about two hours in length, video-recorded by the *USC Shoah Visual History Foundation* during the 1990s. The CSYE team is currently in the process of transcribing these interviews and reviewing the transcriptions, and testimonies are being published on a dedicated website ([yiddishcorpus.org](http://yiddishcorpus.org)) in audio and text formats when their transcripts are completed. The CSYE enabled us to create a corpus of the first hour of testimonies by fifteen speakers from the Unterland region (including nine that were part of Nove (2021), and eight speakers from Central Poland for analysis (USC Shoah Foundation Visual History Archive 2022a,b).<sup>5</sup> An additional benefit of this data source is the possibility of expanding the sample size as more testimonies are processed (Nove 2023, Nove & Sadock submitted). The present study is the first utilization of the data emerging from the project.

### 4.1 Data processing

Recordings in our corpus were transcribed in ELAN according to the protocols established by the CSYE by a team that includes both authors, and the transcriptions were reviewed. A pronunciation dictionary was created based on unique words in the corpus and was used to align the audio and text into word and sound segments using the Montreal Forced Aligner *train and align* function (McAuliffe et al. 2017). Formant frequencies and duration measures were extracted from the sound files and TextGrids using a Praat plug-in called Fast Track (Barreda 2021).

---

<sup>5</sup>As of this writing, most of the testimonies analyzed in this study are pending final review prior to publication on the web site.

The rest of the data processing, analysis, and visualization was conducted using R software (R Core Team 2021). The data were filtered to remove the most reduced function words and other reduced words. Mahalanobis distance was calculated by speaker for each vowel using the `tidy_mahalanobis()` function in the `joeyr` package (an implementation of `mahalanobis()`) (Stanley 2020), and the highest 5% were discarded. The vowel midpoints (median taken from the 40–60% vowel duration segment) were then normalized using the modified Nearey method utilized by Labov and colleagues in *The Atlas of North American English* (Labov et al. 2006) via the `norm_anae()` function in the `joeyr` package (Stanley 2020), and the data were filtered to include only the target vowels.<sup>6</sup>

Two analyses were conducted using this data output. The first, designed to replicate the results of Nove (2021), calculated the duration differences in the vowels of fifteen Unterland speakers, eight of them female, and checked for a gender effect. Table 1 shows a list of the fifteen Unterland speakers by last name, birth year, gender, and the locations (city and country) in which they were raised; these locations are shown on the map in Figure 2.<sup>7</sup>

The objective for the second analysis was to test the hypothesis in Nove (2021) about subdialectal divergence within CY across these two regions. To this end, we compared male speakers raised in interwar Central Poland (Warsaw and nearby areas) and the Unterland. In this comparative analysis, we only included the male speakers from the Unterland for consistency (since we did not yet have any data from Polish females to analyze). Table 2 shows a list of the six Polish speakers and Figure 3 shows their respective places of origin.

Using the methodology described here, we extracted a total of 27,163 vowels, 19,864 from the Unterland and 7,299 from Central Poland, for analysis. Table 3 and Table 4 show the breakdown of vowels analyzed in each class for each sub-region.

---

<sup>6</sup>One reviewer expressed a concern that survivors might have modified their speech in a way that would have impacted their vowels, at least in the very beginning of the interview, due to the perceived formality of the encounter. While some sociolinguists choose to discard the first five minutes of an interview for this very reason, we were loath to do so given that only the first hour of these interviews had been transcribed at the time and we did not want to lose any of the valuable data. Instead, we listened to the testimonies carefully and used our judgments about whether the speakers' dialects were typical for the location. While doing so, we identified one speaker (Robak) who used the more standard pronunciation [aɪ] in place of the expected CY form [a:]. For this reason, all tokens of the relevant long-short vowel pair by this speaker were excluded from our analysis. Other than the aforementioned issue with one particular speaker and a number of isolated instances of accommodation or interference of prestige forms, we did not encounter any systematic deviances that would skew the analyses.

<sup>7</sup>Note: Name label positions on the map are shifted slightly to avoid overlap.

Table 1: Fifteen Unterland speakers by last name, birth year, gender, location in which they were raised, and country in which they were interviewed.

Speaker	Birth Year	Gender	City Raised (Yiddish)	Presently	Interview Country
Berger	1925	F	Beregsaz	Berehovo, UA	Israel
Bitterman	1923	F	Svalyeve	Svaliava, UA	U.S.
Burekhovich	1920	F	Boronyave	Boronyava, UA	U.S.
Erps	1922	M	Kretshenif	Crăciunești, RO	U.S.
Fishman	1919	M	Ungvar	Užhorod, UA	Israel
A. Fried	1917	M	Satmar	Satu Mare, RO	Israel
D. Fried	1913	F	Rezavlye	Rozavlea, RO	U.S.
Gancz	1924	M	Siget	Sighetu Marmației, RO	Sweden
Heimlich	1925	F	Mishkolts	Miskolc, HU	Australia
Herskowitz	1922	M	Siget		U.S.
Katz	1924	M	Satmar		U.S.
Polak	1922	F	Tetsh	Tyachiv, Ukraine	Australia
Preizler	1914	F	Siget		Israel
Rosenfeld	1910	M	Satmar		Canada
Taub	1911	F	Satmar		Israel

Table 2: Six Polish speakers by last name, birth year, gender, location in which they were raised, and country in which they were interviewed.

Speaker	Birth Year	Gender	City Raised	Presently	Interview Country
Popowski	1918	M	Varshe	Warsaw, PL	Argentina
Robak	1922	M	Varshe		Poland
Sherman	1925	M	Apt	Opatów, PL	Israel
Silver	1911	M	Varshe		Australia
Scheinberg	1912	M	Varshe		U.S.
Zylberberg	1922	M	Varshe		Argentina

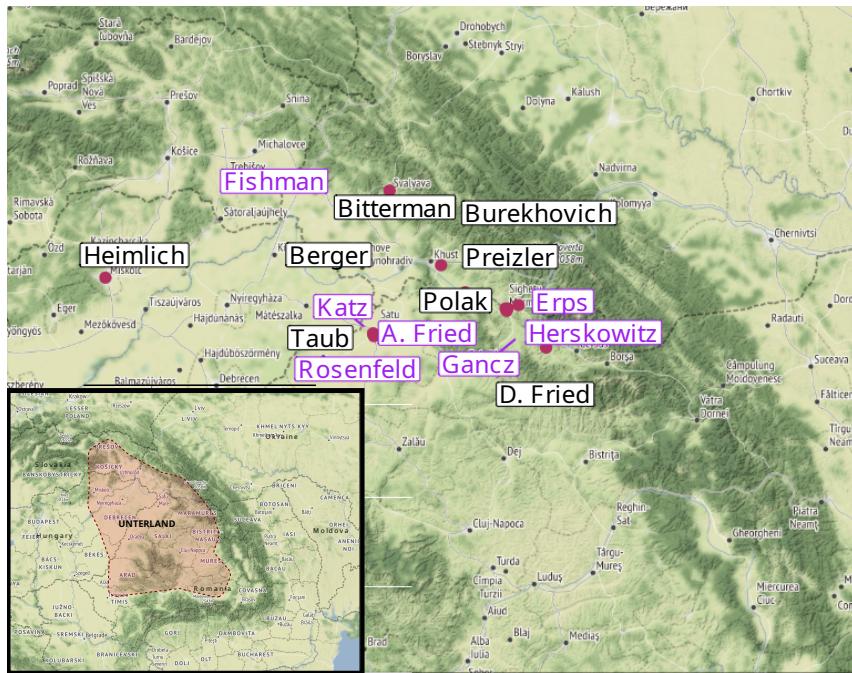


Figure 2: Map with points showing cities where fifteen Unterland speakers were raised and labels displaying the last names of female speakers in black text and male speakers in purple. Insert map delineates the approximate boundaries of the Unterland within the larger region.

Table 3: Breakdown of Unterland vowels analyzed by vowel category.

vowel	count	% of sub corpus
i:	2552	0.13
i	4806	0.24
u:	2316	0.12
u	2118	0.11
a:	1686	0.08
a	6386	0.32

Table 4: Breakdown of Central Poland vowels analyzed by vowel category.

vowel	count	% of sub corpus
i:	1003	0.14
i	1744	0.24
u:	775	0.11
u	667	0.09
a:	766	0.10
a	2344	0.32

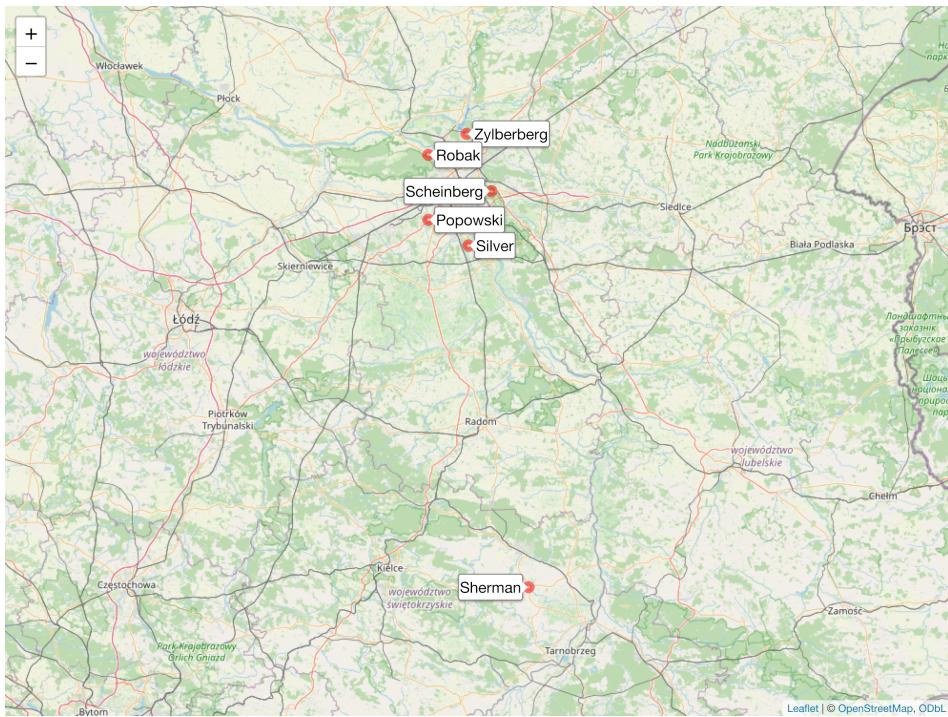


Figure 3: Map with points showing cities where the six Polish speakers were raised and labels displaying the speakers' last names.

## 4.2 Statistical analysis

To test for differences across genders and regional groups, linear mixed-effects regression models were fit for each pair of vowels using the `lmer()` function in the `lme4` package (Bates et al. 2013). The Satterthwaite approximation in the `lmerTest` package (Kuznetsova et al. 2017) was used to calculate all  $p$ -values. Each model had decadic log of vowel duration as the dependent variable and included the following variables and interactions as fixed effects:

1. Interaction: vowel  $\times$  gender (Unterland only)
2. Interaction: vowel  $\times$  corpus
3. Number of segments in the word
4. Preceding segment (silence, vowel, or consonant coded for voice, manner, and place of articulation)

5. Following segment (silence, vowel, or consonant coded for voice, manner, and place of articulation)
6. Interview country (to infer influence of postwar contact language)

Random intercepts:

7. Speaker
8. Word

To quantify the relative overlap between vowel pairs on the quality dimension, we grouped the data by gender and region and performed a MANOVA, with F1 and F2 as the dependent variables and vowel, phonological context, and duration as independent variables, generating a Pillai score for each pair (Hay et al. 2006, Nycz & Hall-Lew 2013). The MANOVA output also includes a *p*-value for the Pillai score that indicates whether the difference between the two vowel clusters is significant.

## 5 Results

### 5.1 Unterland

#### 5.1.1 Vowel duration

The results of the Unterland analysis replicate the findings of Nove (2021) in that the duration differences of all vowel pairs are remarkably small. Figure 4 shows mean vowel duration (in milliseconds) for all Unterland speakers, calculated separately by gender group and labeled with the mean difference for each pair. Female speakers also exhibit smaller differences for /i/ and /a/ than their male counterparts, and their differences fall below 50 milliseconds for all vowel pairs.

Regression analyses (LMM) indicate that these differences are statistically significant: the models for duration show a significant effect of gender in /i/ and /a/. In Figure 5 the gender differences can be seen in the different slopes of the lines on plots of estimated means extracted from the regression models, with vowel on the x-axis and duration on the y-axis.<sup>8</sup> The results of the LMMs are shown in the appendix (Section 7).

---

<sup>8</sup>The reason why the female speakers in this study have longer vowel durations overall is probably because several of the female speakers have particularly slow speech rates.

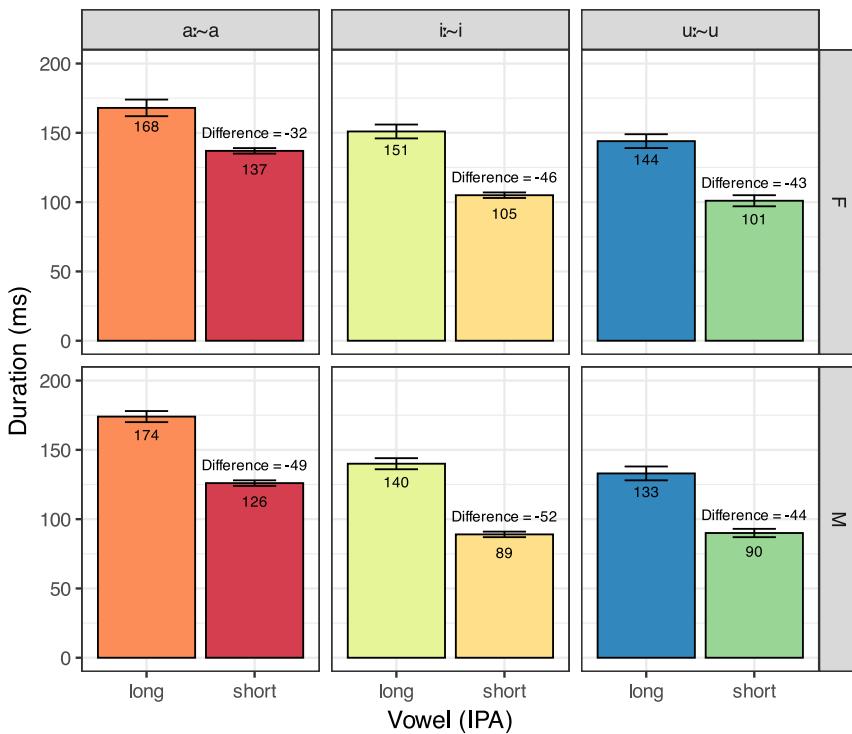


Figure 4: Unterland mean duration faceted by vowel pair (columns) and by gender (rows), with 95% confidence interval standard error bars. Annotations on the bars indicate mean duration for each vowel, and the durational differences between the vowels in each pair are shown above the short vowel bar.

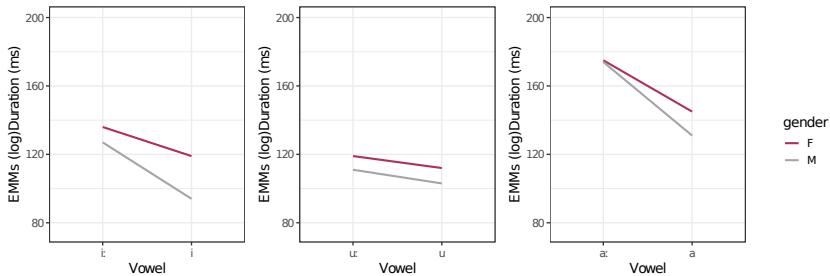


Figure 5: Unterland estimated marginal means of duration LMM for all vowel pairs, with vowel on the x-axis and duration (back-transformed from decadic log) on the y-axis, faceted by gender (rows), with lines connecting the vowels, colored by gender group.

### 5.1.2 Vowel quality

Our analysis of vowel quality among Unterland speakers did not yield any informative patterns thus far. Figure 6 is a contour plot of F1 and F2 values, faceted by gender, in which the lines represent density of the data. Among the female speakers there appears to be more lowering of [i] relative to [i:], but the short vowels of both pairs are more centralized among the male speakers. (Note that the bimodal distribution in the male /a/ is caused by one outlier speaker whose vowel space is particularly small, hence the higher /a/s.)

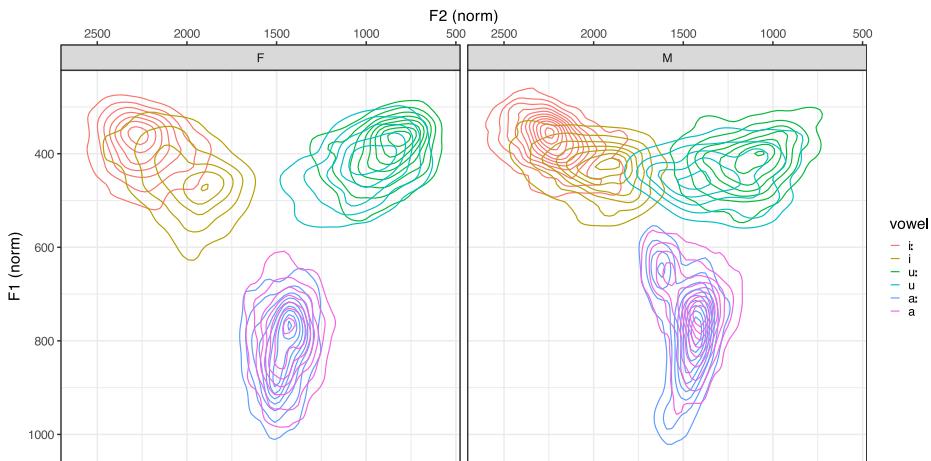


Figure 6: Contour plots of all vowel tokens ( $N = 19,864$ ) by Unterland speakers by F1 and F2 and density, faceted by gender group.

The results of the MANOVAs are in Table 5. Pillai scores range from 0 to 1, with 1 indicating no similarity between the two clusters and 0 indicating no difference. Based on these scores, the quality of the long and short vowels in the high vowel pairs are similar, and they are nearly identical in /a/. Female speakers appear to be more merged in all vowel pairs than males; however, the differences between u:~u and a:~a are minimal.

In examining the data, we noted a great deal of interspeaker variability in the spectral overlap of the long-short high vowels in both gender groups. This is reflected in by-speaker Pillai scores (not shown here), which show a wide range within each group. However, we did not find a correlation between by-speaker Pillai scores and duration difference in these vowel pairs.

To illustrate the kind of interspeaker variability we observed, we present Figure 7, which plots the peripheral vowels of the most and the least merged speakers in our dataset by F1 and F2. The most merged speaker (Taub, on the right)

Table 5: Pillai scores derived from a MANOVA measuring spectral overlap for all vowel pairs by gender group.

vowel	gender	Pillai	<i>p</i> -value
i:~i	F	0.203	0.000
	M	0.313	0.000
u:~u	F	0.096	0.000
	M	0.169	0.000
a:~a	F	0.012	0.000
	M	0.049	0.000

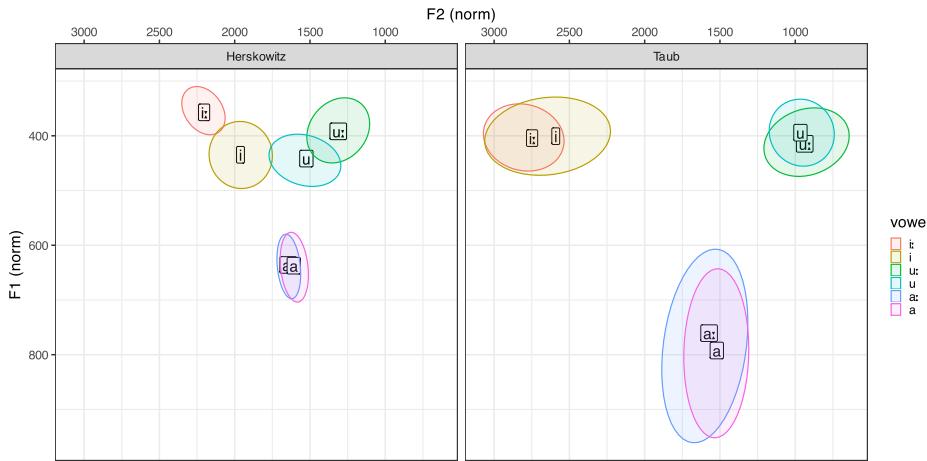


Figure 7: Vowel tokens ( $N = 2,472$ ) of two speakers (Herskowitz, M and Taub, F) plotted by F2 on the x-axis and F1 on the y-axis. Square labels containing IPA symbols represent the location of the vowel means, and ellipses show 68% confidence in the mean.

happens to be female, and the one with the most separation (Herskowitz, on the left) is male. Their Pillai scores are shown in Table 6.

It also happens that Mrs. Taub immigrated to Israel after the war and Mr. Herskowitz resettled in the United States. Recall that these interviews were conducted about fifty years after the war, and that these two speakers are now (presumably) also speakers of Israeli Hebrew and American English, respectively. Note also the similarity between Herskowitz's vowel system and the English tense-lax system in the high vowels. Modern Hebrew, on the other hand, has

Table 6: Pillai scores derived from a MANOVA measuring spectral overlap for all vowel pairs by two speakers for all vowel pairs.

speaker	vowel	Pillai	<i>p</i> -value
Herskowitz	i:~i	0.584	0.000
	u:~u	0.498	0.000
	a:~a	0.065	0.000
Taub	i:~i	0.059	0.001
	u:~u	0.064	0.012
	a:~a	0.028	0.013

no contrast in these vowels. There is ample evidence in the literature on second language acquisition that even minimal experience with an L2 can influence a speaker's L1 (e.g., Chang 2011). While this analysis has revealed nothing definitive regarding vowel quality thus far, an expanded dataset and more detailed analyses, including an analysis of vowel trajectories rather than just the mid-point measures, might reveal additional patterns of variation.

## 5.2 Central Poland

### 5.2.1 Vowel duration

We turn now to the second aim of our study, a comparison of the durational distinctions in Unterland and Polish CY vowel pairs. Looking at the mean duration with long-short differences by corpus (Figure 8), we can immediately see larger differences in the high vowels of the Polish speakers, but not in /a/.

Regression modeling indicates that these variances are statistically significant. That is, the models for duration show a significant effect of region in the /i/ and /u/ pair. These differences can be seen by examining the slopes on the plots below (Figure 9), which show estimated means of duration by vowel for each pair, with lines connecting the long-short vowels colored by regional group. The results of the LMMs are shown in the appendix.

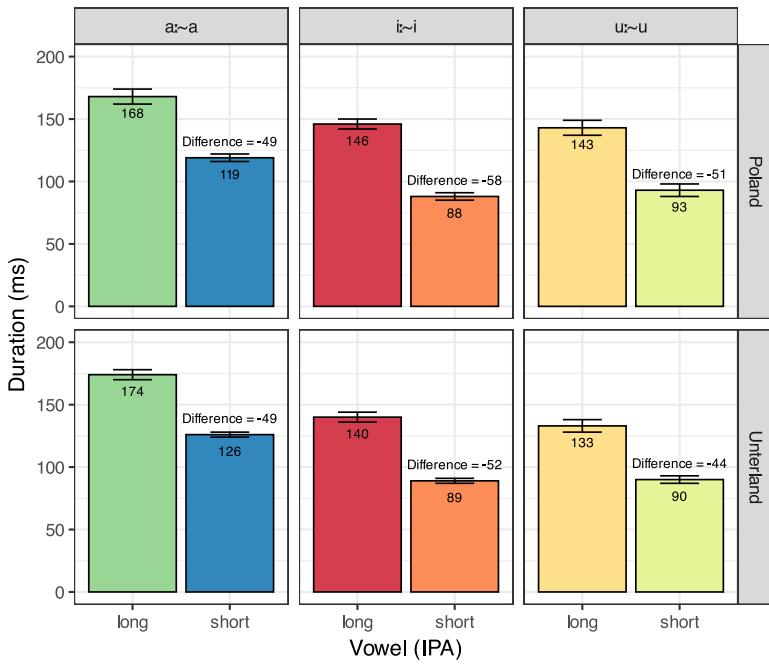


Figure 8: Mean duration for male speakers faceted by vowel pair (columns) and by sub corpus (rows), with 95% confidence interval standard error bars. Annotations on the bars indicate mean duration for each vowel and the durational differences between the vowels in each pair are shown above the short vowel bar.

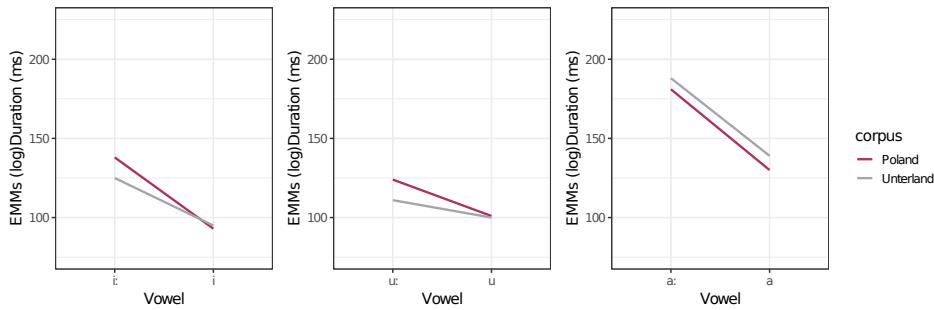


Figure 9: Estimated marginal means of duration LMM, with vowel (long vs. short) on the x-axis and duration (back-transformed from decadic log) on the y-axis, faceted by vowel pair, with lines connecting the vowels colored by corpus (Poland vs. Unterland).

### 5.2.2 Vowel quality

Comparing the two regions on the quality dimension, we once again find our results do not account for the results obtained for vowel duration. A vowel plot of all male speakers, faceted by regional group, is shown in Figure 10. Here too there appears to be a substantial amount of interspeaker variability in the amount of spectral overlap of the long-short vowels in the high vowel pairs within each group. While the high front vowels look a bit more separated in the Polish corpus, the high back vowels appear more separated in the Unterland corpus. Pillai scores calculated for each region, shown in the Table 7, suggest that these differences are indeed very small.

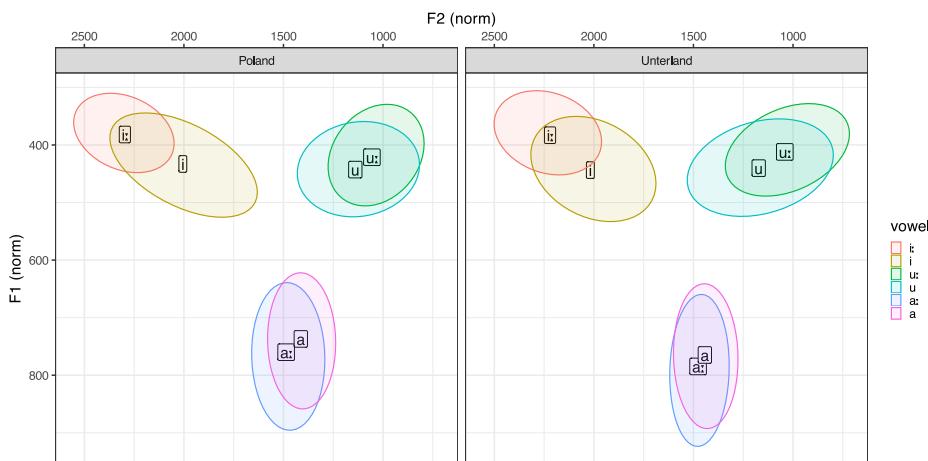


Figure 10: Vowel tokens ( $N = 17,405$ ) of all male speakers plotted by F2 on the x-axis and F1 on the y-axis, faceted by regional group (Poland vs. Unterland). Square labels containing IPA symbols represent the location of the vowel means, and ellipses show 98% confidence in the mean.

It is important to note that in this study we did not analyze a phonological process that reportedly affects the quality of a subset of the target vowels: the insertion of epenthetic schwa between certain long vowels and particular coda consonants.<sup>9</sup> Previous analyses by Nove (2021), and our own impressions from

<sup>9</sup>Jacobs (1993) proposes that such vowel diphthongization is motivated by a phonological avoidance of syllable overlength and formulates two separate but related phonological rules, which he calls *breaking* and *drawl*, to describe the process. Garellek (2020), however, treats it as a single phenomenon and offers a phonetic explanation for the occurrence or nonoccurrence of schwa insertion in such contexts.

Table 7: Pillai scores derived from a MANOVA measuring spectral overlap of male speaker for all vowel pairs by regional group (Poland vs. Unterland).

corpus	vowel	Pillai	<i>p</i> -value
Poland	i:~i	0.325	0.000
	u:~u	0.073	0.000
	a:~a	0.073	0.000
Unterland	i:~i	0.241	0.000
	u:~u	0.101	0.000
	a:~a	0.021	0.000

the data, suggest that the absence or presence of this conditioned diphthongization varies greatly across the CY dialect region. Thus, a careful comparison of vowel quality in these two regions may require that the entire trajectory of the vowels, not merely their midpoints, be analyzed. Plans are underway to conduct such a study on an expanded dataset. We thus refrain from positing anything definitive about vowel quality pending the completion of that analysis.

## 6 Discussion

In this study we analyzed the acoustic correlates of an historical length contrast in CY across two dialect regions: Central Poland and the Unterland. We discovered significant differences in duration distinction across gender groups within the Unterland (in /i/ and /a/), and among male speakers across regional groups (in /i/ and /u/). The results provide some tentative answers to the research questions that motivated this study and give rise to several others.

The small durational differences found in the long-short vowel pairs of speakers in the Unterland region, with females exhibiting significantly shorter differences than males, suggest a change in progress with females in the lead, consistent with the oft-observed trend for female speakers to lead in linguistic innovation. That the Polish (male) speakers have significantly larger durational differences than the Unterlender (male) speakers reinforces this interpretation, especially when considering the relative recency of the establishment of Unterland Jewish communities versus the long and continuous presence of Jews in Central Poland. Taking into consideration other circumstantial factors, including the mountainous barrier between Poland and the Unterland and the degree

of multilingualism, linguistic assimilation, and geopolitical turmoil in the latter region, we seem to have the optimal conditions for linguistic innovation and dialect divergence.

Thus far we do not have an explanation for why this feature was affected and why in this particular direction. Another puzzle is the fact that while the gender effect shows up in the /i/ and /a/ pairs in the Unterland, the regional differences are in the /i/ and /u/ pairs. There is a substantial amount of interspeaker variation in the duration of /u/ among Unterland speakers, which may account for the lack of gender effect there. On the other hand, as noted above, /u/ differs from the other two vowels in that the contrast is not an inherited feature of proto-Yiddish but came about through a more recent split conditioned by phonological environment. Indeed, although it is treated as a length distinction in most of the literature, the phonemic status of short /u/ is disputed by some (e.g., Beider 2015). We believe that the different behavior of /u/ compared to the other two vowel pairs is related to its non-phonemic status, and we are currently experimenting with other methods to further explore these differences. As for the absence of a regional effect in /a/, it is possible that Unterland males have simply not yet reached the threshold of significance. Additionally, /a/ is different from /i/ and /u/ in one very important respect: the contrast is represented orthographically. There is also a strong mental awareness of the diphthongal origin of /a:/, and it is not uncommon for speakers to pronounce it as a diphthong in read or formal speech. Moreover, the dialects that have /ai/ for this vowel (primarily NEY) are considered more prestigious. In fact, one of the speakers in our Polish corpus had to be excluded from the analysis of /a/ because he often pronounces long /a/ as [ai]. So it is likely that the higher rate of literacy among Yiddish-speaking males in that era had the effect of preserving the contrast in /a/ to some extent, accounting for this effect.

## 7 Conclusions

A desideratum of Sadock & Masor (2018) was a better understanding of intra-dialectal differences within the CY dialect region. Our finding of differences in durational distinctions between the Central Poland and Unterland regions is a definitive step in that direction. We are currently working on an expanded version of this study, which includes female speakers from Central Poland, to shed more light on the gender patterns within and across both regional groups. We are also hoping to eventually broaden the geographic reach of this study farther

into Poland, and especially into Galicia,<sup>10</sup> the region immediately to the north of the Unterland, across the Carpathians. In addition to vowel length, we have plans to analyze other phonetic features, such as presence or absence of vowel breaking, differences in /l/ and /r/ quality, and postvocalic /r/ deletion.

It must be noted that the data from the CSYE, while extremely rich in a variety of ways, also inherently contain some limitations for linguistic analysis. For one, at the time of the interviews, nearly all the speakers had spent decades living far away from the cities and towns in which they grew up and had presumably acquired the language of their host country to some extent. It is well known that language can change across a person's lifespan and that exposure to new languages can influence a speaker's first language. Additionally, the interviewers themselves spoke a variety of dialects, often markedly non-natively, and it is impossible to assess what effect, if any, this had on the survivors' speech. Nevertheless, given that the Holocaust effectively destroyed the Yiddish-speaking communities of Eastern Europe, these interviews constitute the best available corpus of high-quality recordings of people who grew up in prewar European Yiddish-speaking communities.

Yiddish is one of very few Germanic dialects that have not been subjected to systematic acoustic analyses, which is unfortunate, given that careful classification of its dialects depends on knowledge of their phonological features. While the destruction of the European Yiddish-speaking homeland certainly complicates such endeavors, this study illustrates how archival recordings can be used to evaluate previous claims about Yiddish dialects and discover additional nuance within and among them. The CSYE will be, upon completion, a source not merely of a large amount of high-quality data suitable for acoustic analysis but also of data from speakers in areas that have previously not been studied, especially marginal and transitional dialect regions.

Finally, but perhaps most importantly, this comparative study functions as an important link between the Yiddish of today and that of the prewar era, making it possible to evaluate claims about linguistic innovation in contemporary Yiddish vowels. As such, it fills an important gap in Yiddish linguistics. Indeed, in a study comparing the peripheral vowels of three generations of contemporary Hasidic Yiddish speakers in New York, Nove (2021) uses data from this archive as a baseline and finds that among the first New York-born generation, the durational differences actually increased in the high vowels relative to the immigrant generation – an apparent reinforcement of a tenuous contrast. Among subsequent

<sup>10</sup>Most of Galicia is within CY territory, although the easternmost region of it, including Kilemey and Ternopol (Kolomyia and Ternopil' in Ukraine) is usually grouped with SEY because of the realization of the reflex of MHG *ei* as /ej/, not /aj/ there.

generations, however, Nove observes a gradual shift in the quality of the short high vowels, with the youngest generation exhibiting a clear qualitative (tense-lax) contrast in these vowel pairs similar to the English (u~v, i~ɪ). This study illustrates how careful studies of prewar Yiddish dialects can shed light on the interplay between the internal and external forces that drive language change. The observed changes in the Hasidic Yiddish vowel system reflect both the persistence of inherited linguistic features and the way that the language is adapting to a changing sociolinguistic context. Such insights are valuable not only for Yiddish linguistics but for a broader understanding of how minority diaspora languages develop in new language contact environments.

## Appendix

Table 8: Results of linear mixed model assessing durational distinction for the high vowels [i:] and [i] in the Unterland

Predictors	Duration (ms)			
	Est.	SE	t	p
(Intercept)	2.20	0.05	45.95	<b>&lt;0.001</b>
vowel IPA [i]	-0.06	0.01	-5.11	<b>&lt;0.001</b>
gender [M]	-0.03	0.04	-0.75	0.455
num seg	-0.01	0.00	-5.69	<b>&lt;0.001</b>
pre context [corNAS]	-0.04	0.02	-2.17	<b>0.030</b>
pre context [dorGLI]	-0.15	0.03	-5.43	<b>&lt;0.001</b>
pre context [dorLIQ]	0.01	0.14	0.06	0.950
pre context [labNAS]	-0.06	0.02	-2.94	<b>0.003</b>
pre context [UNK]	0.01	0.02	0.66	0.511
pre context [V]	-0.09	0.03	-3.54	<b>&lt;0.001</b>
pre context [VcorOBS]	-0.01	0.02	-0.49	0.621
pre context [VdorOBS]	-0.11	0.03	-3.29	<b>0.001</b>
pre context [VlabOBS]	-0.06	0.02	-3.54	<b>&lt;0.001</b>
pre context [XVcorOBS]	-0.10	0.01	-7.78	<b>&lt;0.001</b>
pre context [XVdorOBS]	-0.12	0.02	-5.93	<b>&lt;0.001</b>
pre context [XVlabOBS]	-0.12	0.02	-6.44	<b>&lt;0.001</b>
pre context [XVlarOBS]	0.07	0.02	2.88	<b>0.004</b>

Continued on next page

Predictors	Duration (ms)			
	Est.	SE	t	p
post context [corLIQ]	0.12	0.02	7.82	<0.001
post context [corNAS]	-0.09	0.02	-5.43	<0.001
post context [dorGLI]	0.06	0.12	0.55	0.583
post context [dorNAS]	-0.08	0.02	-3.57	<0.001
post context [labNAS]	-0.10	0.02	-4.73	<0.001
post context [UNK]	0.26	0.03	7.43	<0.001
post context [V]	0.06	0.03	1.82	0.068
post context [VcorOBS]	-0.02	0.02	-1.24	0.214
post context [VdorOBS]	-0.04	0.02	-1.79	0.074
post context [VlabOBS]	0.10	0.02	5.41	<0.001
post context [XVdorOBS]	0.02	0.02	1.26	0.209
post context [XVlabOBS]	-0.04	0.03	-1.46	0.145
post context [XVlarOBS]	0.08	0.05	1.58	0.115
country interview [Canada]	0.02	0.08	0.19	0.853
country interview [Israel]	0.07	0.05	1.32	0.186
country interview [Sweden]	0.03	0.08	0.37	0.710
country interview [U.S.A.]	0.01	0.05	0.14	0.886
vowel IPA [i] × gender [M]	-0.08	0.01	-8.02	<0.001

Random Effects	
$\sigma^2$	0.03
$\tau_{00}$ word	0.01
$\tau_{00}$ speaker	0.00
ICC	0.25
N speaker	15
N word	1054

Observations	7358
Marginal R <sup>2</sup>	0.283
Conditional R <sup>2</sup>	0.462

Table 9: Results of linear mixed model assessing durational distinction for the high vowels [u:] and [u] in the Unterland

Predictors	Duration (ms)			
	Est.	SE	t	p
(Intercept)	2.30	0.06	37.09	<0.001
vowel IPA [u]	-0.03	0.02	-1.19	0.236
gender [M]	-0.03	0.05	-0.71	0.476
num seg	-0.02	0.00	-4.48	<0.001
pre context [corNAS]	-0.05	0.02	-2.21	0.027
pre context [dorGLI]	-0.18	0.05	-3.40	0.001
pre context [dorLIQ]	0.02	0.11	0.21	0.835
pre context [labNAS]	-0.15	0.04	-3.64	<0.001
pre context [OTHER]	-0.28	0.07	-3.81	<0.001
pre context [UNK]	0.03	0.03	1.20	0.230
pre context [V]	-0.01	0.03	-0.33	0.743
pre context [VcorOBS]	-0.08	0.03	-2.85	0.004
pre context [VdorOBS]	-0.10	0.06	-1.77	0.078
pre context [VlabOBS]	-0.01	0.03	-0.25	0.799
pre context [XVcorOBS]	-0.08	0.02	-3.68	<0.001
pre context [XVdorOBS]	-0.09	0.03	-3.43	0.001
pre context [XVlabOBS]	-0.10	0.03	-3.55	<0.001
pre context [XVlarOBS]	-0.06	0.08	-0.71	0.475
post context [corNAS]	-0.03	0.03	-1.12	0.261
post context [dorGLI]	-0.14	0.12	-1.11	0.266
post context [dorNAS]	-0.20	0.03	-6.03	<0.001
post context [labNAS]	-0.11	0.04	-2.72	0.007
post context [UNK]	0.19	0.03	5.96	<0.001
post context [V]	-0.03	0.03	-0.77	0.439
post context [VcorOBS]	0.02	0.03	0.45	0.656
post context [VdorOBS]	-0.16	0.03	-4.83	<0.001
post context [VlabOBS]	0.01	0.04	0.31	0.757
post context [XVcorOBS]	-0.05	0.03	-1.97	0.049
post context [XVdorOBS]	-0.08	0.03	-2.31	0.021
post context [XVlabOBS]	-0.14	0.03	-4.63	<0.001
post context [XVlarOBS]	-0.01	0.07	-0.11	0.916

Continued on next page

Predictors	Duration (ms)			
	Est.	SE	t	p
country interview [Canada]	-0.06	0.10	-0.55	0.581
country interview [Israel]	0.01	0.06	0.15	0.883
country interview [Sweden]	-0.05	0.10	-0.51	0.608
country interview [U.S.A.]	-0.05	0.06	-0.76	0.448
vowel IPA [u] × gender [M]	-0.01	0.01	-0.43	0.668

Random Effects	
$\sigma^2$	0.04
$\tau_{00}$ word	0.01
$\tau_{00}$ speaker	0.01
ICC	0.25
N speaker	15
N word	453
Observations	4434
Marginal R <sup>2</sup>	0.214
Conditional R <sup>2</sup>	0.410

Table 10: Results of linear mixed model assessing durational distinction for the low vowels [a:] and [a] in the Unterland

Predictors	Duration (ms)			
	Est.	SE	t	p
(Intercept)	2.25	0.05	49.43	<b>&lt;0.001</b>
vowel IPA [a]	-0.08	0.01	-6.06	<b>&lt;0.001</b>
gender [M]	0.00	0.04	-0.12	0.904
num seg	0.00	0.00	0.48	0.628
pre context [corNAS]	-0.05	0.01	-3.39	<b>0.001</b>
pre context [dorGLI]	0.04	0.06	0.76	0.449
pre context [dorLIQ]	0.03	0.11	0.27	0.786
pre context [labNAS]	-0.05	0.02	-3.09	<b>0.002</b>
pre context [OTHER]	-0.02	0.07	-0.30	0.761
pre context [UNK]	-0.01	0.01	-1.05	0.296
pre context [V]	-0.02	0.02	-1.36	0.174

Continued on next page

Predictors	Duration (ms)			
	Est.	SE	t	p
pre context [VcorOBS]	-0.06	0.02	-3.84	<b>&lt;0.001</b>
pre context [VdorOBS]	-0.12	0.02	-5.06	<b>&lt;0.001</b>
pre context [VlabOBS]	-0.01	0.02	-0.75	0.455
pre context [XVcorOBS]	-0.06	0.01	-4.55	<b>&lt;0.001</b>
pre context [XVdorOBS]	-0.11	0.02	-6.72	<b>&lt;0.001</b>
pre context [XVlabOBS]	-0.09	0.02	-4.24	<b>&lt;0.001</b>
pre context [XVlarOBS]	-0.16	0.02	-6.58	<b>&lt;0.001</b>
post context [corNAS]	-0.09	0.02	-5.99	<b>&lt;0.001</b>
post context [dorGLI]	0.26	0.14	1.80	0.072
post context [dorNAS]	-0.17	0.02	-8.35	<b>&lt;0.001</b>
post context [labNAS]	-0.13	0.02	-6.40	<b>&lt;0.001</b>
post context [V]	-0.03	0.08	-0.36	0.722
post context [VcorOBS]	0.01	0.02	0.50	0.616
post context [VdorOBS]	0.07	0.04	1.65	0.098
post context [VlabOBS]	0.09	0.02	3.78	<b>&lt;0.001</b>
post context [XVcorOBS]	0.00	0.01	0.26	0.798
post context [XVdorOBS]	-0.04	0.02	-2.25	<b>0.025</b>
post context [XVlabOBS]	0.01	0.03	0.40	0.686
country interview [Canada]	0.05	0.08	0.59	0.557
country interview [Israel]	0.06	0.05	1.26	0.207
country interview [Sweden]	0.04	0.08	0.51	0.608
country interview [U.S.A.]	0.03	0.05	0.60	0.550
vowel IPA [a] × gender [M]	-0.04	0.01	-4.43	<b>&lt;0.001</b>
Random Effects				
$\sigma^2$	0.02			
$\tau_{00}$ word	0.01			
$\tau_{00}$ speaker	0.00			
ICC	0.31			
N speaker	15			
N word	911			
Observations	8072			
Marginal R <sup>2</sup>	0.198			
Conditional R <sup>2</sup>	0.448			

## References

- Barreda, Santiago. 2021. Fast track: Fast nearly automatic formant-tracking using Praat. *Linguistics Vanguard* 7(1). 20200051. DOI: 10.1515/lingvan-2020-0051.
- Bates, Douglas, Martin Mächler, Benjamin M. Bolker & Steven C. Walker. 2013. lme4: Linear mixed-effects models using eigen and S4. *R package version 1.1-7*. <http://cran.r-project.org/package=lme4>.
- Beider, Alexander. 2015. *Origins of Yiddish Dialects*. New York: Oxford University Press.
- Birnbaum, Solomon Asher. 1923. Übersicht über den jiddischen Vokalismus. *Zeitschrift Für Deutsche Mundarten* 18(1/2). 122–130.
- Birnbaum, Solomon Asher. 1934. Di historiye fun di alte U-klangen in Yidish [The history of the u-sounds in Yiddish]. *YIVO Bleter* 6. 25–60.
- Birnbaum, Solomon Asher. 1979. *Yiddish: A survey and a grammar*. Toronto: University of Toronto Press.
- Bleaman, Isaac L. & Chaya R. Nove. In press. *The corpus of spoken Yiddish in Europe: Goals, methods, and applications*. Language Documentation & Conservation.
- Chang, Charles B. 2011. Systemic drift of L1 vowels in novice L2 learners. In Wai-Sum Lee & Eric Zee (eds.), *Proceedings of the 17th international congress of phonetic science (ICPhS XVII)*, 428–431. Hong Kong: International Phonetic Association.
- Cooper, Levi. 2019. Polish hasidism and Hungarian orthodoxy in a borderland: The Munkács rabbinate. In François Guesnet, Howard Lupovitch & Antony Polonsky (eds.), *Poland and Hungary: Jewish realities compared* (Polin: Studies in Polish Jewry 31), 199–224. New York: Oxford University Press. DOI: 10.3828/liverpool/9781906764715.003.0010.
- Garellek, Marc. 2020. Phonetics and phonology of schwa insertion in Central Yiddish. *Glossa: A Journal of General Linguistics* 5(1). 1–25. DOI: 10.5334/gjgl.1141.
- Glasser, Paul. 2008. Regional variation in southeastern Yiddish historical inferences. In Marvin Herzog, Ulrike Kiefer, Robert Neumann, Wolfgang Putschke & Andrew Sunshine (eds.), *EYDES: Evidence of Yiddish documented in European societies*, 71–83. Tübingen: Max Niemeyer Verlag.
- Hay, Jennifer, Paul Warren & Katie Drager. 2006. Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics* 34(4). 458–484. DOI: 10.1016/j.wocn.2005.10.001.
- Herzog, Marvin I. 1965. *The Yiddish language in Northern Poland: Its geography and history*. The Hague: Mouton.

- Jacobs, Neil G. 1990. Northeastern Yiddish gender-switch: Abstracting dialect features regionally. *Diachronica* 7(1). 69–100. DOI: 10.1075/dia.7.1.05jac.
- Jacobs, Neil G. 1993. Central Yiddish Breaking and Drawl. In David Goldberg (ed.), *The field of Yiddish: Studies in Yiddish language, Folklore and literature, fifth collection*, 99–119. Evanston, Illinois: Northwestern University Press.
- Jacobs, Neil G. 2005. *Yiddish: A linguistic introduction*. Cambridge: Cambridge University Press.
- Jelinek, Yeshayahu A. 2007. *The Carpathian diaspora: The Jews of Subcarpathian Rus' and Mukachevo, 1848–1948*. Distributed by Columbia University Press. New York: East European Monographs.
- Kahle, David & Hadley Wickham. 2013. ggmap: Spatial visualization with ggplot2. *The R Journal* 5(1). 144.
- Katz, Dovid. 1982. *Explorations in the history of the Semitic component in Yiddish*. University of London.
- Keren-Kratz, Menachem. 2019. Global politics and the shaping of Jewish religious identity: The case of Hungary and Galicia. *Jewish Political Studies Review* 30(3/4). 100–119.
- Komoróczy, Sonja Rahel. 2018. Yiddish in the Hungarian setting. In Lily Kahn (ed.), *Jewish languages in historical perspective*, 92–108. Boston: Brill.
- Krogh, Steffen. 2012. How Satmarish is Haredi Satmar Yiddish? In Marion Aptroot, Efrat Gal-Ed, Roland Gruschka & Simon Neuberg (eds.), *Leket: Yidishe Shtudyes Haynt/Jiddistik Heute/Yiddish studies today*, 483–506. Düsseldorf: Düsseldorf University Press.
- Kuznetsova, Alexandra, Per B. Brockhoff & Rune H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82(13). 1–26. DOI: 10.18637/jss.v082.i13.
- Labov, William, Sharon Ash & Charles Boberg. 2006. *The atlas of North American English: Phonetics, phonology and sound change*. Boston: de Gruyter. DOI: 10.1515/9783110167467.
- Labov, William & Maciej Baranowski. 2006. 50 Msec. *Language Variation and Change* 18(3). 223–240.
- McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner & Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kald. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 498–502. DOI: 10.21437/Interspeech.2017-1386.
- Nove, Chaya R. 2021. *Phonetic contrast in New York Hasidic Yiddish vowels: Language contact, variation, and change*. The Graduate School & University Center, City University of New York. (Doctoral dissertation).

- Nove, Chaya R. 2023. Regional differences in the length contrast of Central Yiddish vowels. In *Paper presented at the Linguistic Society of America (LSA) Annual Meeting*. Denver, CO.
- Nove, Chaya R. & Benjamin Sadock. Submitted. *Regional differences in Central Yiddish vowel length: Central Poland and the Underland*.
- Nycz, Jennifer R. & Lauren Hall-Lew. 2013. Best practices in measuring vowel merger. In *Proceedings of Meetings on Acoustics*. San Francisco. DOI: 10.1121/1.4894063.
- Polonsky, Antony. 2010. *Warsaw*. YIVO Encyclopedia of Jews in Eastern Europe.
- Prilutski, Noah. 1920. *Tsum yidishn vokalizm: Yidishe dyalektologishe forshungen. Materyaln far a visnahftlekher gramatik un far an etimologish verterbukh fun der yidisher shprakh*, vol. 4. Warsaw: Nayer Farlag.
- R Core Team. 2021. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.r-project.org/>.
- Rubin, Israel. 1972. *Satmar: An island in the city*. New York: Crown Publishing Group.
- Sadock, Benjamin & Alyssa Masor. 2018. Bobover Yiddish: “Polish” or “Hungarian”? *Journal of Jewish Languages* 6. 89–110.
- Schäfer, Lea. 2022. A language is its dialects: The geolinguistics of Eastern Yiddish. In Carmen Reichert, Bettina Bannasch & Alfred Wildfeuer (eds.), *Zukunft der Sprache – Zukunft der Nation? Verhandlungen des Jiddischen und Jüdischen im Kontext der Czernowitz Sprachkonferenz* (Condition Judaica 97), 37–68. Berlin: de Gruyter Oldenbourg.
- Stanley, Joseph A. 2020. *joeyr: Functions for vowel data*. <https://github.com/JoeyStanley/joeyr>.
- Švorc, Peter. 2020. *Budapest or Prague? Jews in Eastern Slovakia and Subcarpathian Rus' at the turn of the 20th century*. Prešov: UNIVERSUM-EU.
- USC Shoah Foundation Visual History Archive. 2022a. *17 survivors from Poland. Tapes 1 and 2*. USC Shoah Foundation. [yiddishcorpus.org](http://yiddishcorpus.org).
- USC Shoah Foundation Visual History Archive. 2022b. *Unterland corpus: 21 survivors from Austria-Hungary. Tapes 1 and 2*. USC Shoah Foundation. [yiddishcorpus.org](http://yiddishcorpus.org).
- Weinreich, Max. 1973. *Di Geshikhte fun der yidisher Shprakh*. New York: YIVO Institute for Jewish Research.
- Weinreich, Uriel. 1964. Western traits in transcarpathian Yiddish. In Lucy S. Dawidowicz, Alexander Erlich, Rachel Erlich & Joshua A. Fishman (eds.), *For Max Weinreich on his seventieth birthday*, 245–264. The Hague: Mouton de Gruyter.

*Chaya R. Nove & Benjamin Sadock*

Weinryb, Bernard D. 1973. *The Jews of Poland: A social and economic history of the Jewish Community in Poland from 1100 to 1800*. Philadelphia: The Jewish Publication Society of America.

# Chapter 7

## Tracking language change in real time: Challenges for community-based research in the 21<sup>st</sup> century

Katharina Pabst<sup>a,b</sup>, Sam Brunet<sup>b</sup>, Alison L. Chasteen<sup>b</sup> & Sali A. Tagliamonte<sup>b</sup>

<sup>a</sup>Radboud University <sup>b</sup>University of Toronto

In this paper, we discuss our experiences conducting a longitudinal research study involving trend and panel data in the 21<sup>st</sup> century. Our goal is to document both challenges and opportunities for researchers engaging in real-time community-based research. The data for this project come from the *Language in Later Life* project, an interdisciplinary research project investigating the language of healthy adults in Toronto during the transition from later life to retirement. We focus on strategies for recruiting panel speakers and finding matches for our trend sample. One key finding is that it is not just crucial to find ways to keep in touch with participants, but also with research assistants and fieldworkers, who were essential for (re-)locating participants almost twenty years after the first point of data collection. While our project was conducted in 2018–2019, i.e., before the COVID-19 pandemic, we argue that many of the obstacles we encountered still apply, and might even be exacerbated, making it more important than ever to reflect on our methods for tracking language change in real time.

### 1 Background

Ever since Labov's foundational study of sound change in Martha's Vineyard (Labov 1963), many researchers have been using synchronic data from different age groups to study language change in progress. However, in recent years, there



has been an increasing interest in supplementing this work with real time studies. The reason for this is twofold: For one thing, real time data can provide insights into how language change proceeds over time and within the same individuals. Moreover, there are more opportunities for this kind of work thanks to the increasing number of community-based studies that can be revisited as well as alternative data sources that have become available, both online and through archives (Sankoff 2018: 298).

Generally, researchers distinguish two types of real time studies: panel studies, which follow the same individuals over time, and trend studies, which examine the behavior of different individuals who are matched for social factors such as age, gender, and education across two or more points in time (Sankoff 2005: 1003).

Although real time studies may be more feasible than ever before, Cieri & Yaeger-Dror (2017: 53) point out that they are still inherently “difficult, time consuming, and expensive” since it takes tremendous financial and organizational resources to track down the same individuals, especially after several years or decades, or to recruit participants that match the original sample in the case of trend studies. Despite these difficulties, a growing number of studies have succeeded in collecting real time data in a variety of settings, including Swedish in Eskilstuna (Sundgren 2001), Brazilian Portuguese in Rio de Janeiro (Naro & Scherre 2003), French in Montreal (Sankoff & Blondeau 2007, Wagner & Sankoff 2011, Sankoff & Wagner 2020), Danish in Copenhagen and various smaller communities (Gregersen 2009, Gregersen et al. 2009), English in Springville (Cukor-Avila & Bailey 2011, 2017), English in Tyneside (Buchstaller 2016, Buchstaller et al. 2017, Mechler & Buchstaller 2019), and Swabian in Stuttgart and Schwäbisch Gmünd (Beaman 2021a,b, Beaman et al. 2021, Beaman & Tomaschek 2021). However, practical advice about how researchers have managed to accomplish this herculean task is rare. One notable exception is Gillian Sankoff’s (2017) paper on the origins of the Montreal French project, one of the first and most comprehensive longitudinal corpora in the field, which was conceptualized in the early 1970s. In this paper, Sankoff outlines many of the challenges in building the original corpus as well as the follow-up studies, including the development of sampling strategies and guidelines for ensuring confidentiality and research access. She also discusses many challenges that are more specific to longitudinal projects, including the difficulties in tracing participants many years after the original project, attrition due to mortality and changed life circumstances, and how lack of resources prohibited long-term planning for research in the future.

Based on our own experiences, we know that anticipating these kinds of challenges can make a world of difference for researchers hoping to engage in real time studies, which is why we decided to build on Sankoff’s paper by sharing

our experiences in conducting our own longitudinal, real-time study. The goal of this paper is to document the challenges we faced, outline our strategies for overcoming them, and share practical advice for success. We note that our data was collected in 2018–2019, i.e., before the COVID-19 pandemic, but many of the same challenges still apply in the post-COVID era. In fact, we would argue that the ongoing public health situation likely exacerbates many of the obstacles we experienced.

## 2 Our starting point: The Toronto English Archive (TEA)

The starting point for our longitudinal work is Sali Tagliamonte's Toronto English project (Tagliamonte 2003–2006). Like the Montreal French project, this project was not originally planned to be longitudinal. The foundations for this work were laid in the early 2000s: Having returned to Toronto after a 20-year hiatus, Tagliamonte noticed tremendous changes were taking place in the local variety of English and decided to document the evolving situation in a community-based study focused on language change in progress. Most of the data collection for the project took place in 2003–2004. Smaller data samples were completed in 2002, 2004 and 2006 in course-based experiential learning projects focused on adolescents and young adults. In order to be included in the study, participants had to be born and raised in Toronto and speak English as their first or dominant language.

## 3 The Language in Later Life project

### 3.1 Project goals

Over fifteen years later, after several years of discussion, two researchers, Sali Tagliamonte (a linguist) and Alison Chasteen (a social psychologist) developed a plan to investigate the language of healthy adults in the transition from retirement to later life (Chasteen & Tagliamonte 2018–2020). The research project entitled *Sociolinguistic and Psychological Impacts of Language in Later Life* was designed specifically as both a window into language variation and change in later life, which is severely understudied, and as a way to learn more about socio-cognitive adjustments and experiences of individuals later in life. The data collection for this project mostly took place in 2018–2019. Together with our research

team, which included Graduate Student Project Manager Katharina Pabst and Undergraduate Student Project Manager Samantha Brunet, we set out to collect both panel and trend data from people who were born and raised in Toronto. The project's broad goal was to reinterview as many individuals as possible who were 30 years or older at the time of the original interviews in the early 2000's. In practice, this meant we were looking for 99 individuals. Given previous accounts of how difficult it is to track people down after many years, we cautiously proposed to re-interview at least 14 individuals who were part of the original study.

The second goal was to conduct a trend study. To keep this part of the project manageable, we aimed to find matches for everyone who was 50 years old or older in 2002–2004 with a new sample in 2018–2020. This meant seeking out 51 new individuals who were 50 years or older. To provide a comparative perspective on our panel participants, we also aimed to recruit matches for them that were consonant in terms of age, gender, and education, both as they were in 2002–2004 and as they were at the time of data collection in 2018–2019.

### **3.2 Project design**

The project comprises several sub-components, including a panel study (henceforth abbreviated with P) and a trend study (henceforth abbreviated with T). In the following, we will briefly describe these projects in more detail. In this context, numeral 1 refers to data that was collected in 2002–2004 and numeral 2 refers to data that was collected in 2018–2019.

The goal of the project was to allow for four types of comparison (see Table 1):

- The first comparison is a panel study, in which the same 14 individuals are interviewed in 2002–2004 and 2018–2019. During the second interview, panel participants were approximately 14–17 years older than in the original interview (P1–P2). This comparison offers insight into lifespan change.
- The second comparison is a trend study, including different individuals that were matched for age, gender, and education at two points in time (T1–T2). This comparison mirrors change in time.
- The third comparison involves the panel participants as they were in 2002–2004 and a set of social twins that matches them in terms of age, gender, and education as they were in 2002–2004 (P1–MP1). This comparison offers insight into the panel participants and their same age cohort in 2002–2004.

- Finally, the fourth and last comparison comprises the panel participants as they were in 2018–2019 and a set of social twins that matches them in terms of age, gender, and education in 2018–2019 (P2–MP2). This comparison offers insight into how the panel participants in later life compare with their own age cohort in later life, in 2018–2019.

Table 1: Project design

Type of comparison	Time 1 (2002–2004)	Time 2 (2018–2019)
Panel: Same individuals	P1	→P2
Trend: Different individuals, same age, gender, and education	T1	→T2
Cross-sectional comparisons: Different individuals, same age, gender, and education as P1 and P2		→MP1, MP2 ←

### 3.3 Recruitment strategies

The biggest challenge for our project was to find the people who were to be interviewed a second time, which would enable us to probe lifespan change. Table 2 provides an overview of our recruitment strategies, in order of their success. There were 99 individuals in the original sample who met the criteria for a second interview. We note that some participants were found using more than one method, which is why the numbers do not add up to 99.

First, we conducted obituary searches only to find that 22 individuals had died between the original interviews and the time of data collection. Given the advanced age of our target population and consistent with Sankoff's (2017) reports of attrition due to mortality, this was not unexpected. However, we emphasize how significant this fact is in terms of limiting the number of possible participants in a real time study across many years.

Our most successful strategy was getting in touch with former participants by seeking out and contacting the original interviewers for more information and also for their help in recruiting the second interviews. Most of the original fieldworkers were students and research assistants in Sali Tagliamonte's lab in the early 2000s and many of them had interviewed individuals in their social networks. Fortunately, they had often kept in touch with the same individuals over the years. Further, they had maintained connections with Sali Tagliamonte.<sup>1</sup> Using this method, we were able to find six participants, four of whom agreed to a second interview, for a success rate of 66.7%.

Because the original sampling strategy targeted specific neighborhoods in Toronto, we had kept records of participants' addresses. Therefore, we were able to use a reverse address lookup website named 411.ca where one can enter a Canadian address and determine if a name and/or phone number are attached to that address.<sup>2</sup> In total, we found 23 participants this way. We successfully connected with twelve of them and managed to recruit six of them for the second interview, an overall success rate of 26.1%.

Another successful strategy was using LinkedIn, a professional networking website. Samantha Brunet created a profile specifically for this purpose and sent potential candidates a contact request explaining the nature of our project. We found ten former participants this way, two of whom were successfully contacted. In the end, one agreed to a second interview, for a success rate of 10%.

We also sent letters to all fifty-three individuals for whom we had valid addresses and who we believed to still be alive using their last known addresses from 2002–2004. We successfully got in touch with fourteen individuals this way, three of whom agreed to return, resulting in a success rate of 5.7%. A substantial number of letters were returned to us. Oftentimes, the returned envelopes

---

<sup>1</sup>We do not mean to make light of how difficult this strategy was. Only a handful of the original interviewers were willing or able to help given their jobs and family situations. One former interviewer had to be flown in from a US city for the weekend. However, a side product was the pleasant reconnection and updates between Sali Tagliamonte and her former students, suggesting that maintenance of social ties with research team members is an asset for community-based research.

<sup>2</sup><https://411.ca> Data for this paper retrieved Monday, January 9, 2023 6:44am.

included helpful information. For example, one letter stated that the addressee passed away in 2010. Another one mentioned that the addressee had moved years ago, indicating that the address we had on file was no longer usable. Taken together, the relatively low success rate of this approach leads us to conclude that it would be promising to collect email addresses for participants in future research projects since email contact information tends to remain more stable than geographic location.

Other strategies we used included looking for individuals on Facebook or searching for them on Google. These approaches allowed us to find five individuals each; however, not a single one was contacted successfully. For Facebook, this is likely due to the site's privacy settings, which automatically route messages from individuals outside of one's network into a spam folder, which most people rarely check. For the remainder, it is difficult to assess the poor outcome because we have no way of knowing if the communications were successfully received by the individuals intended.

Despite the extraordinary difficulties in tracking former participants, we managed to find and communicate with 34 of them. However, only 14 of them agreed to be part of our study. Those who declined to participate offered a wide variety of reasons, ranging from having very busy schedules to no longer living in the area. In many cases, the participants were quite old and felt unable to put in the requisite time and energy. In one case, a former participant had developed cancer and was undergoing intensive chemotherapy. One person was adamant that they had "nothing to talk about," while another felt self-conscious about the messiness of their apartment. The latter concern is why we decided to offer participants the opportunity to come to the lab instead if they preferred, which gained us several participants. Next, we turn to the strategies we used to recruit the trend participants and social twins for the panel participants. Our original goal was to recruit all individuals using the Adult Volunteer Pool, a database maintained by the Psychology Department at the University of Toronto, which has contact information for individuals 50 years or older who have also indicated interest in participating in paid research studies. We were able to recruit 22 individuals this way. However, one challenge that we did not anticipate was how difficult it was to find matches for the social characteristics. The psychology database does not include information on participants' first language, or where they were born and raised – both of which were important selection criteria for our study. Given that Toronto has had tremendous in-migration in the past two decades, many of the participants in the Adult Volunteer Pool were not born in the area or moved to Toronto before the age of 10. Another complication was that most individuals in this database are over 55 years old, making it difficult

to find matches for the younger panel speakers who were in early adulthood or middle-aged in 2002–2004.

Table 2: Overview of recruitment strategies for panel participants.

Success rate	Notes
<b>Connection to former interviewers</b>	
Found: 6	Former students who conducted
Successfully contacted: 6	the interviews in 2002–2004
Recruited: 4	
Success rate: 66.7%	
<b>411.ca</b>	
Reliable phone numbers: 23	Reverse address lookup
Successfully contacted: 12	
Recruited: 6	
Success rate: 26.1%	
<b>LinkedIn</b>	
Found: 10	Professional networking website,
Successfully contacted: 2	created profile and sent
Recruited: 1	participants a contact request
Success rate: 10%	
<b>Letters to addresses from 2002–2004</b>	
Letters sent: 53	Used addresses we had
Successfully contacted: 14	on file from 2002–2004
Recruited: 3	
Success rate: 5.7%	
<b>Facebook</b>	
Found: 5	All five participants who were
Successfully contacted: 0	found were closer to 30 during
Recruited: 0	the original interviews;
Success rate: 0%	privacy settings forbid direct messaging
<b>Google search</b>	
Found: 5	Last resort, only useful
Successfully contacted: 0	in that we found one
Recruited: 0	participant who had moved away
Success rate: 0%	

Table 3: Overview of recruitment strategies for trend participants.

Method	Number of individuals recruited	Notes
Adult Volunteer Pool	22	Database used by Psychology Department at U of T, has contact info for individuals 50+
Forever Young magazine	32	Magazine for individuals 50+ who live in the GTA
Social media (Facebook, Reddit)	10	Toronto-specific Facebook groups, r/Toronto
Friends of friends	6	Research team reached out to friends who matched criteria of participants we did not have matches for

Therefore, we pursued additional avenues for recruitment. For example, we posted two ads in “Forever Young”, a magazine distributed across the Greater Toronto Area, aimed at older adults age 50+ and primarily read by individuals more than 70 years old. In both May and July 2019, we purchased advertising space. In May, we posted an ad seeking older adults aged 50–95 and received hundreds of calls and emails from interested locals. In July, we more specifically advertised to people aged 50–65, and received considerably less attention (only around 25 calls/emails). In total, we were able to secure 32 matches this way. We found 10 more matches by posting on Toronto-specific Reddit and Facebook groups. The remaining six participants were recruited by reaching out to friends who matched the criteria of participants we had not yet successfully matched.

### 3.4 Data

Table 4 shows our final participant sample. In the end, we were able to recruit 14 panel participants, exactly meeting our original goal. We were further able to

recruit 50 out of 51 trend participants we were aiming for. The original sample included one participant who did not finish high school, which is much rarer nowadays due to laws governing minimum education requirements in Ontario and Canada more generally, preventing us from finding a suitable match. As for the social twins, we succeeded in finding matches for our panel participants as they were in 2018–2019, but did not find matches for all 14 of them as they were in 2002–2004. However, Sali Tagliamonte is currently engaged in another large-scale project in Toronto (Tagliamonte 2018–2024) that targets younger speakers, some of whom will hopefully be able to fill these gaps.

We should note that most individuals in our study are considered white according to census criteria (but they comprise a mix of British, European, and Asian backgrounds). There is only one biracial black and white speaker in the panel sample. In contrast, due to shifting demographics in Canada in the intervening 20 years or so, the trend sample has representation of other ethnic groups, but even here the overwhelming majority is also white (as defined above). The criterion of having been born and raised in the city of Toronto continues to restrict the sample to a predominantly white population.<sup>3</sup> The sample is largely balanced between male- and female-presenting speakers, with female-presenting speakers being slightly overrepresented.

Table 4: Speaker sample

Type of comparison	Time 1 (2002–2004)	Time 2 (2018–2019)
Panel: same individuals	n=14 Age 34–73	n=14 Age 49–86
Trend: different individuals same age and gender	n=51 Age 51–92	n=50 Age 50–89
Cross-sectional comparisons		Panel Match 1 n=6 Age 50–73 Panel Match 2: n=14 Age 47–90

In sum, we used different recruitment methods for different participant groups, relying heavily on existing information about the participants we had collected during the original interviews from the early 2000s. The strategy that led to the most success was networking with former fieldworkers. Taken together, these

<sup>3</sup>In more recent research based in a high school, sampling included locally born and raised youth as well as those that had come from other countries to study the impact of immigrant populations on the local vernacular (Tagliamonte forthcoming).

two aspects of our methodology demonstrate two key messages for future research: 1) the critical need to document information in research, not simply about the participants, but also research assistants and fieldworkers; and 2) the importance of maintaining social ties with research assistants and even the participants themselves (see Wagner & Tagliamonte 2019).

The least successful strategy for recruitment was social media, likely because of the age group we were targeting, i.e., the older adult sector. Given the dramatic changes in social media platforms, certain generations may not be as engaged on certain platforms (e.g., Facebook or Twitter) nor possibly as open to communication in one modality or the other. Note that recent work shows that platforms such as Instagram can be crucial for targeting younger speakers (Nesbitt & Watts 2022). For the older adults we were seeking, targeted ads (especially those in print media) were more successful. The key message here is to determine the regular communication habits of the generational sector being targeted for study.

Drawing on existing participant pools such as the Adult Volunteer Pool can be incredibly helpful. However, if these are not run by linguists, they may not include the macrosocial information linguists need to decide whether speakers match inclusion criteria, such as *born and raised in Toronto*.

## 4 Community-based research during and after COVID-19

As mentioned earlier, all our trend and panel data were collected in 2018–2019, before the COVID-19 pandemic put a sudden end to most in-person data collection over the period 2019–2021. During this time many projects shifted to online data collection. Researchers interested in conducting a longitudinal research project might wonder how the pandemic may have affected the data from longitudinal research covering that period. In the years immediately following the pandemic shut-down, several research studies have shown that online data collection is a viable alternative (Carmichael et al. 2022, Hall-Lew et al. 2022, Leemann et al. 2020, Sneller 2022, Sneller et al. 2022). These studies confirm that online data collection was not disruptive to underlying linguistic patterns. Such research also demonstrates the feasibility of longitudinal work repeated sampling in communities that have been studied in the past.

In the summer of 2021, Zoom interviews with Toronto youth confirmed for us that at least when working with young people, ZOOM data is a viable alternative to in-person conversations. While interactional patterns in an online environment can be affected during speaker changes, when the participant talks, the language appears to be relatively unaffected by the modality (Tagliamonte,

p.c. 31-8-2021). Further, Gardner & Kostadinova's (2024) study demonstrates that there are no significant differences in the internal constraints on one of the most diagnostic stylistic variables, *-ing* variation, in a consistent comparison of frequency of forms and linguistic constraints across modalities. The efficacy of online interviewing was further confirmed in interviews done with older adult participants on a recent fieldtrip to northwestern Ontario (Tagliamonte p.c. 30-7-2022). In fact, older people who have mobility issues, find it difficult to travel, or are immunocompromised, greatly enjoyed participating in interviews from the comfort of their own homes. However, it should be noted that some participants required the support of a younger family member to help them with the technology. These studies demonstrate online modalities offer certain sectors of the population an ideal option for participating in linguistic studies.

## **5 Practical advice**

Doing longitudinal community-based work is challenging, but there are ways to ensure greater success. In this project, Sali Tagliamonte had kept detailed records of interview content and documentation of names and addresses of the interviewees and research assistants who conducted the interviews from the previous project.<sup>4</sup> While the original purpose was to create a foundation for future research and to maintain contacts in the speech community, this information made it possible to seek out participants for the current project. However, the time span (18 years), mobility of individuals, and changes in life circumstances made it very difficult nonetheless to find and recruit them. On a related note, Katharina Pabst, who conducted most of the interviews for this project, found that reading through the transcripts of the first interviews was also an ideal way to get a better idea of what topics panel speakers might be interested in talking about during the second interviews. For example, one participant mentioned being an avid photographer during the first interview. When Katharina entered the participant's house for the second interview, she noticed many gorgeous photos on the wall. Asking about these was an ideal way to break the ice and develop rapport with the participant.

Given how difficult it was to track down the original participants, we recommend adding the protocol of asking participants whether they can be contacted

---

<sup>4</sup>Our ethics protocol stipulated that we intended to contact previous participants from the 2003 study. While the original ethics protocol stipulated participant anonymity for research purposes, ongoing contact with the project director was encouraged. We acknowledge that data protection requirements may differ across countries and disciplines.

in future research studies and obtain information such as phone numbers, social media contacts, and email addresses.<sup>5</sup>

Many of the panel speakers we re-interviewed did not remember taking part in the original project, sometimes leading to suspicion on their part. While all the authors regularly give outreach talks to share the results of our work, a new aim of our research modus operandi will be to be more deliberate about finding ways to stay in touch with participants and share the results of unfolding research findings in our work. Public lectures, newsletters, and social media are an ideal means to make that happen.

Both for the sake of funding applications and your own mental health, we recommend setting realistic goals for the kind of community you are working with and the age group. Sankoff (2017) warned that attrition was the highest for older speakers, and as our project has shown, there was indeed substantial attenuation of the target population, not only due to mortality but also pervasive health issues that made participating in the project difficult for many different reasons.

Joining forces with researchers in social psychology has been an outstanding success, allowing us to collect new types of data that will eventually shed light on the relationship between cognitive development and language use in a way that has never been possible before. It has also taught us how to scientifically study the experiences of aging and ageism through sociolinguistic data, demonstrating how fruitful combining methods and resources can be (see Chasteen et al. 2022). The success of this research relies as much on careful planning as it does on an open-minded scientific stance on the part of the researchers, flexibility, and resourcefulness. We also greatly benefitted from informal conversations with other researchers who have conducted longitudinal, community-based research. Our goal in this paper has been to pay forward our experiences and engage in continued conversations about methodological challenges and opportunities.

## Acknowledgements

We gratefully acknowledge the support of the Social Science and Humanities Research Council (SSHRC), including Research Grant # 410-2003-0005 (Linguistic Changes in Canada Entering the 21<sup>st</sup> Century) to Sali A. Tagliamonte and Research Grant #430-2018-000026 (Sociolinguistic and Psychological Impacts of Language in Later Life) to Alison L. Chasteen and Sali A. Tagliamonte. Further,

---

<sup>5</sup>In current research practise, permission for future contact is stipulated in the consent procedures.

we thank the government of Ontario and the Department of Linguistics at the University of Toronto for an Ontario Trillium Scholarship to Katharina Pabst. Finally, we extend our gratitude to the many research assistants of the University of Toronto Variationist Sociolinguistics Lab and the Intergroup Relations Lab who collected and processed the data for this project.

## References

- Beaman, Karen V. 2021a. Exploring an approach for modelling lectal coherence. In Hans van de Velde, Nanna Haug Hilton & Remco Knooihuizen (eds.), *Language variation – European perspectives VIII: Selected papers from the tenth International Conference on Language Variation in Europe (ICLaVE10)*, 135–160. Amsterdam: John Benjamins.
- Beaman, Karen V. 2021b. Identity and place in linguistic change across the lifespan: The case of Swabian German. In Arne Ziegler, Stefanie Edler, Nina Kleczkowski & Georg Oberdorfer (eds.), *Urban matters: Current approaches of international sociolinguistic research* (Studies in Language Variation), 27–60. Amsterdam: John Benjamins.
- Beaman, Karen V., Harald Baayen & Michael Ramscar. 2021. Deconfounding the effects of competition and attrition on dialect across the lifespan: A panel study investigation of Swabian German. In Karen V. Beaman & Isabelle Buchstaller (eds.), *Language variation and language change across the lifespan: Theoretical and empirical perspectives from panel studies*, 235–64. New York, NY: Routledge.
- Beaman, Karen V. & Fabian Tomaschek. 2021. Loss of historical phonetic contrast across the lifespan: Articulatory, lexical, and social effects on sound change in Swabian. In Karen Beaman & Isabelle Buchstaller (eds.), *Language variation and language change across the lifespan*, 209–34. New York, NY: Routledge.
- Buchstaller, Isabelle. 2016. Investigating the effect of socio-cognitive salience and speaker-based factors in morpho-syntactic life-span change. *Journal of English Linguistics* 44(2). 199–229. DOI: 10.1177/0075424216639645.
- Buchstaller, Isabelle, Anne Krause, Anja Auer & Stefanie Otte. 2017. Levelling across the life-span? Tracing the FACE vowel in panel data from the North East of England. *Journal of Sociolinguistics* 21(1). 3–33. DOI: 10.1111/josl.12227.
- Carmichael, Katie, Lynn Clark & Jennifer Hay. 2022. Lessons learned: The long view. *Linguistics Vanguard* 8(s3). 353–362. DOI: 10.1515/lingvan-2021-0050.
- Chasteen, Alison L. & Sali A. Tagliamonte. 2018–2020. *Sociolinguistic and psychological impacts of language in later life. Research grant*.

- Chasteen, Alison L., Sali A. Tagliamonte, Katharina Pabst & Samantha Brunet. 2022. Ageist communication experienced by middle-aged and older Canadians. *International Journal of Environmental Research and Public Health* 19(4). 2004. DOI: 10.3390/ijerph19042004.
- Cieri, Christopher & Malcah Yaeger-Dror. 2017. Alternative sources of panel study data: Opportunities, caveats and suggestions. In Suzanne E. Wagner & Isabelle Buchstaller (eds.), *Panel studies of variation and change*, 53–72. New York, NY: Routledge.
- Cukor-Avila, Patricia & Guy Bailey. 2011. The interaction of transmission and diffusion in the spread of linguistic forms. *University of Pennsylvania Working Papers in Linguistics: Selected Papers from N W A V* 39 17(2). 41–49.
- Cukor-Avila, Patricia & Guy Bailey. 2017. The effect of small Ns and gaps in contact on panel survey data. In Suzanne E. Wagner & Isabelle Buchstaller (eds.), *Panel studies of variation and change*, 181–212. New York, NY: Routledge.
- Gardner, Matt Hunt & Viktorija Kostadinova. 2024. Gettin' sociolinguistic data remotely: Comparing vernacularity during online remote versus in-person sociolinguistic interviews. *Linguistics Vanguard*. DOI: 10.1515/lingvan-2022-0069.
- Gregersen, Frans. 2009. The data and design of the LANCHART study. *Acta Linguistica Hafniensia* 41. 3–29. DOI: 10.1080/03740460903364003.
- Gregersen, Frans, Marie Maegaard & Nicolai Pharao. 2009. The long and short of (ae)-variation in Danish: A panel study of short (ae)-variants in Danish in real time. *Acta Linguistica Hafniensia* 41. 64–82. DOI: 10.1080/03740460903364086.
- Hall-Lew, Lauren, Claire Cowie, Catherine Lai, Nina Markl, Stephen J. McNulty, Shan-Jan S. Liu, Clare Llewellyn, Beatrice Alex, Zuzana Elliott & Anita Klingler. 2022. The Lothian Diary Project: Sociolinguistic methods during the COVID-19 lockdown. *Linguistics Vanguard* 8 (s3). 321–330. DOI: 10.1515/lingvan-2021-0053.
- Labov, William. 1963. The social motivation of a sound change. *Word* 19(3). 273–309. DOI: 10.1080/00437956.1963.11659799.
- Leemann, Adrian, Péter Jeszensky, Carina Steiner, Melanie Studerus & Jan Messerli. 2020. Linguistic fieldwork in a pandemic: Supervised data collection combining smartphone recordings and videoconferencing. *Linguistics Vanguard* 6 (s3). DOI: 10.1515/lingvan-2020-0061.
- Mechler, Johanna & Isabelle Buchstaller. 2019. (In)stability in the use of a stable variable. *Linguistics Vanguard* 5 (s2). DOI: 10.1515/lingvan-2018-0024.
- Naro, Anthony Julius & Maria M. P. Scherre. 2003. Estabilidade e mudança lingüística em tempo real: A concordância de número. In Maria Da Conceição de Paiva & Maria E. Duarte (eds.), *Mudança lingüística em tempo real*, 47–62. Rio de Janeiro: Capa.

- Nesbitt, Monica & Akiah Watts. 2022. Socially distanced but virtually connected: Pandemic fieldwork with Black Bostonians. *Linguistics Vanguard* 8 (s3). 343–352. DOI: 10.1515/lingvan-2021-0049.
- Sankoff, Gillian. 2005. Cross-sectional and longitudinal studies. In Ulrich Ammon, Norbert Dittmar, Klaus J. Mattheier & Peter Trudgill (eds.), *Sociolinguistics: An international handbook of the science of language and society*, 2nd edn., vol. 2, 1003–1012. Berlin & New York, NY: de Gruyter Mouton.
- Sankoff, Gillian. 2017. Before there were corpora: The evolution of the Montreal French project as a longitudinal study. In Suzanne E. Wagner & Isabelle Buchstaller (eds.), *Panel studies of variation and change* (Routledge studies in language change), 21–52. New York, NY & London: Routledge.
- Sankoff, Gillian. 2018. Language change across the lifespan. *Annual Review of Linguistics* 4. 297–318. DOI: 10.1146/annurev-linguistics-011817-045438.
- Sankoff, Gillian & Hélène Blondeau. 2007. Language change across the lifespan: /r/ in Montreal French. *Language* 83(3). 560–588. DOI: 10.1353/lan.2007.0106.
- Sankoff, Gillian & Suzanne E. Wagner. 2020. The long tail of language change: A trend and panel study of Québécois French futures. *Canadian Journal of Linguistics* 65(2). 246–275. DOI: 10.1017/cnj.2020.7.
- Sneller, Betsy. 2022. COVID-era sociolinguistics: Introduction to the special issue. *Linguistics Vanguard* 8(s3). 303–306. DOI: 10.1515/lingvan-2021-0138.
- Sneller, Betsy, Suzanne Evans Wagner & Yongqing Ye. 2022. MI diaries: Ethical and practical challenges. *Linguistics Vanguard* 8(s3). 307–319. DOI: 10.1515/lingvan-2021-0051.
- Sundgren, Eva. 2001. Men and women in language change: A Swedish case study. *NORA: Nordic Journal of Feminist and Gender Research* 9(2). 113–123. DOI: 10.1080/080387401753355344.
- Tagliamonte, Sali A. 2018–2024. *Language change and social change in the early 21<sup>st</sup> century: Canadian English 2002 to 2020*. #435-2019-0053. Research grant. Social Sciences & Humanities Research Council of Canada (SSHRC).
- Tagliamonte, Sali A. 2003–2006. *Linguistic changes in Canada entering the 21<sup>st</sup> century*. Research grant. Social Sciences and Humanities Research Council of Canada (SSHRC). #4102003-0005.
- Tagliamonte, Sali A. Forthcoming. *My story in history: A high school story of Toronto adolescents*.
- Wagner, Suzanne E. & Gillian Sankoff. 2011. Age grading in the Montréal French inflected future. *Language Variation and Change* 23(3). 275–313. DOI: 10.1017/S0954394511000111.

Wagner, Suzanne E. & Sali A. Tagliamonte. 2019. What makes a panel study work? Researcher and participant in real time. In Suzanne E. Wagner & Isabelle Buchstaller (eds.), *Panel studies of variation and change*, 213–32. New York, NY: Routledge.



# Chapter 8

## Corpus-based Low Saxon dialectometry

Janine Siewert<sup>a</sup>, Yves Scherrer<sup>a,b</sup> & Martijn Wieling<sup>c</sup>

<sup>a</sup>University of Helsinki <sup>b</sup>University of Oslo <sup>c</sup>University of Groningen

In this corpus-based study, we explore how the similarity of Low Saxon dialects among each other and to the state languages Dutch and German has changed from the 19<sup>th</sup> century to today. In particular, we want to investigate if the traditional classification into an eastern and a western group is visible in the data and if the Low Saxon dialects can be found to diverge at the Dutch-German border.

We apply principal component analysis and hierarchical clustering to n-grams of characters, part of speech tags and morphological features and observe divergent developments at the separate levels. As a reflection of different orthographic traditions, a noticeable distance between Dutch Low Saxon and German Low Saxon can be attested at the character level. At the PoS and morphological level, we however find a particular closeness between Dutch Low Saxon and the northern dialects from Germany, while we see German Westphalian in an outlier position. A shift towards the state languages can be observed at the PoS level, but the overall distance between Dutch Low Saxon and German Low Saxon does not seem to markedly increase at the three levels we studied.

### 1 Introduction

In the context of a research project on dialectal variation in Low Saxon<sup>1</sup>, we investigate how the similarity of larger dialect areas has changed from the 19<sup>th</sup> to the 21<sup>st</sup> century. Our study is based on corpus data that we have collected earlier and covers both the Dutch and the German side of the Low Saxon language area.

---

<sup>1</sup>Also referred to as “Low German”.



Traditional dialect classifications, such as the ones presented by Schröder (2004), have largely relied on the occurrence of certain phonological and morphological traits. In this study, we however take a corpus-based approach and zoom in on three levels of the language: orthographic, morphological, and syntactic. Among possible differences at these three levels, we particularly want to investigate how the corpus-based results relate to the more traditional classifications and to the language contact situation. Therefore, we are interested in 1) the traditional east-west division, and 2) what the effect of the Dutch-German border is on the development of Low Saxon in recent decades.

## 2 Low Saxon dialect classification and variation

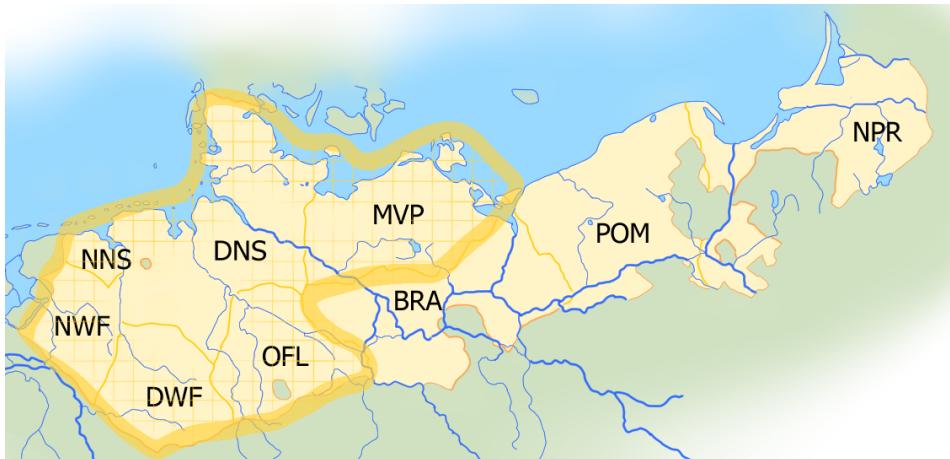


Figure 1: Low Saxon dialects: NNS: Dutch North Saxon, NWF: Dutch Westphalian, DNS: German North Saxon, DWF: German Westphalian, MVP: Mecklenburgish-West Pomeranian, OFL: Eastphalian, BRA: Brandenburgish, POM: East Pomeranian, NPR: Low Prussian; self-created map

Low Saxon is a West-Germanic language primarily spoken in northern Germany and the north-eastern Netherlands. Despite official recognition in both countries, there is no common standard variety. As a result, Low Saxon speakers will typically speak and write their own dialect(s) in all language use contexts.

Figure 1 shows the Low Saxon language area in the early 20<sup>th</sup> century with its major dialect groups. In this paper, we will focus on the encircled dialects.

The dialect division of German Low Saxon follows the traditional classification presented by Schröder (2004) and Stellmacher (1983). This classification is pri-

marily based on phonological and morphological traits such as vowel phoneme mergers and the plural ending of verbs in the present tense. Based on this plural suffix, Stellmacher (1983: 240) divides German Low Saxon into an eastern and a western group, where *-et* is used in the west and *-en* in the east. Of the dialects included in our research, only Mecklenburgish-West Pomeranian (MVP) belongs to this eastern group.

A more fine-grained division is normally used for Dutch Low Saxon (see for instance Bloemhoff et al. 2019: 20), possibly due to the much smaller size of the language area. In this cross-border comparison, we wanted to keep the size of the areas of comparison roughly in line with one another, but at the same time be able to look at internal differences within Dutch Low Saxon. As a compromise, we opted for a two-fold North-South division. Dutch North Saxon (NNS) is only represented by texts from Groningen in our dataset, while the southern part, called “Dutch Westphalian” (NWF) here, includes texts from Achterhoeks, Veluws, Twents, Sallands and Drents. This North-South division has a rough correspondence with German North Saxon (DNS) and German Westphalian (DWF).

An important isogloss for dividing North Saxon from Westphalian is the merger of Proto-Germanic \*â and lengthened \*a in the north contrasting with the preservation of distinct phonemes in the Westphalian dialects (Niebaum 2008, Bloemhoff et al. 2019). This distinction can still be found on both sides of the border. In addition, the northern group of Dutch Low Saxon has a Frisian substrate which is likely why Jellinghaus (1892) refers to it as “Friso-Saxon” and contrasts it with “Saxon” dialects.

## 2.1 Orthographic differences

In addition to dialectal variation, we find a large variety of spelling traditions and personal writing habits. While regional language institutes or societies, publishers or individual language enthusiasts devise and spread their own spelling systems today, explicit normification was not this widespread in the 19<sup>th</sup> century. Even today, every Low Saxon speaker does not adhere to the spelling that a local publisher from the same region might prefer. For instance, while textbooks from Mecklenburg-West Pomerania (MVP) (e.g., Herrmann-Winter 2006) tend to use the spelling by Herrmann-Winter, and textbooks from several areas of German North Saxon (DNS) (e.g., Arbatzat 2016, Hiestermann & Konen-Witzel 2021) can often be seen to follow the SASS<sup>2</sup> spelling (Kahl & Thies 2009), some of the people who provided texts for our dataset had other spelling preferences.

---

<sup>2</sup>Not an abbreviation but based on the family name of Johannes Saß.

Local spelling traditions tend to draw inspiration from the majority language orthography to different degrees and in different ways. We will illustrate this with a few examples from our dataset. Sentence (1) is from the Netherlands and sentences (2) and (3) from Germany. This can easily be recognised by the capitalisation of nouns (e.g., *Kaarten*, *Ogen*, *Dage*, *Magister*) or lack thereof (*leu*, *weerde*, *ding*). While not all Low Saxon writers from Germany capitalise nouns, it is a common trait.

- (1) De leu veult gemènlik eerst de weerde van 'n ding, as ze 't nig hebt.  
'In general, people only become aware of the value of things when they do not have them.'
- (2) Arfest neem twe Kaarten to de eerst Klaß, un as ik daröver grote Ogen maak, lach he un meen, dat kunn darop staan, ik schull man instigen.  
'Arfest bought two first class tickets, and when I looked on at him with astonishment, he laughed and said that may be written there but I should just get on.'
- (3) Eunige Dage später frogere de Magister, biu de veuer Johrestyien herren.  
'A few days later, the teacher asked what the four seasons were called.'

Sentence (1) furthermore shows two grapheme-phoneme correspondences typical for Low Saxon writing from the Netherlands: <eu> for /ø/ and <z> for /z/. In texts from Germany, one would rather expect <ö(ö)> and <s>.

In Sentence (2) we find the letter <ß> for /s/,<sup>3</sup> again a strong indicator of the text coming from Germany as this letter is likely not used in Dutch Low Saxon. In addition, we see a difference in the adaptation of German spelling rules between Sentence (2) and (3). In the word *Johrestyien* 'seasons' (compare German *Jahreszeiten*) in (3), the <h> is used as part of a digraph marking a long vowel. According to some local spellings such as the SASS (Kahl & Thies 2009), the <h> should indeed be used as a length marker if it is written in the German cognate. However, such a principle is not followed by all Low Saxon writers in Germany as can be seen in Sentence (2). If this writer had adhered to the same rule, we would find <nehm> 'took' and <stahn> 'stand' instead of <neem> and <staan>, compare *nahm* and *stehen* in German.

---

<sup>3</sup>Here used according to the old German orthography, where the <ß> replaced <ss> in coda position.

## 2.2 Language contact

Another layer of variation is added by the influence from the different state languages, Dutch and German. Goossens (2019) indeed observed a replacement of certain local characteristics with features from the majority languages. He finds this to primarily affect the lexical level, but influence can be found at the phonological, morphological, and syntactic level as well.

Influence on grammatical gender, a morphological feature relevant for our current research, is described by Bloemhoff et al. (2019: 107–109). Whereas in Dutch, masculine and feminine gender have merged to a common gender, there are Dutch Low Saxon dialects such as Twents, where this distinction has been preserved. Bloemhoff et al. (2019) however find that speakers of Twents are increasingly unsure which of the two genders to assign to nouns with common gender in Dutch.

In his comparison of two neighbouring dialects at the border, Smits (2011) observes two different tendencies. While he finds more structural language loss, i.e., convergence towards the state language, on the Dutch side of the border, there is more overall language loss, i.e., language shift from Low Saxon to German, on the German side. Nevertheless, the remaining speakers on the German side retain more structural characteristics of their dialect than their neighbours from the Netherlands.

## 2.3 Morphological and syntactic traits

Next, we will discuss six morphological and syntactic traits where Low Saxon dialects differ from each other and/or from the respective majority language. We focus on the kind of structures that are visible at the PoS and morphological level and where we expect the majority languages to exert influence. The traits are presented in the same order in which we discuss them in Section 5.2, Section 5.3 and Section 6.

### 2.3.1 Lack of the definite article

According to Pheiff (2022), there are Low Saxon dialects where the usage of the definite article is less frequent than in Standard Dutch and German. He observes this to be particularly true for the north of Groningen but has attested the lack of a definite article in contexts where it would be obligatory in the majority languages in other regions as well.

For instance, our dataset contains the following sentence in older German North Saxon (DNS):

- (4) As de Bur 'n Sett up Bedd wäsen weer.  
when the farmer a moment on bed be.PAST-PTCP be.PST.3SG  
'When the farmer had been on the bed for a moment.'

This sentence illustrates the tendency mentioned by Pheiff (2022: 147) that the definite article is missing particularly after prepositions: While no article is used after *up*, it does occur in *de Bur* and *'n Sett*. Proper nouns are another context where the usage of the definite article is less frequent. This is in line with the geographical distribution of the definite article with proper nouns within German as well, where the use of the definite article with personal names decreases the further north you go (Pheiff 2022: 150).

### 2.3.2 Motion verb + (to/and) + infinitive

German and Dutch combine motion verbs with the construction *um zu/om te* to express finality. In Low Saxon, while usage of *üm to* is possible, plain *to* similar to English occurs commonly as well:

- (5) He keem na de Köök to fröhstücken.  
he come.PST.3SG to the kitchen to eat\_breakfast  
'He came to the kitchen to eat breakfast.' (Thies 2010: 74)

In addition, in a few dialects, the infinitive can be connected with *un/en* 'and'. Thies (2010: 76) mentions this construction for Schleswig and assumes Danish influence but, for instance, Wilhelm Wisser from Eastern Holstein uses it in his collection of fairy tales as well:

- (6) Hê schall mal na ehr'n Brôder Män gahn un fragen  
he shall.3SG once to her-M.ACC brother moon go and ask  
den' mal.  
that.M-ACC once  
'He should just go to her brother moon and ask him.' (Wisser 1921: 53)

Furthermore, this construction is attested in Dutch Low Saxon dialects such as Stellingwerfs: *Ie kun mar beter naor de zoolder gaon en vang(en)* die moes 'You better go to the attic and catch the mouse.' (Bloemhoff 2008: 191) and Gronings: *ik goa hin en helpn Kloas* 'I will go and help Kloas.' (van Bree 2008: 114).

### 2.3.3 Case inflection

Low Saxon dialects differ noticeably in how many cases nouns inflect for. In Medieval Low Saxon, we still find nominative, genitive, dative, and accusative (Lasch 1974), but the modern varieties typically display a simplified case system. According to Lücht (2016: 62), East Frisian<sup>4</sup> does not inflect nouns for case. Usage of independent genitive forms is generally very restricted, but most German Low Saxon dialects still distinguish between nominative and accusative in singular masculine nouns, which can be seen in the form of the adjective and the article (Lindow et al. 1998: 144, 191).

Many German Low Saxon dialects have preserved remnants of the dative case that might be used in connection with certain prepositions. In the German North Saxon (DNS) Sentence (7) from our dataset, after *bi*, we see a form of the definite article resembling the masculine accusative. As the expected form in the accusative of this neuter noun would be '*t*', this can be interpreted as a reduced version of the older dative form '*m*'.

- (7) Mi weer de Sunn to grall bi 'n Läsen.  
 me was the sun too bright at the.DAT.SG reading.  
 'The sun was too bright for me while reading.'

*To* is another preposition where we often encounter the definite article of neuter nouns as '*n*' instead of otherwise expected '*t*', e.g., in this German North Saxon (DNS) sentence from our dataset: *Dat's to'n Lachen!* 'That is ridiculous!'

Independent productive dative forms have only been preserved in a few southern German Low Saxon dialects such as South Westphalian (part of DWF). While most northern dialects in Germany distinguish between *de Disch* 'the table' in the nominative and *den Disch* in the non-nominative, parts of German Westphalia exhibit a three-fold distinction between *de Disk*<sup>5</sup> in the nominative, *dem Diske* in the dative, and *den Disk* in the accusative (Lindow et al. 1998: 144–145).

Case distinctions are not commonly mentioned in descriptions of Dutch Low Saxon and neither have we encountered any in our data. Individual fossilised cases in fixed expressions similar to Dutch might however still exist.

<sup>4</sup>Part of German North Saxon (DNS). Not to be confused with the East Frisian language Saterland Frisian.

<sup>5</sup>The nominative form is not given by Lindow et al. (1998: 145), but based on the other forms they present, it likely looks like this.

### 2.3.4 Subjunctive

Within German Low Saxon, there is a North-South divide in subjunctive usage. As can be seen in the i-mutated forms *söl* and *bekäme* in Sentence (8) from German Westphalia, a few southern Low Saxon dialects have preserved distinct subjunctive forms.

- (8) Et söl mi frögn, wank et bekäme.  
it shall.PST.SBJV.3SG me please if-I it get.PST.SBJV-1SG  
'I would be happy if I got it.' (Saltveit 1983: 299)

Productivity however decreases further north. In Sentence (9) from Schleswig-Holstein, we find no i-mutation in *schusst* although *irrealis* meaning can be deduced from the context.

- (9) Du schusst man lewer to Huus gahn hebbfen.  
you.SG shall.PST-2SG but rather to house go have  
'You had better gone home.' (Saltveit 1983: 300)

The past tense forms in many northern dialects in Germany may show i-mutation due to their origin in subjunctive forms, but synchronically, they have taken on the role of the past indicative and thus there is no formal difference between indicative and subjunctive. These indistinct forms in the northern dialects can function as both indicative and subjunctive (Saltveit 1983: 298–301).

In addition, *irrealis* meaning can be expressed by means of auxiliary verbs such as *wilken* 'to want', *warden* 'to become', *schölen* 'shall' and *doon* 'to do', (cf. Lindow et al. 1998, who mainly describe German North Saxon (DNS)).

We are not aware of the existence of distinct subjunctive forms in today's Dutch Low Saxon.

### 2.3.5 Infinitivus pro participio (IPP)

*Infinitivus pro participio* refers to the phenomenon in West-Germanic languages such as Dutch and German of using an infinitive instead of an expected past participle, for example:

- (10) Ich hätte das tun können.  
I have.PST.SBJV-1SG that do.IMP can.IMP  
'I could have done that.'

Schmid (2005: 1) lists “Low German”<sup>6</sup> as one of the West Germanic languages where IPP-constructions do not appear, which is in line with Lindow et al. (1998).

- (11) *Korl hett den Text nich lesen kunnt.*  
 Korl has the-ACC.SG text not read can.PST-PTCP  
 ‘Korl could not read the text.’ (Lindow et al. 1998: 108)

Bloemhoff et al. (2019: 66) present a more varied picture for Dutch Low Saxon. They state that the northern dialects Gronings and Stellingwerfs indeed do not know the IPP-construction, whereas, in the other Dutch Low Saxon dialects, they assume a correlation with presence or absence of the (*g*)e-prefix in the past participle.

The situation for German Low Saxon is also in fact more complex than presented above. Even in the north-western dialects on which the grammar by Lindow et al. (1998) focuses, some speakers today do use the IPP-construction (personal observation), perhaps due to influence from Standard German.

### 2.3.6 Complementiser doubling in subordinate clauses

Similar to Frisian (Popkema 2018: 299), but different from (Standard) German and Dutch, *as* ‘as’ or *dat* ‘that’ can occur as a second complementiser in subordinate clauses, typically after question words. This is well attested in Dutch Low Saxon, cf. the following example from Gronings:

- (12) *Ik mout waitn wel dat ik in hoes krieg.*  
 I must.1SG know who that I in house get.1SG  
 ‘I need to know whom I will get into the house.’ (van Bree 2008: 114)

This phenomenon is not commonly described in grammars for German Low Saxon but Saltveit (1983: 289, 330) briefly mentions the usage of both *dat* and *as* and offers two example sentences. Moreover, we have encountered several examples with *as* in literary works such as Wisser (1921) and Peters (1986).

- (13) *Un därm̩it secht de ol Mann em Beschēd, wodenni as he dat  
 and therewith says the old man him information how as he that  
 maken schall.  
 make shall.3SG*  
 ‘And with this, the old man tells him how he should do it.’ (Wisser 1921: 29)

---

<sup>6</sup>Due to the name choice possibly only referring to varieties from Germany.

Schallert et al. (2018) have found attestations of the variant with *dat* already in Medieval Low Saxon. Its usage is however not restricted to Low Saxon, but they point out that this type of construction occurs throughout West-Germanic and even neighbouring Romance and Slavonic varieties, in particular in the Alpine region.

### 3 Dataset

Our dataset is taken from the PoS-tagged and morphologically annotated version<sup>7</sup> of the LSDC dataset LSDC-morph (Siewert et al. 2022).

The overall dataset covers eight dialect regions from the 19<sup>th</sup>, 20<sup>th</sup> and 21<sup>st</sup> century, but in this study, we focus on these six major Low Saxon dialect groups: Dutch North Saxon (NNS), German North Saxon (DNS), Dutch Westphalian (NWF), German Westphalian (DWF), Eastphalian (OFL) and Mecklenburgish-West Pom-eranian (MVP). The reason for excluding the other eastern dialects Brandenburgish (BRA), East Pomerian (POM) and Low Prussian (NPR) is the relatively low amount of data.

The older part of the dataset consists primarily of copyright-free material available online, mainly on Wikisource, Leopold & Leopold (1882)<sup>8</sup> and the *Twentse Taalbank* (van der Vliet 2021). On the other hand, the modern data for most dialects was personally provided by a variety of local authors. Common genres in the dataset are short stories and short novels, but various other genres such as speeches, religious texts, historical accounts, fairy tales and letters are included as well. For a more detailed description of the content and data collection for the LSDC dataset, please see Siewert et al. (2020).

We split the data into two time periods: 1800–1939 and 1980–today<sup>9</sup>. This split is motivated by the language shift to the respective majority language, which in most regions occurred between the 1940s and 1980s. A practical reason possibly connected to the language shift is the lack of data from the intermediate period. Only from German Westphalian (DWF) we have one text that was marked as published during this period and three additional short texts that might have been published then based on the authors' life dates. Another practical expla-

<sup>7</sup>Please see <https://universaldependencies.org/u/pos/index.html> and <https://universaldependencies.org/u/feat/index.html> for a description of the PoS tags and morphological features.

<sup>8</sup>Digitised by dbnl: [https://dbnl.nl/tekst/leop008sche00\\_01/](https://dbnl.nl/tekst/leop008sche00_01/).

<sup>9</sup>In practice, this likely means roughly 2000–today, but we do not have the exact year of publication/writing of every text that local authors provided.

nation for the lack of data might be that texts from this period are often still copyright-protected and therefore not easily available in digitised format.

We thus have an older period, when Low Saxon was still the dominant language of oral communication, and a newer period, when (a regional version of) Dutch or German has become the primary language in everyday life.

Our Standard German (DEU) and Standard Dutch (NDL) datasets are taken from Universal Dependencies.<sup>10</sup> For German, we used the UD\_German\_HDT treebank (Borges Völker et al. 2019) and, for Dutch, both the UD\_Dutch-Alpino and the UD\_Dutch-LassySmall treebanks (Bouma & van Noord 2017). We filtered the Dutch and German treebanks to remove duplicate sentences.

The overall size of our dataset is shown in Table 1 and our updated version of LSDC-morph dataset is made available at <https://github.com/Helsinki-NLP/LSDC-morph/tree/main/methods2022>.

Table 1: Size of the dataset

	1800–1939		1980–2022	
	Sentences	Tokens	Sentences	Tokens
Dutch North Saxon (NNS)	1,828	44,067	17,796	261,342
Dutch Westphalian (NWF)	4,948	102,656	9,245	136,992
German North Saxon (DNS)	23,075	429,122	3,526	61,174
Mecklenburgish-West	20,007	577,767	3,055	42,434
Pomeranian (MVP)				
German Westphalian (DWF)	17,450	337,871	16,015	273,513
Eastphalian (OFL)	1,684	33,162	8,441	145,792
Standard German (DEU)			185,380	3,489,305
Standard Dutch (NDL)			20,591	306,028

### 3.1 Annotation: Changes made to the Dutch and German UD datasets

Due to the differences in morphological feature annotation of the UD datasets, reannotation of the Standard Dutch (NDL) and Standard German (DEU) data was necessary. For this, we manually modified the annotation of 200 Dutch and German sentences each in order to train annotation models on these. For Dutch, we mainly extended morphological feature annotation, added e.g., PronType, full

<sup>10</sup><https://universaldependencies.org>

marking of person and number for verbs. There were only marginal adaptations for PoS, specifically relating to what is considered a proper noun or a number. For German, we added the differentiation for PronType, we added case marking to nouns (in the original sometimes only marked on the article), and we removed the case label from proper nouns unless they actually were marked.

### 3.2 Tagging

We have made substantial manual corrections to the automatic tokenisation of the LSDC-morph dataset. In addition to wrong or missing sentence splitting after certain punctuation marks, this mainly concerned different orthographic solutions for contractions. For instance, ‘on the-M.SG’ might be realised in the form of *up’n*, *up n* or *upm* among others, which leads to three different realisations at the PoS level: ‘ADP – PUNCT – DET’, ‘ADP – DET’ and ‘ADP’. These were unified to the ‘ADP – DET’ format.

Due to the corrections to the tokenisation, the Low Saxon data needed to be re-annotated just like the Dutch and German data. The automatic tagging was done with the Stanza tagger (Qi et al. 2020)<sup>11</sup> pretrained on large datasets without target annotation and fine-tuned separately for German, Dutch, Dutch Low Saxon (NNS and NWF), northern German Low Saxon (DNS and MVP) and southern German Low Saxon (DWF and OFL) on small sets of manually annotated data. This split of Low Saxon into three training groups was motivated by the increase in performance observed in unrelated lemmatisation and PoS tagging experiments done by us earlier (unpublished).

For training the Stanza tagger, we used the pretrained embeddings from the CoNLL 2017 shared task<sup>12</sup> for Standard Dutch and Standard German and, for Low Saxon, the fastText embeddings by Grave et al. (2018). In addition, we trained our own Low Saxon embeddings on our dataset using GloVe (Pennington et al. 2014). These led to a small improvement in accuracy for Dutch Low Saxon (91% to 92% for PoS and 80% to 83% for features) compared with previous training using fastText embeddings. So, presumably, the much larger fastText embeddings mostly or only represent German Low Saxon. As we furthermore receive a noticeably better model accuracy for northern dialects (96% for PoS and 85% for features) compared with southern German Low Saxon (91% for PoS, 83% for features), despite more southern training data, the fastText embeddings may have been mostly trained on northern German Low Saxon data.

---

<sup>11</sup>The stand-alone version we used is available at <https://github.com/yvesscherrer/stanzatagger>.

<sup>12</sup>Available at [https://stanfordnlp.github.io/stanza/word\\_vectors.html](https://stanfordnlp.github.io/stanza/word_vectors.html)

## 4 Approaches

### 4.1 Data encoding

The three levels under investigation are represented by bigrams and trigrams. We removed n-grams that occurred five times or less as well as n-grams that contained the tags ‘SYM’, ‘\_’, ‘X’ or two consecutive ‘PUNCT’ tags. The motivation for this is that the tags ‘SYM’, ‘\_’ and ‘X’ do not represent linguistic elements of interest, and that we did not want to give too much weight to personal writing habits, such as the use of doubled quotation marks by certain 19<sup>th</sup> century authors.

Character n-grams (68,695,124 overall n-grams, 30,762 distinct ones) approximate the orthographic level but can be assumed to also capture phonological and morphological features. PoS (Part of Speech) n-grams (10,811,793 overall n-grams, 2,794 distinct ones) match the syntactic level, and n-grams of PoS and morphological features (10,207,431 overall n-grams, 68,551 distinct ones) correspond to morphology and morpho-syntax. Table 2 shows character and PoS n-grams that we extract from the older Dutch Westphalian sentence from our dataset *Met un moand!* ‘With(in) a month!’.

Table 2: Extracted n-grams

	bigrams	trigrams
characters	(‘m’, ‘e’), (‘e’, ‘t’), (‘m’, ‘e’, ‘t’), (‘e’, ‘t’, ‘ ’), (‘t’, ‘ ’), (‘ ’, ‘u’), (‘u’, ‘ ’), (‘t’, ‘ ’, ‘u’), (‘ ’, ‘n’), ...	(‘m’, ‘e’, ‘t’), (‘e’, ‘t’, ‘ ’), (‘t’, ‘ ’, ‘u’), (‘ ’, ‘u’), (‘n’), ...
PoS	(ADP, DET), (DET, NOUN), (NOUN, (NOUN, PUNCT))	(ADP, DET, NOUN), (DET, NOUN, PUNCT)

The first trigram of the same sentence combining PoS and morphological features is:

```
(ADP, AdpType=Prep),
(PRON, Case=Acc,Dat|Definite=Ind|Gender=Masc|Number=Sing|PronType=
Art),
(NOUN, Case=Acc,Dat|Gender=Masc|Number=Sing).
```

In these n-grams, the first part of each element is the PoS tag and the second part consists of a concatenation of the morphological features. In dialects that

have lost certain distinctions, such as the accusative and dative distinction in case of the one above, we use combined values, here `Case=Acc,Dat`<sup>13</sup>.

N-grams have been shown to constitute a suitable unit for corpus-based dialectometry. Wolk & Szmrecsanyi (2016) compared the performance on PoS n-grams to manually selected features in their study of British dialects and discover that PoS n-grams lead to a comparable performance. In their Swiss German dialect identification experiments, Malmasi & Zampieri (2017) discover that character n-grams outperform word-based ones.

## 4.2 Dialectometry background

Dialectometry is a branch of dialectology where quantitative, and today commonly computational, methods are used in order to measure the difference between language varieties. In contrast with more traditional dialectological approaches, dialectometry typically makes use of a large aggregate of features instead of individual or a small number of selected features (Wieling & Nerbonne 2015). Nerbonne (2009) stresses the advantage of aggregate similarity as manually selected features are prone to cherry-picking.

In his PhD research, Spruit (2008) used quantitative approaches to investigate syntactic variation in West Germanic varieties spoken in the Netherlands and Flanders. He also compared the syntactic level with pronunciational and lexical differences, and was able to show that while a certain correlation can be found between the different levels, there is no full overlap. E.g., whereas Frisian clearly stands out at the lexical level, it appears fairly similar to the Low Saxon varieties at the syntactic level. (Spruit 2008: 75–78)

From the perspective of our current research, the reanalysis of the Wenker language atlas data by Lameli (2016) is particularly interesting. Based on linguistic variables, he calculates the aggregate similarity between different locations and uses hierarchical clustering to identify larger German Low Saxon dialect regions. The result are three major dialect areas: Northern Low Saxon (North Saxon and Mecklenburgish–West Pomeranian), Southern Low Saxon (Westphalian and Eastphalian) and Brandenburgish. While most of the areas found correspond roughly to the traditional dialect regions, the north-western part<sup>14</sup> of what is traditionally considered Westphalian here clusters with the northern group. This is especially important as it concerns the border region to the Netherlands. If this north-western part of German Westphalian at the Dutch border in fact shares

<sup>13</sup>The nominative value is, however, always kept separate, because all dialects have preserved this distinction at least in the personal pronouns.

<sup>14</sup>The German Westphalian data in our dataset mostly come from the southern and eastern part.

more features with the northern dialects, one might also expect the Dutch Westphalian dialects to resemble the northern group more closely.

### 4.3 PCA and hierarchical clustering

In the experiments presented below, we make use of PCA (Principal Component Analysis) with two dimensions as well as hierarchical clustering with Ward linkage and Euclidean metric. We compute the dialect distances with the help of the scikit-learn Python library (Pedregosa et al. 2011). The input to the PCA and the hierarchical clustering is a matrix of n-gram counts per variety, such as older (1800–1939) Dutch Westphalian (NWF) or contemporary (1980–2022) Eastphalian (OFL). Before clustering, the n-gram counts are tf-idf-normalised, since raw counts are not comparable due to the differences in amount of data per variety. In our context, the abbreviation can be understood as “term frequency – inverse *dialect* frequency”.

We tried out different splits of the dataset to test the dialect-internal consistency and found that the different parts of the same variety indeed form clusters. Neither did the inclusion of Mecklenburgish–West Pomeranian (MVP), the addition of new data in German North Saxon (DNS) nor the reduction of the German (DEU) data to 20,000 sentences cause a noticeable effect on the overall tendencies. Furthermore, we reran the clusterings several times to test their stability and found no major differences between the individual runs. Only in the k-means clustering<sup>15</sup> performed on the PCA-reduced data, we found slight variation in the position of cluster borders next to or between very close varieties.

## 5 Results

We will first present the results of character-based experiments, followed by the PoS-based ones and, finally, the results based on both PoS-tags and morphological features. The figures show the PCA-based results on the left and hierarchical clustering results on the right. The red arrows in the PCA figure indicate the development of a particular Low Saxon dialect from the older period, 19<sup>th</sup> to early 20<sup>th</sup> century, to the more modern period, late 20<sup>th</sup> to 21<sup>st</sup> century.

In addition to the figures, we also discuss n-gram counts relating to particular phenomena. Due to their size, we cannot present the tables here but, the files can be found on GitHub: <https://github.com/Helsinki-NLP/LSDC-morph/tree/main/methods2022>.

---

<sup>15</sup>Not further discussed in this paper.

## 5.1 Character level

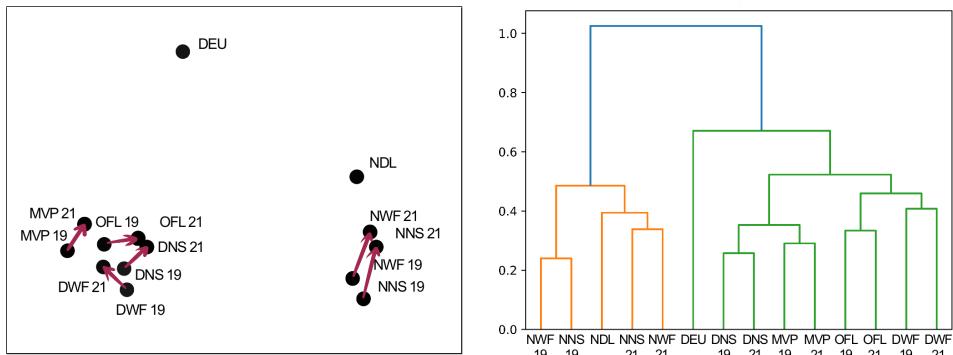


Figure 2: Character level PCA and hierarchical clustering. (NNS: Dutch North Saxon, NWF: Dutch Westphalian, DNS: German North Saxon, DWF: German Westphalian, MVP: Mecklenburgish–West Pomeranian, OFL: Eastphalian)

The character-based PCA results on the left in Figure 2 show a division into three groups: German Low Saxon (DNS, DWF, MVP, OFL), German (DEU), and Dutch (NDL) with Dutch Low Saxon (NNS, NWF). While the Dutch Low Saxon dialects seem to approach Dutch, German Low Saxon does not show a clear trend of converging towards German.

On the right, we see a division according to state, likely reflecting different orthographic traditions. Interestingly, the variants from the Netherlands cluster according to century, with older Dutch Low Saxon in one group and late 20<sup>th</sup> to 21<sup>st</sup> century Dutch and Dutch Low Saxon in the other. The variants from the German side however, branch according to what we would expect based on the analysis by Lameli (2016), with a northern (DNS, MVP) and a southern (DWF, OFL) Low Saxon branch that subsequently divide into the major dialect groups. In Dutch Low Saxon, we thus find diachronic differences more strongly pronounced, which can also be seen in the PCA. In German Low Saxon, in contrast, dialectal differences appear more important than diachronic ones.

## 5.2 PoS level

In Figure 3, visualising the PoS results, we see Dutch (NDL) and German (DEU) forming a common group and the Low Saxon dialects forming another one. Particularly the PCA results show the Low Saxon dialects approaching the modern

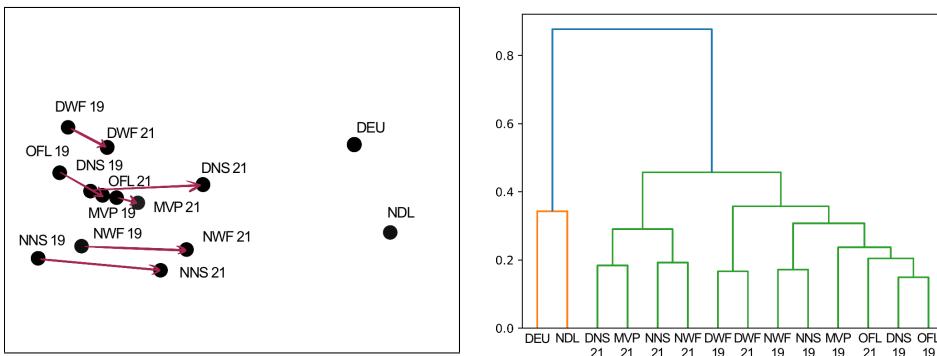


Figure 3: PoS level PCA and hierarchical clustering.

NNS: Dutch North Saxon, NWF: Dutch Westphalian, DNS: German North Saxon, DWF: German Westphalian, MVP: Mecklenburgish-West Pomeranian, OFL: Eastphalian

state languages Dutch and German. This does, however, not increase the distance along the border in all cases. For example, contemporary German North Saxon (DNS 21) is not placed closer to German Westphalian (DWF) than to the contemporary dialects from the Dutch side (NNS 21, NWF 21).

In the hierarchical clustering, we find a somewhat corresponding constellation where the contemporary northern branch from Germany (DNS 21, MVP 21) forms a cluster with contemporary Dutch Low Saxon (NNS 21, NWF 21), while the remaining Low Saxon dialects form another cluster. Within this second cluster, we find German Westphalian (DWF) branching off first while older Dutch Low Saxon (NNS 19, NWF 19) is placed closer to Eastphalian (OFL) and the older variants of the northern branch of German Low Saxon (DNS 19, MVP 19).

In the use or lack of the article described in Section 2.3.1, we indeed see similar trends to what Pheiff (2022) described. For the n-gram ('ADP', 'DET', 'NOUN'), we find the lowest score in older Dutch North Saxon (NNS 19, NWF 19). However, comparing the older part of the dataset to the contemporary one, we can observe the score to increase for all dialects except Mecklenburgish-West Pomeranian (MVP). Similarly, for the n-gram ('ADP', 'DET', 'ADJ'), all Low Saxon variants – both older and contemporary – receive a score lower than Dutch (NDL) and German (DEU), and, again apart from Mecklenburgish-West Pomeranian (MVP), show an increase towards the modern period. Also, the scores for the n-gram ('DET', 'PROPN') show lower values for all Low Saxon varieties compared with Dutch (NDL) and German (DEU). Again, an increase can be attested in all varieties except one. A slight decrease can be seen in German North Saxon (DNS).

Dialect differences can also be found for the infinitive construction with *üm*, cf. Section 2.3.2, represented by the n-gram ('SCONJ', 'PART', 'VERB').<sup>16</sup> We do not find this n-gram in the German North Saxon (DNS) or Mecklenburgish-West Pomeranian (MVP) part of the dataset. Neither do we find it in the older Dutch North Saxon (NNS 19, NWF 19) or older Eastphalian (OFL 19), but this could be due to the small amount of data in these two varieties (cf. Table 1) and the relative rarity of the construction, as it does appear in the larger contemporary dataset. German Westphalian (DWF) receives comparable scores to German (DEU) (0.00015) in both parts of the dataset (0.00011 and 0.00014). Dutch Westphalian (NWF), on the other hand, shows a decrease from 0.00093 to 0.00015, a score similar to German (DEU) and German Westphalian (DWF).

At the PoS level, we cannot observe a clear tendency of Low Saxon growing apart at the political border. While the increase in article usage brings the dialects closer to Standard Dutch (NDL) and German (DEU), this same development occurs on both sides of the border. Instead, we find a particular closeness of Dutch Low Saxon (NNS, NWF) to the northern dialects (DNS, MVP) in Germany and to a somewhat lesser degree to Eastphalian (OFL).

### 5.3 PoS and morphological features

In Figure 4, the distance to Dutch (NDL) is disproportionately increased by the gender feature. Since Dutch does not distinguish between masculine and feminine gender in nouns anymore, these receive the tag *Gender=Fem, Masc*, while Low Saxon and German nouns mostly receive the distinct gender features *Gender=Fem* or *Gender=Masc*. Compare Figure 5, where the *masculine* and *feminine* value of this feature have been replaced by *com* in the whole dataset. This does, however, not seem to obscure the general trends of development within Low Saxon.

An unexpected finding in the PCA is that not only Dutch Low Saxon (NNS, NWF), but also the northern branch of German Low Saxon (DNS, MVP) converges towards Dutch (NDL). Eastphalian (OFL) on the other hand, appears to approach the northern dialects. In the hierarchical clustering, similar to Figure 3, Dutch Low Saxon forms a cluster with Eastphalian and northern German Low Saxon.

We find German Westphalian (DWF) in an outlier position both in the PCA and in the hierarchical clustering. In Figure 5, it can even be seen to cluster with German (DEU) instead of the other Low Saxon dialects. In addition to the

---

<sup>16</sup>Although this does not cover cases with an object or adverb between the *üm* and the *to*, it can still give us a rough idea of the usage.

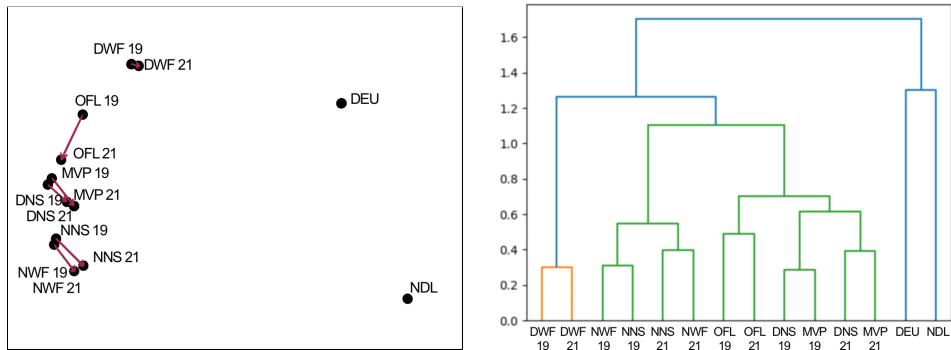


Figure 4: PoS and morphological features, PCA and hierarchical clustering.

NNS: Dutch North Saxon, NWF: Dutch Westphalian, DNS: German North Saxon, DWF: German Westphalian, MVP: Mecklenburgish-West Pomeranian, OFL: Eastphalian

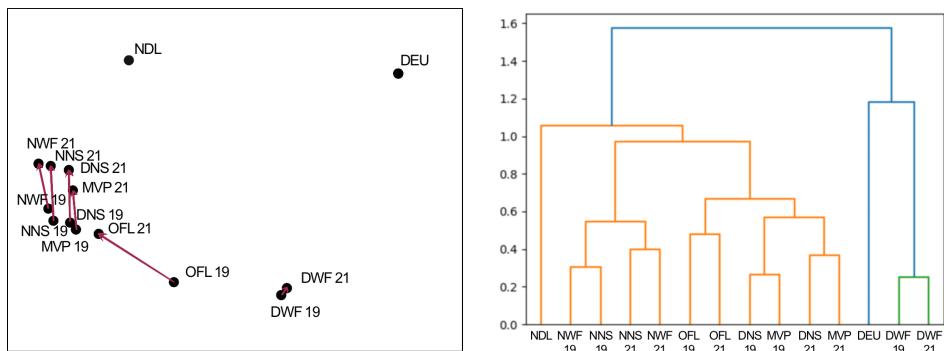


Figure 5: PoS and morphological features, PCA and hierarchical clustering, without feminine–masculine distinction

NNS: Dutch North Saxon, NWF: Dutch Westphalian, DNS: German North Saxon, DWF: German Westphalian, MVP: Mecklenburgish-West Pomeranian, OFL: Eastphalian

greater closeness to German, the amount of change in German Westphalian is also strikingly small. Preservation of the dative case, cf. Section 2.3.3, is certainly one feature that contributes to the relative closeness to German. Among the n-grams with the highest scores for both German and German Westphalian, there are several ones containing the *Case=Dat* feature.

The subjunctive mood described in Section 2.3.4 was unfortunately not learnt properly by the tagger due to its apparent rarity in the training data. As subjunctive forms were tagged as indicative or ambiguous, we cannot draw any conclusions based on our dataset regarding this aspect.

The IPP construction<sup>17</sup> or lack thereof<sup>18</sup> described in Section 2.3.5 is not among the highest scoring features, as only the bigram

('AUX', 'VerbForm=Inf'), ('VERB', 'VerbForm=Inf')

makes it into the top 500 features for older Dutch Westphalian (NWF) and Dutch (NDL), whereas in most cases, the relevant n-grams are not even among the top 1,000 features. Nonetheless, we do observe dialectal differences, as well as language change here. These, however, only partly reflect the description in the literature. The main finding is that the IPP construction, either with *VERB – AUX* or with *AUX – VERB* order, occurs in all Low Saxon dialects and is the preferred construction in most of them. Curiously, it is by far the preferred construction even in both periods of Dutch North Saxon (NNS), while Bloemhoff et al. (2019) claim to not know of the phenomenon. The only varieties showing a clear preference for the non-IPP version, are German North Saxon (DNS) and Mecklenburgish–West Pomeranian (MVP) from the older period. In the contemporary data, despite the non-IPP construction still occurring frequently, its IPP alternative has started to dominate. This confirms our personal observation that the IPP variant is indeed used as well by Low Saxon speakers today.

Despite such individual opposing trends, however, none of the German Low Saxon dialects seem to as a whole increasingly resemble German (DEU), and the distance between Dutch Low Saxon (NNS, NWF) and the northern dialects from Germany (DNS, MVP) remains roughly the same.

## 6 Discussion

As expected, the three levels under investigation lead to rather different results, in terms of both clustering and observable changes in similarity.

<sup>17</sup>Represented by the bigrams ('AUX', 'VerbForm=Inf'), ('VERB', 'VerbForm=Inf') and ('VERB', 'VerbForm=Inf'), ('AUX', 'VerbForm=Inf').

<sup>18</sup>Represented by ('VERB', 'VerbForm=Inf'), ('AUX', 'Tense=Past|VerbForm=Part').

We assume the character level to largely reflect spelling differences (cf. Section 2.1) and these are likely behind the noticeable divide between Dutch Low Saxon and German Low Saxon which we do not see in the other experiments. Nevertheless, among the n-grams with the highest tf-idf values, we also find a few that indicate phonological and morphological differences. For instance, we find that the Low Saxon dialects with preserved word-final *e* (DWF, NWF, OFL) receive a higher score for the n-gram ('e', ' ') than the dialects with *e*-apocope in nouns and verbs (DNS, NNS, MVP). Similarly, the highest score for the n-gram ('t', ' ') is found in contemporary German North Saxon (DNS), which uses -(*e*)*t* as the plural suffix of present tense verbs, while older Mecklenburgish–West Pomeranian (MVP), which typically uses -(*e*)*n* for the same function, receives the lowest score.

The similarity at the character level is likely distorted by the different orthographic traditions in other ways as well. It could for example be the case that the spellings commonly used for Dutch Low Saxon today make it appear disproportionately Dutch compared with the older variants. This would explain the noticeable shift towards Dutch in the PCA as well as the grouping of contemporary Dutch Low Saxon with Dutch in the hierarchical clustering. Dutch Low Saxon clustering according to time period instead of dialect should however not primarily be attributed to orthography, as we see the same phenomenon at the level of PoS and morphological features as well.

One might have expected a comparable closeness between German Low Saxon and German due to the German-based writing traditions. In addition to morphological differences, an explanation might be found in the High-German consonant shift leading to a different distribution and frequency of many phonemes. The overall clustering within German Low Saxon corresponds well to the findings of Lameli (2016), and we see a northern group consisting of German North Saxon (DNS) and Mecklenburgish–West Pomeranian (MVP), and a southern group consisting of Westphalian (DWF) and Eastphalian (OFL).

Despite the common naming, Dutch Westphalian (NWF) could not be shown to be particularly close to German Westphalian (DWF) at any of the levels investigated. Furthermore, the similarity between Dutch North Saxon (NNS) and Dutch Westphalian (NWF) appears to be roughly equal to the similarity between German North Saxon (DNS) and Mecklenburgish–West Pomeranian (MVP). This supports the idea that if, in line with Lameli (2016), the north-western part of German Westphalian shares more features with the northern dialects than with the other Westphalian varieties, the neighbouring varieties on the Dutch side of the border might also share more traits with northern Low Saxon.

The two lines we mainly wanted to investigate – the Dutch-German border and the traditional East-West division in German Low Saxon – did not present themselves as clearly as we had expected.

An East-West division could not be found at any of the levels under scrutiny. On the contrary, German North Saxon (DNS) and Mecklenburgish-West Pomeranian (MVP) actually appear so close that they cluster according to century instead of dialect at both the PoS and the morphological level. It would be desirable to repeat these experiments with additional eastern dialects when more data becomes available.

A certain distance between Dutch Low Saxon (NNS, NWF) and the northern dialects from Germany (DNS, MVP) is observable and might be attributed to influence from the different state languages. The distance is however not bigger than between the northern German Low Saxon dialects and German Westphalian (DWF). In particular, the border does not explain why the northern dialects from Germany would converge towards Dutch (NDL) instead of German (DEU). Thus, the situation appears to be more complex.

On the one hand, the loss of morphological complexity, especially the loss of case inflection, could be one factor behind the increasing similarity between Dutch (NDL) and northern German Low Saxon (DNS, MVP). Another factor that should be taken into consideration is the ongoing strife to codify Low Saxon, in particular in Germany, where textbooks and grammar descriptions for school and adult education have been published in several dialects.

Due to this available documentation, the people who produce written German Low Saxon are probably very aware of the differences between Low Saxon and German. Moreover, they presumably strive to produce what they consider “good Low Saxon”, which is not necessarily the same form of language they would informally speak at home with family and friends but might be an idealised form of their dialect where features distinct from German are preferred.

Furthermore, thanks to the internet and social media in particular, it has become easier to access content in other dialects and from the other side of the border. We are aware that several people who provided data have interdialectal and some even cross-border contact.

As a result from the larger number of speakers, northern German Low Saxon (DNS, MVP) is better represented in media and literature, which could explain why Eastphalian (OFL) appears to be converging towards the northern dialects.

Language skills might play a role as well, as the respective state language is not necessarily the only additional language of which Low Saxon writers have knowledge. Some of the people from whom we received texts are to different degrees proficient in the other state language or in Scandinavian languages as well.

Especially in case of younger speakers, knowledge of English can be assumed and English influence on the Low Saxon of younger second language or heritage learners might be a topic worth investigating.

## 7 Limitations and future research

We have so far based our comparisons of Low Saxon dialect similarity only on PCA and hierarchical clustering. We can therefore not preclude that other clustering approaches or a different implementation of the same clustering methods might yield different results. Furthermore, we would like to test a greater variety of visualisation techniques. For instance, a map format might be easier to grasp, especially for readers less familiar with the Low Saxon language area.

Another factor of uncertainty is the automatic tagging that our work relies on. Upon manual inspection of the n-grams, we have become aware of issues with the tagging of certain features. One of these is the gender feature. Even in the dialects that have preserved a three-fold distinction, the form of a possible accompanying article or adjective does not necessarily fully disambiguate the gender of the noun. This poses a challenge for both the automatic tagger and for the human annotator as the gender of nouns can differ from dialect to dialect and might not be documented for the precise dialect in question. In particular, the Dutch Low Saxon dialects in the process of losing their feminine-masculine distinction pose problems as it is not always obvious when a gender distinction can still be assumed.

The unfortunate fact that our tagger did not learn to recognise the subjunctive has shown that rare features can require careful manual selection of the training data. In the German and Low Saxon sentences randomly selected for manual annotation and subsequent fine-tuning, this feature apparently did not occur frequently enough.

Another tagging-related risk concerns the finetuning to different groups of dialects. Since the finetuning sets consist of only around 200 to 300 sentences per group, some overfitting to a particular subset of the data might have happened. Nevertheless, we considered the finetuning justified due to the substantial spelling variation described in Section 2.1. There are, for instance, several cases of character strings that would receive different PoS tags in different dialects: For instance, *doe* as the personal pronoun of the second person singular in Gronings (part of NNS) receives the tag ‘PRON’; as the definite article in East Westphalian (part of DWF) it is tagged as ‘DET’ and, in addition, it is a possible spelling for the 1<sup>st</sup> person singular present tense of the verb *doon* ‘to do’ in, e.g., German North

Saxon (DNS) and Dutch Westphalian (NWF), when it should be tagged as ‘VERB’ or possibly ‘AUX’. Neither can we rule out that this finetuning might cause the varieties within the same group to appear closer than they are. While the members of the same group do at least not appear artificially close, compare, for instance, the German South Low Saxon group consisting of Eastphalian (OFL) and German Westphalian (DWF) in Figures 4 and 5 or the two periods of Dutch Low Saxon: Dutch North Saxon (NNS) and Dutch Westphalian (NWF) in Figure 3, larger and more diverse finetuning sets would certainly be desirable.

The different size of the dialect regions should not be neglected either. The German Low Saxon regions are noticeably larger than the Dutch Low Saxon ones, and most of the German Low Saxon texts in our dataset are not from areas particularly close to the border. Several of the German North Saxon (DNS) writers come from Schleswig-Holstein or Hamburg, while in the German Westphalian dataset (DWF), East Westphalian and South Westphalian are overrepresented. Varieties from border regions, such as East Frisian and West Munsterlandic, might have exhibited greater similarity with Dutch Low Saxon.

The lack of diversity among the writers is a problem in some dialects as well, in particular in contemporary German Westphalian (DWF) and Eastphalian (OFL), and in older Mecklenburgish–West Pomeranian (MVP). Contemporary Eastphalian is unfortunately only represented by one writer so far, which is why some of the developments we observe might simply be features of this writer’s idiolect. While in contemporary German Westphalian, we have obtained texts from a variety of writers, these mostly represent the older generation born in the first half of the 20<sup>th</sup> century. Even though the texts themselves were published in the late 20<sup>th</sup> or the 21<sup>st</sup> century, their language might be more representative of the older period. This would explain why we see so little change in German Westphalian both at the PoS and the morphological level. The older Mecklenburgish–West Pomeranian part of the dataset also includes texts from a variety of sources but works by Fritz Reuter clearly dominate. The fact that the older part of the Mecklenburgish–West Pomeranian data scores higher in article use than the newer part, could thus possibly result from characteristics of Reuter’s idiolect. The collection of additional data for these dialects would therefore be desirable.

An additional desideratum is the comparison with Dutch and German from the 19<sup>th</sup> – early 20<sup>th</sup> century. While the convergence towards the state languages that we see at the PoS level seems intuitive, we cannot yet rule out the possibility that all three languages develop into a similar direction.

Some structures such as the double complementiser with *as/dat* described in Section 2.3.6 that we initially planned to include, turned out to require lemma information since the PoS bigram *ADV – SCONJ* representing, e.g., *worüm dat*

‘why’, captures several unrelated constructions as well. We are already working on automatic lemmatisation for Low Saxon and are planning to study lexical differences as well in our future work.

## Abbreviations

MMM	German	NDL	Dutch
DEU	German	NNS	Dutch North Saxon
DNS	German North Saxon	NWF	Dutch Westphalian
DWF	German Westphalian	OFL	Eastphalian
MVP	Mecklenburgish-West Pomeranian	PoS	Part of Speech

## Acknowledgements

We would like to thank Kevin Behrens, Behrend Böckmann, Johanna Bojarra, Marita Bojarra, Christiane Ehlers, Marianne Ehlers, Heiko Gauert, Jan Graf, Bernd Lubs, Christian Peplow, Karl Peplow, Gennadi Ratson, Heinrich Siefert and Florian Wille for providing additional texts in the northern dialects from Germany!

This work has been supported by the Academy of Finland through project No. 342859 “CorCoDial – Corpus-based computational dialectology”.

## References

- Arbatzat, Hartmut. 2016. *Platt: Dat Lehrbook*. Hamburg: Quickborn-Verlag.
- Bloemhoff, Henk. 2008. Stellingwerfs. In Henk Bloemhoff, Jurjen van der Kooi, Hermann Niebaum & Siemon Reker (eds.), *Handboek Nedersaksische taal- en letterkunde*, 175–193. Assen: Koninklijke Van Gorcum.
- Bloemhoff, Henk, Philomène Bloemhoff-de Bruijn, Jan Nijen Twilhaar, Henk Nijskink & Harrie Scholtmeijer. 2019. *Nedersaksisch in een notendop: Inleiding in de Nedersaksische taal en literatuur*. Assen: Koninklijke Van Gorcum.
- Borges Völker, Emanuel, Maximilian Wendt, Felix Hennig & Arne Köhn. 2019. HDT-UD: A very large Universal Dependencies Treebank for German. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, 46–57. Paris, France: Association for Computational Linguistics. DOI: 10.18653/v1/W19-8006. <https://aclanthology.org/W19-8006>.

- Bouma, Gosse & Gertjan van Noord. 2017. Increasing return on annotation investment: The automatic construction of a Universal Dependency treebank for Dutch. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, 19–26. Gothenburg, Sweden: Association for Computational Linguistics. <https://aclanthology.org/W17-0403>.
- Goossens, Jan. 2019. „Dialektverfall“ und „Mundartrenaissance“ in West-niederdeutschland und im Osten der Niederlande. In Gerhard Stickel (ed.), *Varietäten des Deutschen: Regional- und Umgangssprachen*, 399–404. Berlin: de Gruyter. DOI: 10.1515/9783110622560-023.
- Grave, Edouard, Piotr Bojanowski, Prakhar Gupta, Armand Joulin & Tomas Mikolov. 2018. Learning word vectors for 157 languages. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis & Takenobu Tokunaga (eds.), *Proceedings of the eleventh international conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). <https://aclanthology.org/L18-1550/>.
- Herrmann-Winter, Renate. 2006. *Hör- und Lernbuch für das Plattdeutsche*. Rostock: Hinstorff.
- Hiestermann, Heike & Katrin Konen-Witzel. 2021. *Snacken, Proten, Kören: Plattdüütsch-Lehrbook für de SEK I*. Hamburg: Quickborn-Verlag.
- Jellinghaus, Hermann. 1892. *Die niederländischen Volksmundarten nach den Aufzeichnungen der Niederländer*. Norden, Leipzig: Diedr. Soltau's Verlag.
- Kahl, Heinrich & Heinrich Thies. 2009. *Der neue SASS: Plattdeutsches Wörterbuch*. Neumünster: Wachholtz.
- Lameli, Alfred. 2016. Raumstrukturen im Niederdeutschen: Eine Re-Analyse der Wenkerdaten. *Niederdeutsches Jahrbuch: Jahrbuch des Vereins für niederdeutsche Sprachforschung* 139. 131–152.
- Lasch, Agathe. 1974. *Mittelniederdeutsche Grammatik* (Sammlung kurzer Grammatiken germanischer Dialekte 9). Halle (Saale): Max Niemeyer Verlag.
- Leopold, Johan A. & Lubbertus Leopold. 1882. *Van de Schelde tot de Weichsel*. Groningen: J.B. Wolters.
- Lindow, Wolfgang, Dieter Möhn, Hermann Niebaum, Dieter Stellmacher, Hans Taubken & Jan Wirrer. 1998. *Niederdeutsche Grammatik*. Leer: Schuster.
- Lücht, Wilko. 2016. *Ostfriesische Grammatik*. Aurich: Ostfriesische Landschaftliche Verlags- und Vertriebsgesellschaft mbH.
- Malmasi, Shervin & Marcos Zampieri. 2017. German dialect identification in interview transcriptions. In Preslav Nakov, Marcos Zampieri, Nikola Ljubešić, Jörg Tiedemann, Shevin Malmasi & Ahmed Ali (eds.), *Proceedings of the Fourth*

- Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, 164–169. Valencia: Association for Computational Linguistics. DOI: 10.18653/v1/W17-1220.
- Nerbonne, John. 2009. Data-driven dialectology. *Language and Linguistics Compass* 3(1). 175–198.
- Niebaum, Hermann. 2008. Het Nederduits. In Henk Bloemhoff, Jurjen van der Kooi, Hermann Niebaum & Siemon Reker (eds.), *Handboek Nedersaksische Taal- en Letterkunde*, 430–447. Assen: Koninklijke Van Gorcum.
- Pedregosa, Fabian, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot & Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12. 2825–2830.
- Pennington, Jeffrey, Richard Socher & Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. <http://www.aclweb.org/anthology/D14-1162>.
- Peters, Friedrich Ernst. 1986. *Baasdörper Krönk*. Wolfgang Lindow & Paul Selk (eds.). Husum: Husum Druck- und Verlagsgesellschaft.
- Pheiff, Jeffrey. 2022. Grammatikalisierung im Raum? Zu Variation und Wandel des Definitartikels in den niedersächsischen Dialekten Groningens und Drents. *Niederdeutsches Jahrbuch: Jahrbuch des Vereins für niederdeutsche Sprachforschung* 145. 130–155.
- Popkema, Jan. 2018. *Grammatica Fries*. Leeuwarden: Afûk.
- Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton & Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. <https://nlp.stanford.edu/pubs/qi2020stanza.pdf>.
- Saltveit, Laurits. 1983. Syntax. In Gerhard Cordes & Dieter Möhn (eds.), *Handbuch zur niederdeutschen Sprach- und Literaturwissenschaft*, 279–333. Berlin: Erich Schmidt Verlag.
- Schallert, Oliver, Alexander Dröge & Jeffrey Pheiff. 2018. Doubly-filled COMPs in Dutch and German: A bottom-up approach. <https://lingbuzz.net/lingbuzz/003979>.
- Schmid, Tanja. 2005. *Infinitival syntax: Infinitivus pro participio as a repair strategy* (Linguistik aktuell/Linguistics today 79). Amsterdam: John Benjamins.

- Schröder, Ingrid. 2004. Niederdeutsch in der Gegenwart: Sprachgebiet – Grammatisches – Binnendifferenzierung. In Dieter Stellmacher (ed.), *Niederdeutsche Sprache und Literatur der Gegenwart*, 35–97. Hildesheim, Zürich, New York: Georg Olms Verlag.
- Siewert, Janine, Yves Scherrer & Martijn Wieling. 2022. Low Saxon dialect distances at the orthographic and syntactic level. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, 119–124. Dublin, Ireland: Association for Computational Linguistics. DOI: 10.18653/v1/2022.lchange-1.12.
- Siewert, Janine, Yves Scherrer, Martijn Wieling & Jörg Tiedemann. 2020. LSDC: A comprehensive dataset for Low Saxon dialect classification. In Marcos Zampieri, Preslav Nakov, Nikola Ljubešić, Jörg Tiedemann & Yves Scherrer (eds.), *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, 25–35. Barcelona (Online): International Committee on Computational Linguistics (ICCL). <https://www.aclweb.org/anthology/2020.vardial-1.3>.
- Smits, Tom. 2011. Dialectverlies en dialectnivellering in Nederlands-Duitse grens-dialecten. *Taal en Tongval* 63(1). 175–196.
- Spruit, Marco René. 2008. *Quantitative perspectives on syntactic variation in Dutch dialects*. Utrecht: LOT.
- Stellmacher, Dieter. 1983. Neuniederdeutsche Grammatik: Phonologie und Morphologie. In Gerhard Cordes & Dieter Möhn (eds.), *Handbuch zur niederdeutschen Sprach- und Literaturwissenschaft*, 238–278. Berlin: Erich Schmidt Verlag.
- Thies, Heinrich. 2010. *SASS Plattdeutsche Grammatik*. Neumünster: Wachholtz.
- van Bree, Cor. 2008. Syntaxis. In Henk Bloemhoff, Jurjen van der Kooi, Hermann Niebaum & Siemon Reker (eds.), *Handboek Nedersaksische Taal- en Letterkunde*, 113–133. Assen: Koninklijke Van Gorcum.
- van der Vliet, Goaitsen. 2021. *Twentse Taalbank*. <http://www.twentsetaalbank.nl/>. Accessed: (15 December, 2021).
- Wieling, Martijn & John Nerbonne. 2015. Advances in dialectometry. *Annual Review of Linguistics* 1(1). 243–264.
- Wisser, Wilhelm. 1921. *Wat Grotmoder vertelt: Ostholsteinische Volksmärchen*. Jena: Eugen Diederichs.
- Wolk, Christoph & Benedikt Szmrecsanyi. 2016. Top-down and bottom-up advances in corpus-based dialectometry. In Marie-Hélène Côté, Remco Knooijhuizen & John Nerbonne (eds.), *The future of dialects: Selected papers from Methods in Dialectology XV*, 225–244. Berlin: Language Science Press.

# Chapter 9

## Leaner, cleaner, and full of attitude

Allison Burkette<sup>a</sup> & Lamont Antieau<sup>a</sup>

<sup>a</sup>University of Kentucky

Early fieldwork for the Linguistic Atlas of the United States and Canada (later, the Linguistic Atlas Project, or LAP) consisted of 6- to 8-hour-long interviews designed to elicit lexical, phonological, and grammatical targets from native informants in predominantly English-speaking communities throughout North America. Field-workers for the project were trained to ask questions that were usually in the form of descriptive phrases that tasked the speaker to name the item being described (these were often framed by the phrase *what would you call [description]*) or fill-in-the-blank questions that asked the speaker to supply the target as the missing word (*if a glass fell on a hard floor and shattered, you would say the glass \_\_\_\_\_*). For the earliest surveys, responses were written down in the International Phonetic Alphabet (IPA), but later interviews were tape recorded in their entirety and the responses transcribed from the recordings. The ultimate goal of the earliest surveys was the mapping of dialect boundaries, particularly the mapping of individual linguistic features.

Over time, however, the goals and structure of the LAP interview have changed. The overarching goal of LAP interviews has shifted from an interest in isoglosses and dialect boundaries to a desire to record variation and to investigate the correlations between that variation and specific social and regional groups. Today's LAP interviews take the form of conversational interviews that still seek names for specific foods, animals, weather phenomena, etc., as well as morphological and syntactic forms, but also address the contemporary sociolinguistic interest in perceptions and attitudes. This paper provides details of the new hybrid LAP interview, one whose format offers the best of all worlds: a set of dialectological targets that will facilitate comparisons across LAP surveys and through time, the free-flowing conversation prized by traditional sociolinguistics for grammatical and phonological analysis, and questions aimed at getting at interviewees' attitudes and beliefs about language use in their communities. Interviews have already been conducted in Kentucky with this new framework.



## 1 Opening remarks

In this paper, we describe current efforts at the University of Kentucky to adapt traditional Linguistic Atlas Project (LAP) methods to reflect a new generation of speakers and technology and to attract a new generation of students and future researchers. As an important part of this investigation, we aim to produce results that are comparable with those collected by earlier LAP projects, so one of our tasks has been to place all results on a similar digital playing field. This has entailed organizing and cataloging many kinds of older LAP materials (e.g., paper, recordings of various media types, transcriptions, etc.) and digitizing that which had not previously been digitized at the University of Georgia. This massive endeavor was fueled by the desire to make LAP materials more accessible to researchers in dialectology and sociolinguistics who are interested in studying language variation and change.

Just as importantly, if not more so, we wish to encourage students to become interested in and engaged in all aspects of the collection and preservation of LAP data, including fieldwork, thus continuing the tradition of using the LAP as a laboratory for new and future linguists. The motivation for doing so is grounded in the belief that the best way to learn how to do something is through hands-on, practical experience (and the belief that to know the Atlas is to love the Atlas).

## 2 Background

### 2.1 The Linguistic Atlas Project

The LAP began in 1929 when Hans Kurath was asked by the American Dialect Society to undertake a large-scale survey of American English. Kurath began training fieldworkers in 1930 and with them began the systematic investigation of variation in regional American English. Fieldwork began with the Linguistic Atlas of New England (LANE) in 1931, which was then followed by the Linguistic Atlas of the Middle and South Atlantic States (LAMSAS), and the Linguistic Atlas of the North Central States (LANCS). These early LAP surveys employed face-to-face interviews to elicit speakers' responses to a lengthy questionnaire that contained over 800 prompts aimed at capturing variation in regional vocabulary, pronunciation, and grammar. The bulk of these early interviews took the form of questions aimed at eliciting specific responses, such as *what do you call [description]* or fill-in-the-blank questions that asked the speaker to supply the target as the missing word (*if a glass fell on a hard floor and shattered, you would say the glass \_\_\_\_*). We know now, however, that many fieldworkers also integrated more conversational prompts into the interviews, such as *tell me about*

\_\_\_\_\_, which prompted longer, sometimes narrative, answers. Although a smattering of brief recordings exists from these early surveys, the main mode of data collection took the form of on-the-spot, handwritten International Phonetic Alphabet transcription of words or phrases. These transcriptions, collectively referred to as field pages, have been the main source of data for most of the LAP studies carried out to date.

Data collection for the LAP has never been a linear process; the bulk of the LAMSAS interviews took place in the 1930s and '40s and the bulk of LANCS interviews in the 1950s and '60s, but there were other surveys being conducted concurrently, such as the interviews with Gullah speakers conducted by Lorenzo Dow Turner in 1933 and the Linguistic Atlas of Hawai'i (LAH), directed by C.M. Wise during his year as a visiting professor at the University of Hawaii in 1950.

Historically, the aim of the LAP interview was, as Kurath (1972: 13) put it, to collect "the speechways of the folk", with the ultimate goal of the earliest surveys being the delineation of dialect boundaries, an exercise interwoven with the creation of maps of individual linguistic features. Over time, however, the goals and structure of the interview have changed. The overarching goal of LAP interviews has shifted from an interest in isoglosses and dialect boundaries to a desire to record variation and to look at the correlations between that variation and specific social and regional groups. Starting in the late 1960s with the Linguistic Atlas of the Gulf States (LAGS), LAP fieldworkers under the direction of Lee Pederson implemented an interview structure that encouraged a more conversational exchange.

With LAGS in the process of publication (Pederson et al. 1986-1993), Pederson set his sights on bringing together existing LAP databases from across the continent while also surveying large parts of the western United States that had not been previously covered by earlier fieldwork (Pederson 1996a). As part of Pederson's plan for the project (Pederson 1990), fieldworkers would tape-record three-hour interviews with natives of communities in the West in their entirety. These recordings would then be transcribed in more or less standard orthography, thus allowing for the interviews to be processed using digital tools to compile wordlists, to get wordcounts, to investigate keywords in context, and the like. To this end, he sought to make interviews much shorter (down to 360 targets over the course of three hours) while still ensuring they would be largely comparable to earlier interviews in terms of lexical and phonetic targets. Prompts for the grammatical features that had been targeted by earlier worksheets were eliminated, and instead the worksheets placed greater emphasis on conversation from which grammatical variants might emerge more naturally. Additionally, some of the tasks/targets that might make informants self-conscious, such as

counting and providing the names of days and months, as well as names for body parts, were moved to late in the interviews. Finally, there was a focus in the interviews on the language and culture of the Western states.

Fieldwork for the Linguistic Atlas of the Western States (LAWS) began with interviews of 15 native informants in Wyoming in 1988 (Pederson & Madsen 1989), before fieldwork commenced in Colorado and Utah for a Linguistic Atlas of the Middle Rockies (LAMR: Pederson 1996a, Antieau 2012, Antieau & Darwin 2013), as well as El Paso, Texas (Hamilton-Brehm 2003), and southern California. Pederson then went on to use the same methods in his fieldwork in the Caribbean islands of St. Kitts and Nevis (Baker & Pederson 2013).

While many basic assumptions and methods of the LAP have remained constant through the years, in part to ensure that more recently collected data is comparable to the earliest LAP data, LAP methodology has evolved in several ways. For instance, while the earliest LAP components often sought nonmobile, older, rural males (NORMs) for interviews, there has been some effort to make the informant pool more inclusive with regard to the sex of the informant, their education level, profession, etc. NORMs also tended to be white, and LAGS, in particular, made an effort to have a more balanced representation with respect to race. While the LANE and LAMSAS informant pools evidence a gross underrepresentation of women and people of color, LAGS informants were chosen in a more demographically representative manner. For example, LANE had two African American informants (out of 416 informants total) and LAMSAS had 41 (out of 1162), but LAGS contains data from 241 African American speakers (out of 914).

The process of audio recording was integrated into LAP interviewing early on, albeit sporadically and not exhaustively, and continued to varying extents in many of the subsequent LAP components, but the tools for recording and the motivation for doing so have changed over time. In 1940, LANE collected speech samples from over 400 informants in the field on machines capable of recording to aluminum disk at 78 rpm. These disks are stored at the Special Collections Library at the University of Kentucky, with copies at the Library of Congress, where digital copies are also archived. For the Linguistic Atlas of the Upper Midwest (LAUM), Allen collected speech samples of many of his informants (Allen 1973: 2–3) on wire and tape.<sup>1</sup> Some interviews conducted for

---

<sup>1</sup>Unfortunately, the location of the disks that Allen's wire and tape recordings were transferred to has been unknown for quite some time, and although four of Allen's wire recordings apparently existed in 1967, wire is considered to be an unstable and, hence, temporary audio-recording format (Howren 1967: 32–33). Furthermore, none of these recordings and/or copies are in the current LAP inventory.

LAMSAS and LANCS, as well as the Linguistic Atlas of the Pacific Northwest (LAPNW), were recorded using reel-to-reel tape; these tapes are located at the University of Kentucky, and many have already been digitized. All Linguistic Atlas of Oklahoma (LAO) and LAGS interviews were recorded in their entirety on reel-to-reel tape, thus replacing the phonetic transcription in the field that up until then had always been the primary method of recording data. The targets on these tapes were later transcribed on paper. LAGS tapes, as well as digitized LAO and LAGS recordings, are housed at the University of Kentucky. LAWS interviews were taped on audio cassette, and both the cassettes and the digitized files are archived at the University of Kentucky. Finally, the digitized interviews from St. Kitts and Nevis are also stored at the University of Kentucky.

### 3 The New Atlas Interviews

#### 3.1 Motivation for the work

The main goal of today's LAP team is threefold. We want to make the historical LAP data available and accessible online for researchers and other interested parties. For researchers in sociolinguistics, specifically, we want to demonstrate that the LAP data can offer the best of all worlds: a set of dialectological targets that will facilitate comparisons across the LAP surveys and through time, the free-flowing conversation prized by traditional sociolinguistics for grammatical and phonological analysis, and a task that will elicit interviewees' language regard factors about language use in the area (see Preston 2018, 2019). And finally, today's LAP still hopes to collect the "speechways of the folk" but wants to make sure that all folk are represented.

Much of what we intend to do in Kentucky adheres to the framework of the most recent of the major LAP components, LAWS, in that we aim to audio record interviews with native informants in their entirety, transcribe the interviews, and rely on the resulting transcriptions as the primary data for analysis. At the same time, we want to incorporate advancements in recording that have occurred since the creation of the LAWS project in the late 1980s – notably, being able to record straight to digital in the field – and thus eliminate the need for digitization after the fact, freeing up time for other tasks.

#### 3.2 Grid units

Since an important aspect of this new work is to produce data that is comparable to older LAP data, using the same grid units established by the LAP for earlier

work (in areas where earlier work was done) is appropriate, barring any significant changes in the political boundaries (such as state and county lines) of that region. (For more on grid units in LAP, see, e.g., Kurath 1973: 39–40, Allen 1973: 23–24, Pederson 1990: 4) This will allow for real-time studies in the spirit of Johnson's (1996) work using LAMSAS interviews conducted in the states of Georgia and North and South Carolina in 1930 and her own follow-up interviews in those same states in 1990. Because the state of Kentucky and its political boundaries have not undergone a major shift since fieldwork was conducted in the mid-20<sup>th</sup> century as part of the LANCS project, we are using the same ones established in the state for that project.

### **3.3 The perceptual component**

Attention to language regard entails the collection and study of attitudinal data that contribute to our understanding of the viewpoints, beliefs, and ideologies that people have about language and language use (Preston 2018, 2019). Although aspects of regard have not before been sought in most LAP work, our work in Kentucky will address regard directly. The goal of incorporating interview techniques from perceptual dialectology, such as the draw-a-map task, into the new LAP interview is to trigger conversational data that serendipitously contains explicit and implicit clues to language regard.

The extended conversations from the map-drawing task as well as other interactions that may arise will offer many opportunities to report and analyze language regard issues. Recent work in anthropological and semanto-pragmatic approaches to conversational data also suggest that implicit regard factors may be extracted from data in this way. Irvine (2001: 25) explains: “the best place to look for language ideology may lie in the terms and presuppositions of metapragmatic discourse, not just in its assertions”. We have already piloted such analytic approaches, even with older LANCS data (George 2022, Passarelli 2023) and are convinced that we will encounter not only valuable overt assertions about language regard but also will find many opportunities to look at implicit or nonasserted material found in both semantic and pragmatic presuppositions and implicatures (Preston 2019, Preston & Evans 2023).

### **3.4 Community and informant selection**

Along the same lines, the communities surveyed for earlier projects should be considered as locations for new interviewing, although communities that are close in proximity and character to the original communities will serve as well.

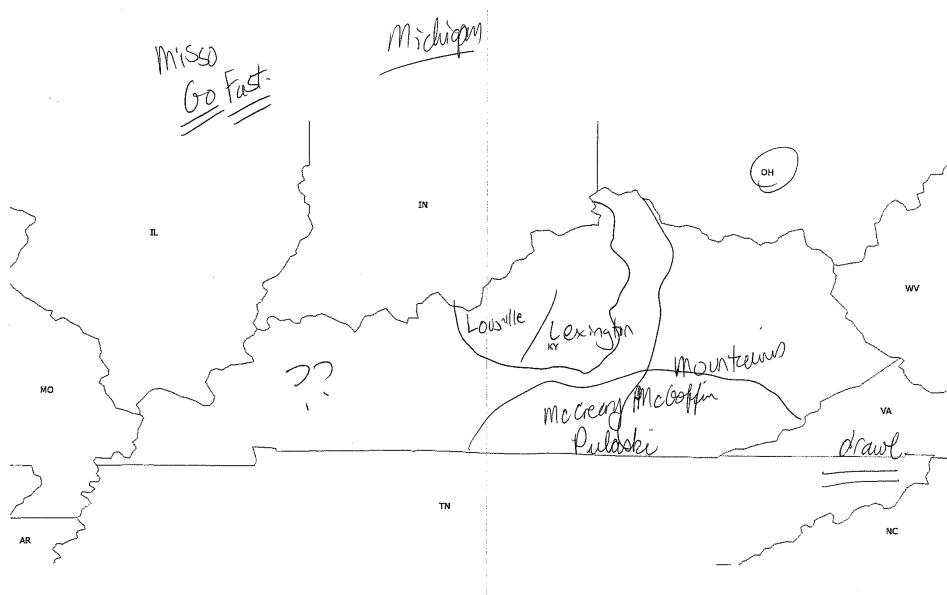


Figure 1: Hand-drawn perceptual dialectology map of a 66-year-old white female Louisville respondent from 2022.

Some of the communities that were sites of earlier LAP interviews no longer exist, sometimes only in name but other times as distinct political entities, having been absorbed by neighboring towns or cities, or becoming ghost towns through population loss, often as a result of the depletion or devaluation of the economic attraction that brought people there in the first place. Such situations obviously require the use of another community within the established grid unit for fieldwork.

With respect to informant selection, finding new informants based on their sharing similar demographic characteristics with earlier informants might be ideal for focusing only on linguistic change; however, our interest in language variation and in the LAP being representative of more diverse populations of communities is at odds with a methodology that tended to seek out NORMs, or at least elderly informants, for interviewing. We are also aware that students, especially those whose major interest in conducting an Atlas interview is for course credit, tend to not wander outside their comfort zone and instead lean on family and friends for the purposes of their interview. Thus, at this point we are mainly interested in informants who identify as natives of the communities they represent; older informants are preferred but not required, and informants with some

knowledge of farming (or ranching, depending on the location) terminology are also desired, as the worksheets include many such terms.

In Figure 2, the primary area where we have conducted and will continue to conduct new interviews is indicated by shading. Thus far, most of the new interviews have been conducted with speakers from Kentucky's two largest cities, Louisville and Lexington.

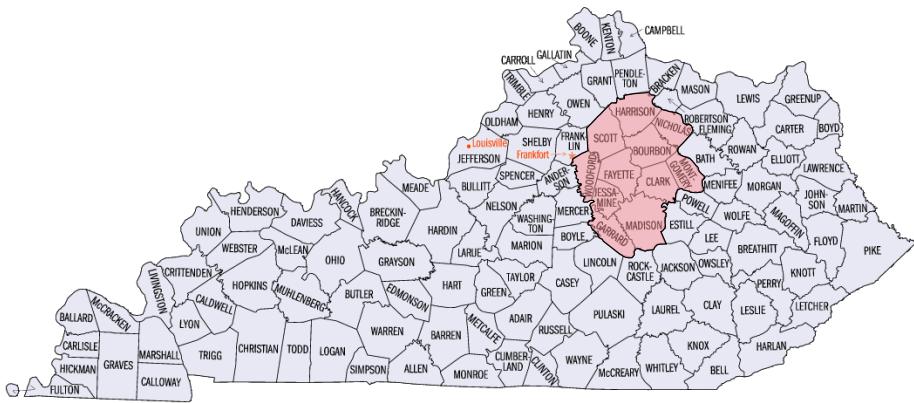


Figure 2: Primary area of focus for new Kentucky interviews. Original map from [https://commons.wikimedia.org/wiki/File:Kentucky\\_counties\\_map.gif](https://commons.wikimedia.org/wiki/File:Kentucky_counties_map.gif).

### 3.5 Worksheets

In that they play a pivotal role in any LAP investigation, we have extensively reviewed previous LAP worksheets to create the best worksheets possible for conducting the new Atlas interviews. In addition to the LANE worksheets, from which all LAP worksheets have been derived, the new Kentucky worksheets are a blend of two sets of worksheets in particular: LANCS and LAWS. More specifically, our aim is to collect data largely comparable to that collected in the LANCS fieldwork in Kentucky in the mid-20<sup>th</sup> century so we can look at linguistic change. At the same time, we want worksheets that will create a comfortable situation for fieldworker and informant alike, an environment in which the speech that is elicited not only yields lexical and phonological data comparable to earlier data but that can be used for syntactic and even pragmatic analysis as well. Additionally, we want to take advantage of this opportunity to incorporate a new component not previously dealt with explicitly in LAP interviews: data for perceptual dialectology (Preston 2018, 2019).

We began our creation of the new Kentucky worksheets with the LAWS worksheets (Pederson 1996b) as our base because of those worksheets being intended for shorter interviews that did not directly target grammatical features, for the most part, and in fact tried to delay the questions that might make speakers self-conscious, such as counting and reciting days of the year, names for parts of the body, and labels for different social categories, until later in the interview. The 12 semantic domains of the LAWS worksheets align with our own interests in such things as the vocabulary of house design and construction, as well as furnishings, kinship terminology, flora and fauna, and weather terms, although Pederson's motivation for the number and size of these domains (three domains per each side of two 90-minute cassettes) no longer rigidly applies to the digital recording we will be doing, which is discussed further below.

At the same time, some changes have been made to the LAWS format for new Kentucky fieldwork. First, some of the items that were appropriate for fieldwork in the West, such as geographical and weather-related differences unique to that region, were replaced by items more suitable to life in the Bluegrass State, many of which came from the LANCS worksheets. Worksheets will continue to be refined as we conduct interviews and find questions that yield nothing of interest or where we find evidence of new forms that we want to elicit.

Additionally, Pederson introduced a task to LAWS interviewing whereby an informant was handed pictures of a saddle and bridle and asked to name their parts. Although this activity experienced only limited success in Rocky Mountain fieldwork, there is enough interest in horses in various parts of Kentucky to suggest that continuing to perform such a task might be worthy of consideration.

We have continued Pederson's work of streamlining worksheets to the amount of information that can be elicited in the shortest amount of time possible and to strive for the least self-conscious speech possible while at the same time reducing the workload back in the Atlas office so that interviews can be made available to researchers more quickly. For instance, the first page of the LAWS worksheets asked for several pieces of information – most notably, the names of informants and their addresses – that hold questionable value for the interview and were probably only there for logistical reasons that can be, and, at least since the second author's fieldwork in Colorado in 2001–2004, have been collected by other and better means (most notably in the mandatory collection of signed human subject consent forms). This information was redacted in transcription and in the digitization of taped interviews before being posted for the public anyway. Thus, the current practice eschews asking informants for such information and only asks them for their zip code, as it establishes some idea of place and holds value in terms of the pronunciation of numbers.

The LAWS worksheets have some issues of redundancy and the potential to lead informants at the beginning of the interview, specifically on the first page, between the gathering of logistical information and the elicitation of targets in the first semantic domain. For instance, following the LAWS worksheets, a fieldworker asks for information about various members of the family, such as what their relationship was to the informant (e.g. mother, father, sister, brother), where they were born, and how they made their living, before going on to other biographical information pertaining to the informants, such as the schools they attended, their own profession, groups they were part of, etc. Then, the linguistic elicitation begins, with the first semantic domain targeting words for family, education, church, etc. The current worksheets instead integrate the gathering of both types of data with the use of an early *shotgun* prompt: *Tell me about your family growing up*. In this way, the terms informants used for family members can be elicited more organically and avoids the priming that might otherwise occur (e.g. *what did you call your mother growing up?*), while also leaving room for follow-up prompts for specific targets that the informant may not have used in their initial response (e.g. *you mentioned your mother; did you have other terms for her growing up?*).

As this will typically be the first, and oftentimes only, LAP interview that students will have the opportunity to conduct, as well as the fact that the worksheets include targets that may be unfamiliar to some if not all of these students, a sheet of fully formed questions for eliciting targets has been compiled and is being made available to fieldworkers, to be used in preparation for interviewing and/or used as an aid in the interview when fieldworkers encounter difficulty articulating a prompt on the fly. However, too much reliance on these questions might cause informants to shift from the conversational tone being sought for these interviews to a more formal manner; thus, we stress that fieldworkers should rely on these questions only as a very last resort.

Finally, we recently began to integrate a perceptual component to the LAP interview by engaging informants in the draw-a-map perceptual dialectology task, a technique already carried out in Kentucky by Cramer (2016). In this task, respondents are given a blank map of Kentucky and asked to 1) outline the areas where people speak differently, and 2) write in examples of and comments on the speech and identities of the people in the outlined regions. (See Figure 1 above for an example of a completed map.) Although these data may be treated independently in the perceptual dialectological tradition (e.g., Preston 2010), the primary use of the technique here is to trigger conversational data that serendipitously contains explicit and implicit clues to language regard. The pragmatic and discourse/conversation-analytic techniques outlined in Niedzielski & Preston (2003)

and Preston (2019) will allow us to interpret what our respondents say about all aspects of language, particularly, of course, their own language use and its contrast with others. This aspect of the new LAP interview will provide a broader and more in-depth look at language regard than is normally expressed in short-term language surveys and is an important characteristic of variationist interests in language change.

### 3.6 Audio recording

As mentioned above, various components of the LAP used available means of audio recording to document at least parts of interviews. Digitizing these analog artifacts has been a focus of the LAP for many years and continues to be so. It is important to note at this time that the LAP has six recordings from Kentucky in its holdings, although the quality of the recordings in terms of both listenability and completeness of interview has yet to be determined.

With respect to the new Atlas interviews, at least for the work being done in Kentucky, several methods of recording the interviews were considered. These typically implemented a combination of iPad, audio interface, lightning adapter, and microphone that required phantom power. However, these configurations were assessed as being less than ideal for interviews taking place out of controlled settings, vulnerable in a number of ways to the types of failures that have been an issue in previous LAP recording using simpler setups, and possibly harmful to the casual tone being sought in these interviews. We eventually settled on a configuration that consisted simply of an external microphone (Shure MV188 IOS Digital Stereo Condenser Microphone) that could be plugged directly into an iPad or iPhone, either that the students owned or that could be borrowed from the LAP office. In testing this configuration in the office, the resulting file provided high-quality sound, while also allowing us to bypass the time-consuming digitization process required for older analog recordings of, e.g., the LAGS and LAWS interviews, by going straight to digital.

### 3.7 Processing and analytical tools

Since the advent of the LAWS project near the end of the 1980s, the end-product of LAP interviews has been an audio recording of the interviews in their entirety. This is followed by transcription of the recording in its entirety in standard orthography in machine-readable text, which can subsequently be processed by computational tools such as AntConc or KwicKwic. The finished transcript can also be used as a guide for finding select features in the audio recordings or simply read by the interested party. The 70 transcriptions of the LAMR recordings

were done by Pederson and Antieau, and scribes under their direction, and later edited by Antieau. We are now investigating the use of software to create transcriptions of earlier LAP recordings, e.g. those collected for the LAGS and LAO components, followed by human editing of the manuscripts. As the new Atlas interviews will also be recorded in their entirety, such tools will also be useful for creating transcripts and compiling a corpus of the collected interview transcripts.

## **4 Discussion**

### **4.1 The initial new Kentucky interview**

The first of the new Kentucky interviews was conducted by the second author of this chapter in July 2021. The informant was found through the friend-of-a-friend approach; he was a 92-year-old man who was born in Louisville, Kentucky, but considered Lexington his home since his acceptance at the University of Kentucky, where he earned a degree in engineering. After graduation, he found a job with a Lexington company building roads throughout the Mid-Atlantic and Southeastern regions of the United States. He eventually established his own road construction business in Lexington and continued to do this work throughout the region. By the time of the interview, he had sold his business and retired. The three-hour interview was conducted in the informant's Lexington home, and the fieldworker implemented the revised worksheets and the audio recording configuration described above.

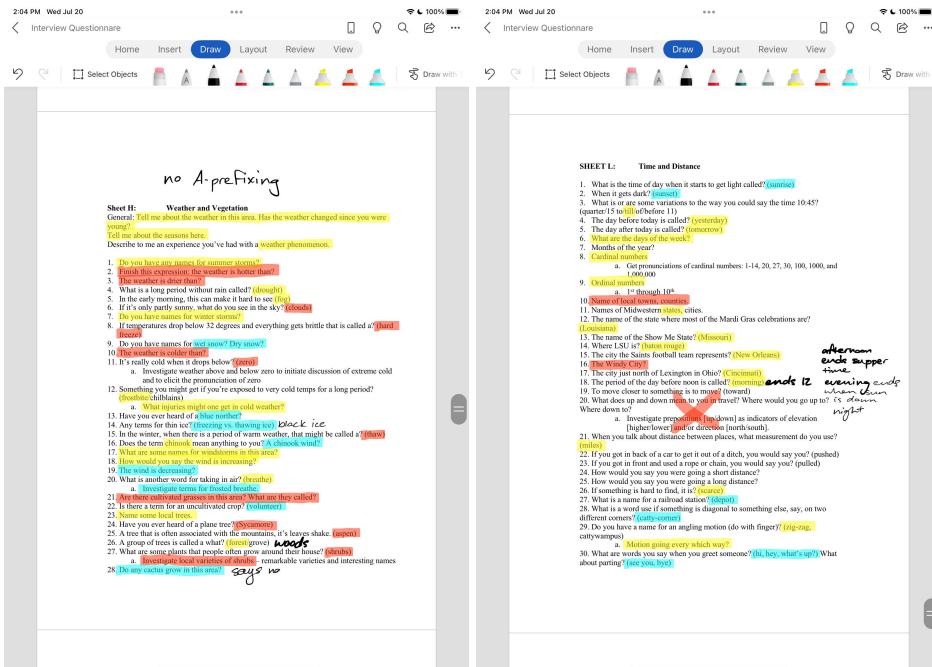
The interview presented no major issues, although the recording has not yet been transcribed in its entirety. The fieldworker had never worked in the area of perceptual dialectology, so he was apprehensive about broaching the subject with the informant and presenting the map for illustrating his perceptions of the speech of people throughout Kentucky and in the surrounding environs, as this task presents itself as a departure from the prompt-response format of most of the interview. However, the informant took to the task immediately, marking areas where he thought people in the surrounding region talked differently than Lexington inhabitants and telling why he believed so (aided greatly by his travel for work throughout the region).

### **4.2 Subsequent interviews**

Subsequent interviews have been conducted by students as part of an effort to both expand the present-day informant pool and to develop training materials

that could be used by researchers interested in contributing to the LAP. Eight interviews with speakers in Lexington and Louisville were conducted by graduate students, and four interviews with speakers from three counties in western Kentucky were conducted by a single undergraduate. The transition to a hybrid interview format has also signaled a technological transition. The LAP has traditionally been heavily dependent on an overabundance of paper. Student interviewees were encouraged to experiment with different means of deploying the new LAP worksheets in the field, with most opting against printed worksheets in favor of digital ones.

The images in Figure 3 are an example of how one student used digital worksheets on a tablet as a way for keeping track of what was occurring in the interview. The student used color coding to note information such as what questions were asked and answered/not answered as well as what subjects were unfamiliar to the speaker. The goal now is to test various means of note-taking and digital map-drawing in the hopes of developing a uniform system that can be included in our training materials and then employed in future interviews.



With continued digitization efforts of older field records, recording, and notes, along with the use of the technologies we now have at our fingertips, we hope to bring the LAP into the 21<sup>st</sup> century.

## 5 Conclusion

In this chapter, we have discussed current efforts to conduct new Atlas interviews in the state of Kentucky so that we can arrive at a better understanding of change and variation in the speech of the Bluegrass State. But also, we hope that we can encourage students looking for research topics (or who are just interested in the speech of the state in which they attend college) to consider adding to LAP coverage of the state.

The method we have arrived at is informed by earlier LAP components, especially LAWS and LANCS, so that we can compare the results of new interviews to older LAP interviews. At the same time, we need to adapt, by tweaking worksheets so that we elicit the best data possible and by adopting newer technologies that create opportunities for analyzing our data in new and exciting ways.

We are also hoping that others will follow our lead and conduct interviews in other regions, states, and communities that were surveyed by earlier LAP components. Over the last several years, we have delved into the LAP archives and have digitized complete collections full of linguistic, cultural, and historical information that has not been properly tapped, such as the Linguistic Atlas of Hawaii and the Linguistic Atlas of the Pacific Northwest, as well as records from the Hudson Valley collected in the early 1940s. We are now nearing the point in this process that we can run investigations of linguistic features on the 5000-plus interviews in the LAP holdings, as evidenced by our recent paper on a-prefixing (Burkette & Antieau 2022), and we will soon be able to do this all online (Burkette forthcoming). The LAP represents a massive amount of time, energy, and effort – on the part of fieldworkers, LAP office workers and administrators, as well as the speakers who gave of their time and community knowledge – and our goal is to see that time appreciated through continued use of LAP materials and the expansion of contemporary LAP interviews.

## Acknowledgements

The authors would like to thank Dennis Preston, Jennifer Cramer, and Shane O’Nan for their contributions to this paper and this work. Thanks also to the two anonymous reviewers for their helpful suggestions.

## References

- Allen, Harold. 1973. *The linguistic atlas of the Upper Midwest, in two volumes*. Minnesota: University of Minnesota Press.
- Antieau, Lamont. 2012. Ascending kinship terminology in middle Rocky Mountain English. *English World-Wide* 33(2). 185–204.
- Antieau, Lamont & Clayton Darwin. 2013. Fatback and gunnysacks: Lexical variation in the Southeastern U.S and the Middle Rockies. *Southern Journal of Linguistics* 37(2). 39–56.
- Baker, Philip & Lee Pederson. 2013. *Talk of St. Kitts and Nevis*. Kitts, Nevis. London & Colombo: Battlebridge Publications.
- Burkette, Allison & Lamont Antieau. 2022. A-prefixing in linguistic atlas project data. *American Speech* 97(2). 167–196.
- Cramer, Jennifer. 2016. *Contested southernness: The linguistic production and perception of identities in the borderlands*. Publication of the American Dialect Society 100. Durham, NC: Duke University Press.
- George, Crissandra. 2022. *Ambiguous appalachianess: A linguistic and perceptual investigation into ARC-labeled Pennsylvania counties*. University of Kentucky. (MA thesis).
- Hamilton-Brehm, Anne Marie. 2003. *A foundational sample of El Paso English*. University of Georgia dissertation.
- Howren, Robert. 1967. Iowa materials for the linguistic atlas of the Upper Midwest. *Books at Iowa* 6(1). 29–35. DOI: 10.17077/0006-7.
- Irvine, Judith T. 2001. “Style” as distinctiveness: The culture and ideology of linguistic differentiation. In Penelope Eckert & John R. Rickford (eds.), *Style and sociolinguistic variation*, 21–43. Cambridge: Cambridge University Press.
- Johnson, Ellen. 1996. *Lexical change and variation in the Southeastern United States, 1930–1990*. Tuscaloosa: The University of Alabama Press.
- Kurath, Hans. 1972. *Studies in area linguistics*. Bloomington: Indiana University Press.
- Kurath, Hans. 1973. *Handbook of the linguistic geography of New England*. New York: AMS.
- Niedzielski, Nancy & Dennis R. Preston. 2003. *Folk linguistics*. Rev. pbk. edn. Berlin: de Gruyter Mouton.
- Passarelli, Nicholas. 2023. *Language ideologies in the informant biographies of the linguistic atlas project*. University of Kentucky. (MA thesis).
- Pederson, Lee. 1990. *A plan for the linguistic atlas of the Western states*. Unpublished manuscript. LAP archives, University of Kentucky.

- Pederson, Lee. 1996a. LAMR/LAWS and the main chance. *Journal of English Linguistics* 24(3). 234–249.
- Pederson, Lee. 1996b. LAWCU worksheets. *Journal of English Linguistics* 24(1). 52–60.
- Pederson, Lee & Michael Madsen. 1989. Linguistic geography in Wyoming. *Journal of English Linguistics* 22(1). 17–22.
- Pederson, Lee, Susan L. McDaniel & Carol M. Adams (eds.). 1986–1993. *Linguistic atlas of the Gulf States, 7 volumes*. Athens, Georgia: University of Georgia Press.
- Preston, Dennis R. 2010. Mapping the geolinguistic spaces of the brain. In Alfred Lameli, Roland Kehrein & Stefan Rabanus (eds.), *Language and space: An international handbook of linguistic variation*, vol. 2: Language mapping (Handbooks of Linguistics and Communication Science (HSK) 30/2), 121–153. Berlin, New York: de Gruyter Mouton.
- Preston, Dennis R. 2018. Language regard: What, why, how, whither? In Betsy Evans, Erica Benson & James Stanford (eds.), *Language regard: Methods, variation and change*, 3–28. Cambridge: Cambridge University Press.
- Preston, Dennis R. 2019. How to trick respondents into revealing implicit attitudes: Talk to them. *Linguistics Vanguard* 5(s1). 20180006. DOI: 10.1515/lingvan-2018-0006.
- Preston, Dennis R. & Betsy E. Evans. 2023. Language regard. In Robert Bayley & Erica Benson (eds.), *Needed research in American dialects*, 246–267. PADS 2023.

## **Part III**

# **Dialectology, linguistic identity, and social factors**



# Chapter 10

## “Das ist dann schon total cool zu sagen, *Machanot*”: Revealing speakers’ justifications for linguistic choices

Esther Jahns

Carl von Ossietzky Universität Oldenburg

In this paper I introduce the linguistic-positioning task, a method that reveals speakers’ perspective on their own linguistic choices and the choices of others. It was developed for and will be demonstrated through the research on linguistic choices of German-speaking Jews in today’s Berlin, Germany. Like other contemporary Jewish communities (Bunin Benor & Hary 2018, Kahn & Rubin 2016), German Jews make use of what Bunin Benor (2008: 1064) has defined as a “distinctively Jewish linguistic repertoire”. This repertoire consists mainly of lexical items from Yiddish and Hebrew that are integrated into German. Due to the two different donor languages, the repertoire allows for inter- and intra-speaker variation as several concepts can be expressed either in Hebrew or in Yiddish like Hebr. *Mizwa*, Yidd. *Mizwe* (‘good deed’). Thus, speakers as active agents can express social meaning through their meaningful choices. It will be demonstrated how selective lexical items are used as stimuli to enhance a meta-discussion on linguistic choices. The findings show that language ideologies have the biggest impact on speakers’ choices and on their interpretation on the choices of others. Even though the method was developed for a distinctive multilingual community, it is applicable to reveal speakers’ justifications and explanations concerning variation in other multi- and monolingual communities as well.

### 1 Introduction

Jewish speakers in contemporary Berlin have access to and make use of a “distinctive Jewish linguistic repertoire” (Bunin Benor 2008: 1064) like Jewish speakers



in other contemporary communities (Harry & Bunin Benor 2018). This repertoire consists mainly of lexical items from Hebrew<sup>1</sup> and Yiddish that are integrated into German (see examples (1) and (2)).<sup>2</sup>

- (1) Ich fahre auf *Machane*. (Rebecca, 23:31)<sup>3</sup>  
‘I am going to a Jewish summer camp.’
- (2) Bist du *brojges*? (Petra, 22:25)  
‘Are you annoyed?’  
Hebrew *Machane* (‘summer camp’), Yiddish *brojges* (‘angry, annoyed’).

As research on other contemporary communities has shown, the function of the repertoire is to index alignment to the Jewish community, but also to index subtleties of speakers’ Jewishness, i.e., to align to or distinguish from other Jewish speakers (Bunin Benor 2009: 234–235).

In this study I am investigating the speech pattern of Jewish speakers in 21st century Berlin.<sup>4</sup> This speech pattern could be labeled a (dia)lect. For my approach and research interest the concept of the repertoire is, however, more adequate, as will be explained in more detail in the next section. The focus of this study is the variation the repertoire allows for, how speakers make use of it and how they explain their choices. Therefore, I developed a method to grasp the meaning of variation from the speakers’ perspective and to reveal their justifications and explanation for their linguistic choices and the choices of others. In this article I will describe and exemplify the method through one of the main categories that emerged from the data as affecting speakers’ choices, i.e., language ideologies towards Hebrew and Yiddish. Moreover, I will demonstrate the applicability of the methods for other multi- or monolingual communities where variation occurs.

---

<sup>1</sup>Hebrew encompasses Biblical as well as Modern Hebrew elements. As the aim of the study is to grasp speakers’ perspective, a differentiation is not made as informants themselves generally do not differentiate between Biblical and Modern Hebrew.

<sup>2</sup>Yiddish and Hebrew words that are part of the repertoire are used, if written, in Latin script. I decided to use, if possible, the spelling that speakers proposed or referred to Weinberg (1994, 1973).

<sup>3</sup>All names of the interviewees have been changed.

<sup>4</sup>The method described in this article is part of my study on linguistic choices and language ideologies of German-speaking Jews in contemporary Berlin (Jahns 2024).

## 2 Research background

### 2.1 Multilingualism in Jewish communities in past and present

Jewish communities have almost always been multilingual, from the 6th century BCE until today (Spolsky & Bunin Benor 2006). However, while some aspects of this multilingualism, such as the importance of Hebrew, remained stable, others changed throughout time. Due to conquest and expulsion, Jewish communities migrated within the area of the Mediterranean, but also to other parts of Europe and to the Middle East (Peltz 2010: 136). In the Diaspora, wherever Jewish communities settled, a triglossic pattern emerged. Hebrew-Aramaic<sup>5</sup> was the sacred language reserved for the religious domain, the language of the surrounding society was used for communication with non-Jews and eventually a third language emerged that was used for communication within the Jewish community outside the religious domain. This third language evolved due to language contact between the surrounding varieties, Hebrew-Aramaic, and other Jewish languages that the community brought with it from former areas of settlement. In the literature these vernaculars for in-group communication in Jewish diasporic communities that emerge through contact have been labelled "Jewish languages" (cf. Peltz 2010, Fishman 1981, Gold 1981).

In the case of medieval Germany, the surrounding variety was Middle High German (MHG), or more precisely dialects of MHG. The first Jewish communities that settled in the territory of today's Germany mainly came from France and Italy and brought the Jewish languages they had spoken there with them. The contact with MHG and Hebrew-Aramaic led eventually to the emergence of Yiddish, as the Jewish language used by Jewish speakers in Ashkenaz<sup>6</sup>, and later on throughout western, middle and eastern Europe (Weinreich 2008, Jacobs 2005).

One main aspect of discussion and controversy in the field of comparative Jewish linguistics<sup>7</sup> was, however, defining the object of research and comparison, i.e., to develop unambiguous criteria for Jewish languages (e.g. Rabin 1981, Fishman 1981). Both parts of the label were under discussion, the Jewishness of the respective language as well as the question generally discussed in linguistics,

<sup>5</sup>Aramaic is considered the second important language of religion as important parts of the Talmud are written in it (see Myhill 2004: 111). Therefore, Hebrew-Aramaic is often used as a label for the sacred language of Judaism.

<sup>6</sup>Hebrew name for medieval Germany

<sup>7</sup>Several other labels have been used for the linguistic field that deals with the comparison of Jewish languages, e.g., sociology of Jewish languages (Fishman 1981), Jewish interlinguistics (Wexler 1981).

whether a distinction can be made between language and dialect on linguistic grounds. Concerning the latter, other labels were proposed like “lect” (Gold 1981) or “religiolect” (Hary & Wein 2013).

In the 21st century the differentiation between language and dialect becomes even more crucial concerning Jewish languages, as most of them range towards lesser linguistic distinctiveness in relation to the surrounding language.<sup>8</sup> Some scholars even denied the existence or emergence of Jewish languages in the 21st century at all (Myhill 2004).

Therefore, a definition as a lect could seem more adequate in the 21st century. However, Bunin Benor proposed another approach, namely the “distinctive Jewish linguistic repertoire”, which she defines as “the linguistic features Jews have access to that distinguish their speech or writing from that of local non-Jews” (Bunin Benor 2008: 1064). This approach changes the focus from the language (or lect) to the speech community and the (multilingual) resources to which its speakers have access. Thus, the question of (sufficient) distinctiveness is obsolete, which is not only important for Jewish speech patterns in the 21st century but also crucial when it comes to inter- and intra-speaker variation. Speakers vary not only concerning the elements from different donor languages, but also more generally concerning the number of items from the repertoire they integrate into the respective majority language. It would therefore be difficult or even impossible to determine a specific number of elements as a minimum requirement to consider the speech pattern an instance of the distinctive (dia)lect (Bunin Benor 2010: 166–167). Moreover, even in former times the use of Yiddish or Yiddish items was not stable across speakers, but was influenced by several categories like religiosity, geography, profession and gender (Jahns 2024: 37–38).

The focus on the speakers allowed by the repertoire approach is also in line with 3rd wave sociolinguistics where the speaker is conceptualized as an active agent making use of all the linguistic resources she or he has access to (Eckert 2012, Irvine 2001, Johnstone et al. 2006). Thus, speech patterns in Jewish communities in the 21st century, as well as in the past, are best described and analyzed by the repertoire approach, which I will follow here.

## **2.2 Social meaning and language ideologies in the multilingual repertoire**

The variation the repertoire allows for is the prerequisite for social meaning: If two or more lexical elements denote the same referent, the choice of one of the

---

<sup>8</sup>This is, however, not the case in ultra-Orthodox communities where the daily vernacular is often Yiddish.

possible variants may be used for the social positioning of the speaker. (Jahns 2024: 66) defines social meaning as “any social information speakers give about themselves or their positioning in the social landscape through linguistic means that goes beyond the denotational meaning”. This is, however, only successful if the chosen element is part of what Irvine calls a “system of distinction” (2001: 22). This means the element has to be in contrast to another element in order to be salient and perceived by the hearer, and it has to be interpreted in the way the speaker intended. The interpretation depends on speakers' and hearers' social and linguistic background and experience and their contact with different groups (Irvine 2001: 24). However, this system is not stable but dynamic, and changes through use, like the social meaning or index of a distinctive variant is open to interpretation and re-interpretation (Eckert 2008). To trigger the intended interpretation or to understand what was meant to be expressed is therefore easier within a community that shares practices, ideas, forms of talk or history as “variation constitutes a social semiotic system capable of expressing the full range of a community's social concerns” (Eckert 2012: 94).

It must be noted that Jewish speakers in Berlin are not part of a homogeneous community. In fact, Berlin's Jewish community is highly diverse (Belkin 2017). However, all speakers do share a cultural and religious heritage which has been intertwined with two languages to which they have access due to their Jewishness, even though to different extents. The role that these two languages, Hebrew and Yiddish, have been playing in past and present for the Jewish community worldwide and for Berlin's Jewish community in particular influences the symbolic value that speakers attribute to them, and this value in turn influences their roles. It can therefore be assumed that language ideologies will have an impact on linguistic choices, as they can be defined as “the cultural (or subcultural) system of ideas about social and linguistic relationships, together with their loading of moral and political interests” (Irvine 1989: 255).

The variation or the “system of distinction” (see above) that the repertoire allows for is twofold. Firstly, it is already a choice to use an item from the repertoire instead of an item from German as the majority language. Secondly, the repertoire contains variants for several concepts or referents, mainly from the two donor languages but also from different dialects of Yiddish, like Hebrew *Schabbat Schalom* and Yiddish *Gut Schabbos* or *Git Schabbes* ('Shabbat greeting'). In addition to the assumed role of language ideologies concerning the choices, it is also possible that single elements are perceived as shibboleths for distinctive subgroups within the community or are used to express a certain emotional value. This is again an instance of contrast and distinction in relation to another element from the repertoire.

In sum, due to salience, distinction and contrast, the linguistic repertoire to which Jewish speakers in Berlin have access gives them multiple opportunities to vary their speech and to express social meaning. In the following sections, I will explain how I designed my study to explore the reasons for speakers' choices from their perspective.

### **3 Research design**

#### **3.1 Exploring the field through expert interviews**

The data collection started with expert interviews with nine Jewish leaders in Berlin. Expert interviews can be used in the research design either as the main method or to explore the field by gathering expert knowledge on the topic (Meuser & Nagel 1991, Bogner & Menz 2005). In this case the aim was the latter, as there has not been any research on the speech patterns of contemporary Jewish communities in Germany so far. The main question to be answered through the expert interviews therefore was whether the Jewish community in Berlin actually has access to and makes use of a "distinctive Jewish linguistic repertoire" (Bunin Benor 2008: 1064) like contemporary Jewish communities in other countries (cf. Hary & Bunin Benor 2018). Subsequent aims of the exploration – if the existence of the assumed repertoire could be proven – were to get the experts' insights on variation the repertoire allows for and an actual collection of lexical items that are part of the very repertoire.

Even though Meuser & Nagel (1991: 443) state that it is the researcher who decides according to research topic and aim who can be defined as an expert, the decision about experts concerning language use within the Jewish community was not straightforward. For the aim of this study, I was looking for persons who are especially attentive concerning languages and language use or have a certain sensitivity for linguistic matters (e.g., multilinguals like Russian L1 speakers in Germany, persons working on linguistic questions without being linguists, teachers that are trained to evaluate linguistic expression of pupils) and who would come into contact with a lot of different speakers. The assumption was that group leaders might enhance the use of the repertoire for pedagogical or identity-building reasons,<sup>9</sup> are addressed with items from the repertoire as an expression of alignment, and are in contact with other leaders/groups and might

---

<sup>9</sup>This is for example the case for the website of the Jewish unity-community in Berlin and its monthly journal, where several concepts are used in Hebrew with an explanation in German following: <http://www.jg-berlin.org/>, 14.11.2022.

perceive differences. I tried to capture the diversity of Berlin's Jewry by choosing leaders or representatives from different Jewish groups in order to gather a wide range of variation in the use of the assumed repertoire. The groups differed on various levels, e.g., motivations for grouping (religion, education, leisure, culture), membership categories, regularity of meetings, stability and dynamics of the groups. In addition, the experts themselves differed according to their age, gender, religiosity and L1s; they were rabbis, school and university teachers as well as leaders from different associations.

The expert interviews proved that Jewish speakers in contemporary Berlin make use of a "distinctively Jewish linguistic repertoire", as all experts agreed upon its existence. They also stated unanimously that there is intra- and inter-speaker variation, even though the opinions differed concerning the presumed reasons for it. During the interviews the experts provided items that are, according to them, part of the repertoire, and reported their use of those that have been collected in previous interviews. This collection was further expanded during the main data collection, but the main part was contributed by the experts. As of now the collection comprises 219 lexical items and formulaic sequences (see Jahns 2024) and the online lexicon *Judäo-Deutsches Wörterbuch* (Jahns 2022–present). This is, however, only a part of the repertoire and probably not all items that are part of the repertoire can be elicited in this manner.

However, I argue that the items mentioned by the experts are adequate stimuli to trigger explanations concerning variation and linguistic choices. It can be assumed that the experts mentioned items that were salient to them, as mentioning linguistic elements in meta-communications indicates their salience (Lenz 2010: 96). Auer (2014) distinguishes between three dimensions of salience: physiologically, cognitively, and sociolinguistically conditioned salience. The first one describes the pure distinction in relation to the rest of an utterance, which is probably the case for almost all items from the repertoire, as the majority stem from languages other than German. Cognitively conditioned salience refers to items that are distinguished by the hearer because s/he did not know or did not expect this item in this context. This latter means that the hearer would have used something different in this context or did not expect the speaker to make use of this item, which indicates that this is a case where inter-speaker variation happens. When salience is sociolinguistically conditioned the item stands out because it has a strong emotional value for the hearer which means it is linked with a social evaluation and indexes a certain type of speaker (Auer 2014: 9–10). It can be assumed that the items the experts mentioned are salient to them regarding the second and the third dimension. The experts, who themselves share the repertoire, have been in contact with items they did not know or expect (2nd

dimension) or with items that index a certain type of speaker (3rd dimension). Thus, in addition to the fact that making use of items from the repertoire is already an instance of variation (instead of using an item from majority-German), the items mentioned by the experts can be considered signs of differentiation or difference (Gal & Irvine 2019) or even shibboleths (Busch & Spitzmüller 2021). Therefore, these elements are adequate to trigger explanations about linguistic choices.

### **3.2 Justifying perceived language use and explaining personal choices against the use of others**

The main data collection consisted of semi-structured interviews that included a task based on stimuli that were selected elements from the repertoire. The interviews were conducted with 12 Jewish speakers in Berlin and had a duration between 30–100 minutes. All interviews were recorded and transcribed with f4. The aim of the interviews was to reveal categories that affect speakers' linguistic choices. As the analysis of the interviews was oriented towards the principles of Grounded Theory (Charmaz 2010), the categories were not predefined, but were meant to emerge from the data.

Speakers were between 25 and 59 years old at the time of the interview and had different first languages (German, Russian, Polish, Swiss German, Russian and German). Eight of them considered themselves as religious (ranging from Modern Orthodox to Reform), four as secular. The interviews were conducted, when possible, in locations where the speaker would make use of the repertoire (this was for some speakers their working place) or at least in informal settings like cafés.

The interviews started with four general questions meant as ice-breakers but also as a frame-setter. The first two questions were about the Jewish community in Berlin without explicit reference to linguistic characteristics. Two further questions followed where the speaker was asked about his or her individual positioning within this community.

The main part of the interview consisted of the task which will be described in detail below. The development of the task was influenced by methods from Perceptual Dialectology (Cramer 2016, Preston 1999, Preston 2010). The aim of Perceptual Dialectology (PD) is to grasp laypersons' perspective on (dialectal) variation and where it comes from (Cramer 2016: 1) which is also the aim of this study.

The draw-a-map task is the most prominent task in PD. Laypersons are asked to locate different dialects within their home-country by literally drawing them,

and their boundaries, into an empty map (Preston 1999: XXXIV). My design was based mainly on the draw-a-map task as it is meant to visualize variation and difference in language use.

Yet the researched community are Jewish speakers in Berlin, which means that, like in other urban areas, boundaries between language use of different speakers or groups of speakers cannot be drawn with exact lines or isoglosses, but a visualization might look more like an intersection of sets. For these speakers – and research in urban centers in general – there are parts of the individual's language use that overlap with the use of others and it is not possible to distinguish these uses in a geographical way like it is done in the draw-a-map task. It will be described below how I adapted this method to fit my qualitative research design. Moreover, in addition to identifying speakers' own language use against the use of others, my task also needed to uncover the possible social meaning for the perceived variation as an expression of the individual positioning within the community.

### 3.3 Selection of stimuli

From the collected items that are part of the repertoire of German-speaking Jews in Berlin I selected 78 as stimuli for the main data collection. From these 78, I chose 50 as a minimum for each interview and added items when the interviewee was especially comfortable with the task and eager to continue. According to the principles of Grounded Theory the selection of items was also adapted in the course of the interviews when distinctive items appeared to be particularly appropriate for differentiation. This was especially the case for items from the third category (see below). The items that I selected fall into five categories.

The first category included items that have one or more variant(s) within the repertoire. This applies mainly to concepts that exist in a Hebrew and a Yiddish version in the collection like Hebrew *Challa* and Yiddish *Challe* ('braided bread eaten on Shabbat and holy days'), Hebrew *dawka* and Yiddish *dawke* ('on purpose; just to annoy'). But there were also concepts with more variants like Hebrew *Kippa*, Yiddish *Jarmulke* or *Keppele* and German *Käppchen* ('skullcap'). As the use of items from the repertoire generally indexes alignment with the Jewish community, we can assume that concepts with several variants within the repertoire allow for positioning within the community.

The second category contained items that are, according to the literature or to the experts, typical for speakers from distinct networks within the community. These are items that are assumed to be known mainly by speakers who are, for example, involved in religious learning, like Hebrew *Nafka Minna* ('practical

difference'); or by younger speakers who go to *Machane* ('summer camp'), like Hebrew *Madrich* ('group leader'); or by older speakers whose families have been living in Germany for generations, like Western Yiddish *Barches* ('braided bread eaten on Shabbat and holy days').

Items that triggered strong emotional reactions by some of the experts formed the third category. These were either positive reactions for items that were especially liked, e.g., *brojges* ('annoyed'), *Tuches* ('buttocks'), or negative reactions concerning items that were strongly disliked, e.g., *jiddische Mame* ('Jewish mother'), *Chugist* ('group leader'). These are examples of sociolinguistically conditioned salience. The strong reaction points either towards language ideologies or might be triggered due to the fact that the respective item is perceived as an index for distinctive speakers.

The fourth category consisted of items that were interesting concerning their integration into German. These are either Yiddish items that, due to the linguistic closeness of Yiddish and German, need special flagging in order to be recognized as being part of the repertoire like *Mensch* (Yiddish meaning = 'good, loyal person'; German meaning = 'human being'), or items that are integrated through periphrastic forms, which is also a common strategy of integration in other Jewish languages (see Jacobs 2005: 210–212 for Yiddish). Examples with German auxiliaries *haben* ('to have') and *sein* ('to be') are *Moire haben* ('to be afraid'), *Mazliach sein* ('to succeed').

As a fifth category I included items from other contemporary Jewish repertoires (Klagsbrun Lebenswerd 2016, Bunin Benor & Cohen 2011) in order to answer the question of whether a global repertoire exists or whether we find local interpretations of this very repertoire. This very small category contained local innovations from Jewish Swedish like *goga* ('synagogue') and Yiddish items from the repertoire of American Jews that might be avoided by German-speaking Jews because of the linguistic closeness between Yiddish and German (see above) like *heimisch* ('homey') or *kwetschen* ('to complain')<sup>10</sup>.

### 3.4 Linguistic-positioning task

After the four initial questions that were described above, the main part of the interview started with the task that I developed for this study and labelled *linguistic-positioning task*, which can be applied for research on linguistic variation in other groups and settings as well (see below).

Each item that was chosen for this task was written on a card and the interviewee got her or his pile of cards. The task was then to consider item after item,

---

<sup>10</sup>The German verb *quetschen* has, however, different semantics, meaning 'to squeeze'.

read each item aloud, comment on it, correct it if necessary, and finally classify it according to the speaker's own use. The three possible categories were:

1. I know this item and make use of it.
2. I know this item, but wouldn't make use of it.
3. I don't know this item.

Eventually three piles of different heights emerged, visualizing each speaker's own use (first pile) against the use of others (second and third pile). The second pile consisted of items the speaker consciously chose not to use. So, like the draw-a-map task in PD, speakers made a sort of virtual map of their language use. It is also possible to understand this task as an inversion of the Language Situations-method introduced by Wiese (2020) where speakers have to imagine a specific situation and how they would describe an event in this very situation (i.e., spoken message to a friend, spoken witness statement). In the linguistic-positioning task, speakers get the linguistic element as a trigger and are encouraged to imagine a situation and/or an addressee (e.g., friend, colleague, grandparent) with whom they would make use of this very item or, if they don't use it themselves, think of a person who would. It could therefore also be described as a focused interview where all participants were confronted with the same stimuli (cf. Flick 2007: 195–202).

The structure of the task together with the frame-setting starting questions had several advantages and facilitated imagining contexts in which the item is or was used. First of all, the fact that the interviewee had to read the item on the card aloud was helpful, as elements from the repertoire are mainly used in spoken language.<sup>11</sup> Thus, reading, speaking and hearing the element addressed several senses and made it more likely that at least one of them would trigger the memory of a situation in which it was used by the speaker or by somebody else. Moreover, the card itself was something the speaker had to grasp and pile. This haptic part of the task, as well as the focusing on the item itself, helped to dissolve stress that might emerge in a face-to-face interview situation.

For the interviewer, the biggest advantage of the interview situation (in contrast to a questionnaire) is the opportunity to dig deeper whenever an interesting aspect or (unexpected) emotion emerged. The items that were classified in the second category (known, but not used) were of particular interest as they had

---

<sup>11</sup>In Jahns (2024) it is explained in detail why items were presented in writing even though the repertoire is mainly used in spoken language.

the potential to express social meaning, being interpreted as signs of difference (Gal & Irvine 2019) or shibboleths (Busch & Spitzmüller 2021). Thus, the whole interview was a meta-linguistic discussion that was triggered by elements from the repertoire which were classified during the discussion in order to form a visualization of the speaker's own language use in contrast to the use of others. Through the task, the speaker positioned her- or himself towards other speakers and their use (Spitzmüller 2013), sometimes offering explanations that show the wish of alignment with or distinction from other speakers or speaker groups through the language use. A number of questions were prepared that the interviewer could ask whenever it seemed necessary or interesting. These questions included contexts and domains of use, (typical) users, and integration into German.

## 4 Findings

### 4.1 Categories affecting speakers' linguistic choices

Through the process of coding the transcribed interviews, three main categories emerged that revealed to have an impact on speakers' linguistic choices concerning the elements from the repertoire. Even though a lot of codes and subsequent categories emerged when coding the data, the following three were the most prominent and emerged through all speakers:

- Distinctively Jewish linguistic awareness
- Language ideologies towards Yiddish and Hebrew
- Intra-speaker variation

In order to demonstrate how the method that I developed works, I will give some examples from the second category and describe briefly what should be understood by the two others (for a detailed description see Jahns 2024).

The category "distinctively Jewish linguistic awareness" includes all utterances that demonstrate that speakers have a distinctive awareness concerning the languages that they have access to due to their Jewishness (in this case Yiddish and Hebrew). This awareness can also be described as an advanced knowledge about etymology of the elements as well as possible variants within and across the two languages and grammar, especially concerning Hebrew. Interestingly, this awareness is very important to speakers and highly valued by them.

- (3) *Jesch!*<sup>12</sup> [...] das ist aber auch schon so 'n Codewort, dass man so 'n bisschen Hebräisch kann  
*Jesch!* [...] that has already become sort of a code word to show that you know a little Hebrew. [Mirjam, 08:02]

Concerning the variation within the speakers, first it has to be noted that all speakers make use of items from both languages, sometimes even for the same concepts, e.g., Yidd. *Tallis*, Hebr. *Tallit* ('prayer shawl'). None of them uses exclusively Yiddish or exclusively Hebrew elements, but they do use them to varying extents, with one informant trying to use as little Yiddish as possible. However, no matter how many elements they use, all speakers consider Yiddish elements to be more appropriate for informal situations with family and close friends and report using Hebrew elements predominantly in more formal situations with unknown interlocutors. Thus, the distribution of the elements from the two languages integrated into German follows a pattern that is similar to what Koch & Oesterreicher (1986) describe as language of immediacy vs. language of distance.

## 4.2 Language ideologies towards Yiddish and Hebrew

As hypothesized, language ideologies towards Hebrew and Yiddish turned out to have the strongest influence on speakers' linguistic choices and their interpretation of utterances from others. This means that speakers explained why they would or wouldn't use items often through (explicit or implicit) statements about the respective donor language. These explanations are in line with Silverstein's (1979: 193) definition of language ideologies as a "set of beliefs about language articulated as a rationalization or justification of perceived language structure and use". Interestingly, speakers started their evaluation of an item quite often with a remark concerning its language of origin, even though this was not part of the task. However, this knowledge, that I labeled distinctively Jewish linguistic awareness (see above), was an indispensable prerequisite for using languages and their symbolic value as an explanation. Knowing about the origin of a lexical element and being able to differentiate between the two donor languages was the basis for justifying one's language use through beliefs about the respective language of origin and/or typical speakers who make use of it.

The explanations and justifications uttered by the speakers are not expressions of one single ideology, but form clusters or "language ideological assemblages" (Kroškrty 2018: 134). Generally speaking, it can be said that the ideologies towards Yiddish reflect the more emotional, but also conflicting, relation speakers

---

<sup>12</sup>*Jesch* ('I have got') used as 'Yes! I've got it!'

have with this language. It is conflicting also due to its changing role from a vernacular for the majority of European Jewry to a marker of ultra-Orthodoxy. This emotional relationship with Yiddish has the effect that speakers often try to describe their connection to Yiddish or its role by distinctive concepts or even lexical elements from Yiddish, which I then used as in-vivo codes, e.g., nostalgia or *Stetl* ('little town'). However, to what extent and in what direction the respective ideology influences the speaker's choices depends on her or his individual biography and actual positioning. The two following quotes from my data are examples for the ideology that I labelled "Nostalgia for the *Stetl*" and they demonstrate very clearly how the same ideology can be interpreted in two directions:

- (4) ...und Jiddisch ist sowas, was ich so 'n bisschen mehr mit so Tradition und meiner Herkunft verbinde, weil ich weiß meine Großeltern und Eltern kamen ja auch aus Osteuropa, aus Rumänien und äh mein Vater ist auch mit Jiddisch aufgewachsen. Und dann weiß man irgendwie, wenn man diese Wörter benutzt, dann ist so irgendwie so dieses Alte, dieses Traditionelle, das so 'n bisschen weiterlebt. Das hat sowas ähm Nostalgisches und auch was so nja, vielleicht sowas Symbolisches, Bedeutungsvolles. Deswegen find ich das schon ganz schön dann solche Begriffe auch mal dann mitzunehmen.

... and Yiddish is something that is for me more connected with tradition and my origins, because I know that my parents and grandparents came from Eastern Europe, from Romania and um my father was raised with Yiddish. And then you know when you use these words, then in a way it's like something old, something traditional, lives on a little bit. There's something um nostalgic about it and also something, well maybe symbolic, meaningful. That is why I do like sometimes to pick up these terms. [Rebecca, 1:06:52]

For Rebecca, keeping up some Yiddish words and integrating them into German indexes a link to her past and triggers nostalgia. Petra, on the other hand, completely rejects the use of Yiddish as an index for backwardness. She, like others, uses the Yiddish noun *Stetl* ('little town') as a metonymy for the traditional and religious Jewish life in Eastern Europe. She considers this world vanished and it has no relevance for her kind of Jewishness. According to her, Hebrew, on the contrary, is an adequate expression also for a secular Jewish identity.

- (5) Ich würde sagen, dass ich mich im Hebräischen viel wohler fühle als im Jiddischen, weil ähm, ich glaube, dass das Jiddische eben dann aufgesetzt

wäre, dass man sich 'ne Welt aneignet, die es nicht mehr gibt. [...] Also jedenfalls nicht in den jüdischen Kreisen in denen ich mich bewege, ne?! Das ist keine *Stetl*-Welt und dementsprechend hat das Jiddische da auch keine Relevanz und keinen Platz. Das Hebräische aber schon, das moderne Hebräisch ist für mich ein Ausdruck des Jüdischseins, äh, der mir einfach entspricht, weil der eben auch, äh, säkular geht.

I would say, that I am much more comfortable with Hebrew than with Yiddish, cause hum, I think that Yiddish would then be artificial, that you try to take possession of a world that does not exist anymore. [...] At least not within the Jewish circles that I am moving in, see?! That is no *Stetl*-world and therefore, there is no place for Yiddish in it and Yiddish has no relevance in it. But Hebrew has, modern Hebrew is for me an expression of Jewishness, hum, that fits me, because it, um, it works also in a secular way. [Petra, 42:04]

The quote from Petra also demonstrates another ideology towards Yiddish that is very prominent across my data, namely authenticity (Jahns 2024: 167–174). Like other speakers, Petra seems to accept the use of Yiddish items mainly from an “authentic” speaker, who is a person that has been raised with Yiddish. Taking over Yiddish elements or learning the language later in life is perceived as inauthentic or, as Petra puts it, *aufgesetzt* ('artificial').

The ideologies towards Hebrew are generally more stable, both through time and across speakers, which shows the high prestige Hebrew has been enjoying by almost all speakers. Generally, speakers describe the relation towards Hebrew more through reasoning and do not use lexical items or emotions to do so. The next quote demonstrates the prestige of Hebrew as the language of Judaism and underlines its utility, while Yiddish is considered as less important and only interesting from a cultural or scientific perspective.

- (6) Also das Hebräische ist halt einfach die Sprache des Judentums. So. Und die Tora ist halt eben die Heilige Schrift auf Hebräisch. So. Und ja, und Jiddisch ist natürlich total schön und total interessant usw., aber es ist halt eigentlich, glaube ich mehr aus so 'ner ja, soziologischen Perspektive, kulturell-soziologischen Perspektive interessant zu untersuchen und linguistisch sicherlich auch, aber im, jetzt für die für des Judentum an sich, für den Bestand, sag ich mal, des Judentums ist es jetzt nicht so wichtig, glaub ich, Jiddisch zu können. Also da ist es viel wichtiger Hebräisch zu können, um das Judentum sozusagen aufrecht zu erhalten, weil das Hebräische nun mal einfach die, die Sprache des Judentums ist,

inhaltlich, und immer sein wird. Und Jiddisch, kann halt sein, dass es irgendwie in 50 Jahren keiner mehr kennt oder so, ja?

Well, Hebrew is just simply the language of Judaism. So. And the Torah is of course the holy book in Hebrew. So. And, yes, and Yiddish is of course really nice and really interesting and so on, but it is, however, I think, more from a sociological perspective, cultural-sociological perspective interesting to investigate and sure enough also linguistically, but in, now for Judaism as such, it is, I think, not so important to know Yiddish. Thus, it's more important to know Hebrew to maintain Judaism, so to speak, because Hebrew is just simply the, the language of Judaism, in terms of content, and always will be. And Yiddish, it might be, that nobody will know it in 50 years or so, yes? [Julia, 1:17:34]

This high prestige goes along with the ideology of prescriptivism, which means that it is considered important to use Hebrew words in the “correct” way. This includes the Hebrew stress pattern which differs from German and especially the plural morphemes. When I asked about the integration of Hebrew nouns like *Kippa* (‘skullcap’) and *Machane* (‘Jewish summer camp’), speakers unanimously agreed upon the fact that these nouns should be used with the Hebrew plural inflection, i.e., *Kippot* and *Machanot*, when integrated into German.

In the next quote the speaker describes that he would correct children at a Jewish summer camp who would add the German plural morpheme *-s* to the Hebrew noun *Madrichim* (‘Youth supervisors’) that is already in the Hebrew form (sg. *Madrich* – pl. *Madrichim*).

(7) Wenn 'nen Kind kommt und sagt ,Wo sind die Madrichims?', sag ich ,*Madrichim!*'.

If a child comes to me and says ‘Where are the Madrichims?’, I say ‘*Madrichim!*’. [Aaron, 37:05]

In quote (8) the interviewer explicitly asks whether the adding of a German plural morpheme *-s* would be a possible variant:

(8) Interviewer: Gibts auch jemand der Machanes sagt? Oder könnte man das sagen?

Julia: Also, der wäre sozusagen ziemlich außen als Idiot. Quasi, das ist dann schon total cool zu sagen, so *Machanot*. Da hat man ja Ahnung, wie der Plural gebildet wird im Hebräischen. [...] Jaa. Ne, also das, da würde man sofort korrigiert werden.

Interviewer: Are there people who say Machanes? Or is it possible to say it?

Julia: Well, this person would be rather outed as an idiot. I mean, it's pretty cool then to say, like *Machanot*. You obviously have a clue then how the plural is formed in Hebrew. [...] Yeah. No really, you would definitely be corrected. [21:50]

The quote shows that the prestige the language enjoys within the community is transmitted to the user of the language, or in this case, to the user of elements from the language. The transmission of prestige does, however, apply only if the element is used in the “correct” way. Speakers who can’t show their distinctively Jewish linguistic awareness are stigmatized as outsiders or newcomers.

## 5 Conclusion and applicability

In this study I have shown how Jewish speakers in Berlin make sense of the variation within the linguistic repertoire to which they have access. The variation consists mainly of speakers’ choices between lexical elements from Yiddish and Hebrew or between dialectal variants of Yiddish, but it encompasses also the quantity of items used. The linguistic-positioning task reveals speakers’ perspectives on their linguistic choices leading to categories that affect these choices.

In this study, the importance of knowing about, and differentiating between, the two main donor languages; intra-speaker variation due to context-awareness; and, most importantly, language ideologies towards Hebrew and Yiddish, were revealed as being the most important categories that affect speakers’ choices.

Hebrew elements are considered more neutral but prestigious variants that should be used in a “correct” way. Hebrew allows for the expression of Jewishness also for secular speakers. Yiddish indexes a vanished world that some have nostalgia for, while others reject it. Its use triggers strong emotions in both directions and is reserved for the “authentic” speaker.

However, other categories emerged as well from my data which would allow to answer slightly different questions using the same material, i.e., the transcribed interviews. I would like to give some examples of what these research questions or different perspectives on the data might look like.

First of all, it would be interesting to take a much closer look at individual elements and their use by the interviewees. A detailed statistical analysis of the distribution into the three piles and its comparison across speakers might lead to clusters of use or correlations between speakers sharing characteristics, like age,

L1, and religiosity. In addition, the findings from the three piles could be used to get insights into a core vocabulary that is used almost by everyone and thus is used for indexing alignment to the Jewish community as a whole instead of foregrounding subtleties of different kinds of Jewishness.

The task could also give insights into the salience of distinctive variants. In this study some items were not perceived by all speakers as being part of a distinctively Jewish repertoire. This was the case with elements that are cognates in Yiddish and German and therefore seem not to be salient when integrated into a German sentence. An example is *heimisch* ('homey') even though contexts of use seem to be slightly different in the two languages, or *Macher* ('active person'). It is no surprise that speakers with German as their L1 did not consider them part of the Jewish repertoire, while speakers with Russian as their L1 or who lived outside Germany for a longer time considered themja Yiddish and even expressed an especially positive emotion towards it.

Insights on ongoing language change concerning distinctive elements is another possible outcome of this task. Several items from the repertoire were either classified as known, but not used or not known, which could indicate that they are no longer in use, but have been overheard from the generation of parents or grandparents, e.g., Yiddish *Barches* ('braided bread eaten on Shabbat and holy days').

I have demonstrated that the linguistic-positioning task captures, from a speaker's perspective, justifications and explanations for lexical variation in the multilingual community of Jewish speakers in today's Berlin.

The task, however, is also applicable for multiple other sets of data and other linguistic areas. It is an effective tool to reveal reasons for linguistic choices, including ideologies and prestige of involved languages or varieties. It could also be used to find out about the salience of distinctive elements and whether they are used or understood as shibboleths for distinctive groups. This could, in consequence, give further insights into speakers' positioning in relation to these presumed groups, whether they wish to align to or want to distinguish from them. Moreover, the task is not restricted to lexical variation but can also be applied to pronunciation and syntactic patterns.

The task does not necessarily need to be applied to multilingual groups of speakers but can be used to research monolingual speakers and any kind of variation in their speech, e.g., dialectal variation, youth language, or language of distinct networks variation.

## References

- Auer, Peter. 2014. Anmerkungen zum Salienzbegriff in der Soziolinguistik. *Linguistik Online* 66(4). 7–20.
- Belkin, Dmitrij (ed.). 2017. *#Babel 21: Migration und jüdische Gemeinschaft* (Schriftenreihe des Ernst Ludwig Ehrlich Studienwerks 2). Berlin: Henrich und Henrich Verlag.
- Bogner, Alexander & Wolfgang Menz. 2005. Das theoriegenerierende Experteninterview: Erkenntnisinteresse, Wissensformen, Interaktion. In Alexander Bogner (ed.), *Das Experteninterview*, 2nd edn., 33–70. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Bunin Benor, Sarah. 2008. Towards a new understanding of Jewish language in the twenty-first century. *Religion Compass* 2(6). 1062–1080. DOI: 10.1111/j.1749-8171.2008.00108.x.
- Bunin Benor, Sarah. 2009. Do American Jews speak a “Jewish language”? A model of Jewish linguistic distinctiveness. *The Jewish Quarterly Review* 99(2). 230–269.
- Bunin Benor, Sarah. 2010. Ethnolinguistic repertoire: Shifting the analytic focus in language and ethnicity. *Journal of Sociolinguistics* 14(2). 159–183. DOI: 10.1111/j.1467-9841.2010.00440.x.
- Bunin Benor, Sarah & Steven M. Cohen. 2011. Talking Jewish: The “ethnic English” of American Jews. In Eli Lederhendler (ed.), *Ethnicity and beyond: Theories and dilemmas of Jewish group demarcation* (Studies in contemporary Jewry 25), 62–78. Oxford, New York: Oxford University Press.
- Bunin Benor, Sarah & Benjamin H. Harry. 2018. A research agenda for comparative Jewish linguistic studies. In Sarah Bunin Benor & Benjamin H. Harry (eds.), *Languages in Jewish communities, past and present* (Contributions to the Sociology of Language (CSL) 112), 672–694. Berlin, Boston: de Gruyter Mouton. DOI: 10.1515/9781501504631-026.
- Busch, Brigitta & Jürgen Spitzmüller. 2021. Indexical borders: The sociolinguistic scales of the shibboleth. *International Journal of the Sociology of Language* 272. 127–152.
- Charmaz, Kathy. 2010. *Constructing grounded theory: A practical guide through qualitative analysis*. Los Angeles, California: SAGE.
- Cramer, Jennifer. 2016. Perceptual dialectology. *Oxford Handbook Online* 1. 1–31. DOI: 10.1093/oxfordhb/9780199935345.013.60.
- Eckert, Penelope. 2008. Variation and the indexical field. *Journal of Sociolinguistics* 12(4). 453–476. DOI: 10.1111/j.1467-9841.2008.00374.x.

- Eckert, Penelope. 2012. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology* 41(1). 87–100.
- Fishman, Joshua A. 1981. The sociology of Jewish languages from the perspective of the general sociology of language: A preliminary formulation. *International Journal of the Sociology of Language* 30. 5–16. DOI: 10.1515/ijsl.1981.30.5.
- Flick, Uwe. 2007. *Qualitative Sozialforschung: Eine Einführung* (Rowohlt Enzyklopädie 55694). Reinbek bei Hamburg: Rowohlt Taschenbuch.
- Gal, Susan & Judith T. Irvine. 2019. *Signs of difference: Language and ideology in social life*. Cambridge: Cambridge University Press.
- Gold, David L. 1981. Jewish intralinguistics as a field of study. *International Journal of the Sociology of Language* 30. 31–46. DOI: 10.1515/ijsl.1981.30.31.
- Hary, Benjamin H. & Sarah Bunin Benor (eds.). 2018. *Languages in Jewish communities, past and present* (Contributions to the Sociology of Language/CSL 112). Berlin: de Gruyter Mouton.
- Hary, Benjamin H. & Martin J. Wein. 2013. Religiolinguistics: On Jewish-, Christian- and Muslim-defined languages. *International Journal of the Sociology of Language* 220. 85–108. DOI: 10.1515/ijsl-2013-0015.
- Irvine, Judith T. 1989. When talk isn't cheap: Language and political economy. *American Ethnologist* 16(2). 248–267.
- Irvine, Judith T. 2001. "Style" as distinctiveness: The culture and ideology of linguistic differentiation. In Penelope Eckert & John R. Rickford (eds.), *Style and sociolinguistic variation*, 21–43. Cambridge: Cambridge University Press.
- Jacobs, Neil G. 2005. *Yiddish: A linguistic introduction*. Cambridge: Cambridge University Press.
- Jahns, Esther. 2024. *Diglossic translanguaging: The multilingual repertoire of German-speaking Jews*. Berlin, Boston: de Gruyter Mouton. DOI: 10.1515/9783111322674.
- Jahns, Esther. 2022–present. *Judäo-deutsches Wörterbuch*. <https://jdw.jewish-languages.org>.
- Johnstone, Barbara, Jennifer Andrus & Andrew E. Danielson. 2006. Mobility, indexicality and the enregisterment of "Pittsburghese". *Journal of English Linguistics* 34(2). 77–104.
- Kahn, Lily & Aaron D. Rubin (eds.). 2016. *Handbook of Jewish languages* (Brill's handbooks in linguistics 2). Leiden, Boston: Brill.
- Klagsbrun Lebenswerd, Patric Joshua. 2016. Jewish Swedish. In Lily Kahn & Aaron D. Rubin (eds.), *Handbook of Jewish languages* (Brill's handbooks in linguistics 2), 618–629. Leiden, Boston: Brill.

- Koch, Peter & Wulf Oesterreicher. 1986. Sprache der Nähe – Sprache der Distanz: Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch* 36(1). 15–43. DOI: 10.1515/9783110244922.15.
- Kroskrity, Paul V. 2018. On recognizing persistence in the indigenous language ideologies of multilingualism in two native American communities. *Language & Communication* 62. 133–144.
- Lenz, Alexandra N. 2010. Zum Salienzbegriff und zum Nachweis salienter Merkmale. In Christina Ada Anders, Markus Hundt & Alexander Lasch (eds.), „*Perceptual Dialectology*“: *Neue Wege der Dialektologie*, 2nd edn. (Linguistik: Impulse & Tendenzen 38), 89–110. Berlin, New York: de Gruyter. DOI: 10.1515/9783110227529.1.89.
- Meuser, Michael & Ulrike Nagel. 1991. ExpertInneninterviews: Vielfach erprobt, wenig bedacht. In Detlef Garz & Klaus Kraimer (eds.), *Qualitativ-empirische Sozialforschung: Konzepte, Methoden, Analysen*, 441–471. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Myhill, John. 2004. *Language in Jewish society: Towards a new understanding*. Clevedon: Multilingual Matters.
- Peltz, Rakhmiel. 2010. Diasporic languages: The Jewish world. In Joshua A. Fishman & Ofelia García (eds.), *Handbook of language and ethnic identity*, 2nd edn., vol. 1, 135–152. Oxford, New York: Oxford University Press.
- Preston, Dennis R. 1999. *Handbook of perceptual dialectology*, vol. 1. Amsterdam: John Benjamins.
- Preston, Dennis R. 2010. Perceptual dialectology in the 21st century. In Christina Ada Anders, Markus Hundt & Alexander Lasch (eds.), *Perceptual dialectology: Neue Wege der Dialektologie*, 2nd edn. (Linguistik: Impulse & Tendenzen 38), 1–29. Berlin: de Gruyter.
- Rabin, Chaim. 1981. What constitutes a Jewish language? *International Journal of the Sociology of Language* 30. 19–28. DOI: 10.1515/ijsl.1981.30.19.
- Silverstein, Michael. 1979. Language structure and linguistic ideology. In Paul R. Clyne, William F. Hanks & Carol L. Hofbauer (eds.), *The elements: A parasession on linguistic units and levels*, 193–247. Chicago: Chicago Linguistic Society.
- Spitzmüller, Jürgen. 2013. Metapragmatik, Indexikalität, soziale Registrierung. zur diskursiven Konstruktion sprachideologischer Positionen. *Zeitschrift für Diskursforschung* 1(3). 263–287.
- Spolsky, Bernard & Sarah Bunin Benor. 2006. Jewish languages. In Keith Brown (ed.), *Encyclopedia of language & linguistics*, 2nd edn., vol. 6, 120–124. Amsterdam: Elsevier.

- Weinberg, Werner. 1973. *Die Reste des Jüdischdeutschen*. 2nd edn. (Studia Delitzschiana 12). Stuttgart: Kohlhammer.
- Weinberg, Werner. 1994. *Lexikon zum religiösen Wortschatz und Brauchtum der deutschen Juden*. Stuttgart-Bad Cannstatt: Frommann-Holzboog.
- Weinreich, Max. 2008. *History of the Yiddish language*. New Haven, Connecticut: Yale University Press.
- Wexler, Paul. 1981. Jewish interlinguistics: Facts and conceptual framework. *Language* 57(1). 99–149. DOI: 10.2307/414288.
- Wiese, Heike. 2020. Language situations: A method for capturing variation within speakers' repertoires. In Yoshiyuki Asahi (ed.), *Proceedings of Methods XVI: Papers from the Sixteenth International Conference on Methods in Dialectology, 2017* (Bamberger Beiträge zur Englischen Sprachwissenschaft / Bamberg Studies in English Linguistics 59), 105–120. Berlin: Peter Lang.

# Chapter 11

## Regional prosodic variation in the speech of young urban Russians: Quantitative vowel reduction in Moscow and Perm

Margje Post

University of Bergen

Most young urban Russians speak with little or no local characteristics in their speech, but small regional differences in segmental and prosodic properties are likely to be present. In order to find out whether regional prosodic differences persist in modern urban speech, we compared durational vowel reduction in the speech of adolescents from the capital Moscow and from Perm (Ural). In contemporary Central Standard Russian, non-high vowels in first pretonic position are much less reduced than other unstressed vowels. This reduction in two degrees is strong in Central Russia, but the difference between the degrees appears to be smaller in other parts of the country. Our study shows a large and stable difference in reduction patterns between the two cities, for both male and female speakers, even though it is based on read speech, which tends to be closer to standard language, with less local traits, than other speaking styles.

### 1 Introduction

#### 1.1 Regional variation in Russian

Russian is a language with little geographically based variation, compared to other languages, and the Russian-speaking community has a strong standard language ideology. The dialectal differences have never been large, and, with



the strong centralization tendencies in Russia and expectations to speak proper, standard Russian, most young urban Russians speak with little or no local features in their speech. It is often assumed that educated Russians have a standard pronunciation without local characteristics. However, most researchers acknowledge the existence of locally coloured varieties of Standard Russian (Panov 1967, Krysin 2007, Andrews 2006, Krause 2010, Grammatčikova et al. 2013).<sup>1</sup> The local features in regionally coloured standard speech are mainly restricted to lexical items and minor differences in pronunciation.<sup>2</sup>

When so little variation is heard in public space, it is no surprise that today's youth in Russia, according to a recent folk linguistic study, has restricted knowledge of regional variation in Russian (Vardøy 2021). Grammatčikova et al. (2013: 72) claim nevertheless that Russians can often, after hearing the first word, distinguish speakers from different cities. One can expect, along with Grammatčikova and her colleagues, that much of the ability to recognise a speaker's regional provenance can be accounted for by regional differences in prosody (Grammatčikova et al. 2013, Post 2017). Erofeeva et al. (2000) get the impression that prosodic characteristics are the most important cue to the local colouring of speech from the city of Perm, despite its numerous local characteristics on the segmental level (Erofeeva 2005). However, little is known about the phonetics of regional varieties of Standard Russian, let alone about their prosody. As remarked by Kalenčuk (2021), sociolinguistic studies of regional variants of standard language pronunciation have hardly been conducted. Linguists usually describe only the Moscow and Petersburg norm, and codify only the Moscow norm, based on the idea formulated by Vasilij Černyšev back in 1915 that "educated people in all parts of Russia speak Moscow Russian" (Kalenčuk 2021: 11). Following Iosad (2012: 522), I will call this Central Russian norm *contemporary Central Standard Russian* (CSR).

---

<sup>1</sup>Cyrillic script is latinised following Comrie & Corbett's (1993) transliteration system, apart from toponyms, where the usual English forms are used (Moscow, Perm, Chelyabinsk).

<sup>2</sup>In Auer's (2005) standard-dialect constellations in European languages, Russian is classified among the languages with so-called *diaglossia*, which have a continuum between rural dialects and the standard language (cf. Krause 2010), but Russian is at best a poor example of *diaglossia*. Little is known about the intermediate space between base dialects and the standard. The traditional rural dialects are under strong pressure, and, although the term *regiolect* is in use for Russian (see discussion in Krause 2010), stable intermediate regional varieties appear to be absent or rare.

## 1.2 Vowel reduction and prosodic word structure in Russian

### 1.2.1 Central Standard Russian: Strong nucleus and reduction in two degrees

One of the areas with known regional variation is in the expression of unstressed vowels. Russian has mobile, distinctive stress – word stress can occur on any syllable of the word; cf. *zámok* ‘castle’ vs. *zamók* ‘lock’; *kólokol* NOM.SG ‘bell’ vs. *kolokolá* NOM.PL ‘bells’. Vowel length is not phonologically contrastive. Instead, it plays an important role in the expression and perception of stress (Bondarko 1998: 219). Unstressed vowels are subject to reduction. Most notably, unstressed /o/ merges with /a/, so the words *travá* [tré'va] ‘grass, herb’ and *drová* [dré'va] ‘firewood’ are pronounced with the same vowels. In addition, Central Standard Russian (CSR) has a typologically unusual word prosodic pattern with a heavy nucleus and weak periphery: the first pretonic syllable – i.e., the syllable immediately preceding the stressed syllable – is uncommonly prominent, especially when it contains an open vowel; cf. the [e] in *travá* and *drová*. This syllable forms a salient contrast, together with the stressed syllable, with unstressed syllables in other, weak positions, which are heavily reduced, both in quality and quantity (Zlatoustova 1981, Kodzasov 1999). In addition, the first pretonic syllable is often singled out by a local high tone (Kasatkina 2005). One could argue that word stress is realised over two syllables, as Dubina (2012: 175) claims about Belarusian. This uncommon prominence of the first pretonic vowel leads to vowel reduction in two degrees, called *moderate* and *radical* reduction by Crosswhite (2000), with moderate, degree 1 reduction occurring in the first pretonic syllable and strong, degree 2 reduction elsewhere.<sup>3</sup> In CSR, the word *molokó* ‘milk’ is typically pronounced [mélə'ko], with a short *schwa* in the second pretonic syllable (= radical reduction), and a long, open vowel in first pretonic position (= moderate reduction). Following Potebnja (1866), we can describe the distribution of syllable prominence in the word in CSR by the formula 112'311, where 1 means radical reduction, 2 means moderate reduction and '3 stands for no reduction in the stressed syllable. Empirical studies of CSR have confirmed that Potebnja's 112'311 formula corresponds to a three-way distinction in duration (e.g., Bondarko et al. 1966, Zlatoustova 1981, Kuznecov 1997, Barnes 2006, Knjazev 2006). The durational differences between the syllables relative to stress are highest for non-high vowels (Zlatoustova 1981, Bondarko et al. 1966), i.e., for unstressed /a, o/ after non-palatalised consonants. Only for these vowels a qualitative distinction between

<sup>3</sup>Moderate, degree 1 reduction is also found in other positions that allow long vowel durations, notably, in onsetless syllables and, optionally, in phrase-final open syllables (Barnes 2006, Iosad 2012: 534). Kuznecov (1997) discerns an additional, third degree of quantitative reduction for posttonic vowels, based on the statistics of his phonetic study.

the two positions into two different allophones is discerned. This two-degree qualitative reduction of unstressed /a, o/ into [ɐ] and [ə] is incorporated into the prescriptive pronunciation standard (Avanesov 1984).<sup>4</sup> It is also an integral part of Russian L2 course books.<sup>5</sup> The two-degree reduction in duration, however, is not mentioned.

Due to this typologically uncommon reduction pattern, vowel reduction in Russian is a well-known and much studied area in phonetics and phonology.<sup>6</sup> Most literature is based on, and accounts for, Central Standard Russian speech. Although two-degree reduction in duration is claimed to be a common feature of East Slavic (Dubina 2012), the difference between degree 1 and degree 2 varies. The next sections will discuss the durational reduction patterns in other varieties of Russian.

### 1.2.2 Traditional Russian rural dialects

As to traditional rural dialects, the first pretonic prominence – and the related two-degree reduction – varies from very prominent in some Central Russian dialect groups (in which the rhythmic structure can be described as 113'311 according to Potebnja's formulae, e.g., group V in Figure 1), to (almost) absent in parts of the Russian North (111'211; group VIII in Figure 1). Grammatčikova et al. (2013) state that some dialects do not have a prosodic nucleus of the first pretonic and tonic syllable at all, but Vysotskij (1973) claims that all traditional Russian dialects have it to some degree (1973: 34,36).<sup>7</sup> Vysotskij (1973) measured the rhythmic word structure in a large number of Russian rural dialects by examining consonant and vowel durations. He distinguishes 15 varieties of Russian, both rural dialects and varieties of Russian spoken in Moscow, which all have a different

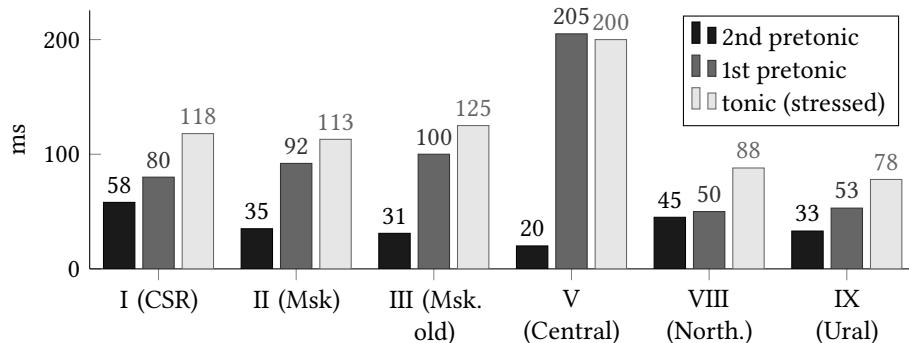
<sup>4</sup>Most older literature uses the symbol [ʌ] for the first degree reduction of /a, o/ after non-palatalised consonants; cf. Iosad (2012) for a discussion.

<sup>5</sup>This is not restricted to books at the university level. An example of a Russian L2 course book for secondary schools where two-degree reduction is taught, is Hertz et al. (2005).

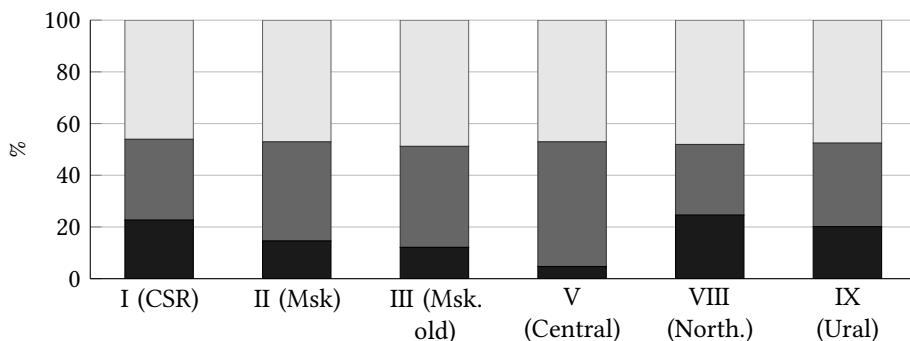
<sup>6</sup>Examples of empirical phonetic research on vowel reduction in CSR are Zlatoustova (1981), Bondarko et al. (1966), Kuznecov (1997), Padgett & Tabain (2005), Barnes (2006), Kocharov et al. (2015). Examples of phonological accounts are Crosswhite (2000), Iosad (2012), Mołczanow (2015).

<sup>7</sup>The fact that Vysotskij found a durational difference between the second and first pretonic vowels in all dialects does not necessarily entail that they all have two-degree reduction, at least not at a phonological level. Besides, some of the recorded speakers might have accommodated their speech to the interview situation and have spoken with more two-degree reduction than they would have done when speaking with locals. Paufošima (1978) suspects some of her dialect speakers of trying to copy the word rhythm of perceived Standard Russian, with some being more successful at this than others.

word rhythm pattern. The durations in a selection of these groups are given in Figure 1.



(a) Mean absolute vowel durations (from Vysotskij 1973)



(b) Mean vowel durations relative to the total vowel duration in the word (from Vysotskij 1973)

Figure 1: Mean vowel durations in second pretonic, first pretonic and tonic (stressed) position, in Standard Russian pronunciation (CSR, group I), in Moscow vernacular speech (group II), in the old Moscow norm (III) and in three dialect groups (V, VIII and IX), from Vysotskij (1973: 38), partly also in Bethin (2006: 131). The vowels are /o, a/ after non-palatalised plosives and fricatives in words with the structure CV-CV-CVC, in utterance-final and utterance-medial position. Vysotskij measured 100–150 word tokens per speaker (1973: 33), but it is not clear if more than one speaker per group was analysed.

The most extreme groups are group V and group VIII. In group V, found in the Vladimir-Volga basin in Central Russia (discussed by Bethin 2006: 113), the durations of the second and first pretonic vowels are extremely different. Group VIII, on the other hand, situated in the North of European Russia, shows hardly any difference between the pretonic vowels. The remaining 13 groups in Vysotskij's study have intermediate values, including Standard Russian, which is Vysotskij's

type I (Figure 1, first group). Type II represents the Moscow vernacular, with a larger two-degree difference than in CSR (group I). Group III shows the old Moscow norm, with an even larger difference, closer to the rural Central Russian dialects. Finally, group XI is included in Figure 1 because it represents the Ural region, where we recorded part of our own data (see Section 1.3 below).

In a large area in Southwestern Russia, and in a neighbouring region across the border in Belarus and Ukraine, the vowel reduction pattern is complicated by vowel dissimilation. In these dialects, the quality and duration of the first pretonic vowel, and in some cases its tone, depend on the quality of the stressed syllable, for the two neighbouring vowels must be different in both quality and quantity.<sup>8</sup>

### 1.2.3 Modern urban Russian

Most speakers of Russian today do not speak a traditional rural dialect, however. A handful of studies have addressed ongoing phonological changes in rural areas of Russia, among them Paufošima (1978) on a Northern dialect and Kochetov (2006) on rural speech from the Perm area. As to modern urban Russian, there might still be differences in relative vowel durations between speakers from Moscow and St. Petersburg, although most differences between Moscow and St. Petersburg speech that have been reported in older literature have disappeared, as Verbickaja (1977) found out in a phonetic study of the speech of 150 speakers from St. Petersburg (Leningrad) and 50 speakers from Moscow – at least between educated speakers in formal settings. In Verbickaja's study, the two-degree reduction is still somewhat stronger in Moscow than in Saint Petersburg standard speech, due to its relatively shorter vowels in second pretonic and posttonic position (Verbickaja 1977, see also Nikolaeva 1977, Kuznecov 1997).

Padgett & Tabain (2005) found substantial interspeaker variation in the expression of two-degree reduction among speakers of Standard Russian, which might be due to differences in geographical provenance. They recorded speakers

<sup>8</sup>In dialects with vowel dissimilation, a stressed low vowel [a] is never preceded by a low vowel, but by a mid vowel *schwa*, whereas high vowels are preceded by a low vowel. The low vowels are substantially longer than the high vowels, leading to a length trade-off between first pretonic and tonic vowel (see Čekmonas 2001, Al'muxamedova 1985, Kasatkina 2005, Bethin 2006, Savinov 2013, D'jačenko 2015 for studies of dialects with dissimilative vowel systems that take into account their duration). Iosad (2012) suggests that all vowel dissimilation is first and foremost a dissimilation in length. Speakers from Moscow have no qualitative dissimilation, but clear quantitative dissimilation: a length trade-off between first pretonic and tonic vowel, for the first pretonic low vowels have by far the longest duration when preceding short, stressed high vowels (Kasatkina 2005, Iosad 2012).

of Standard Russian living in Australia with various geographical provenance, tacitly assuming that people who are assessed to speak standard Russian, have the same pronunciation. Their participants from Moscow, Saint Petersburg and Kiev had larger, and clearly categorical, differences in duration than their other participants, who were born in China or had a mixed geographical background (Padgett & Tabain 2005: 42, Fig. 4). However, a speaker of Russian does not have to grow up in Central Russia to express strong two-degree reduction: The difference between the vowels was very large for Barnes' participant from Ufa, the capital of Bashkortostan (Barnes 2006).

Two preliminary studies have measured unstressed vowel durations in large cities other than Moscow and Saint Petersburg (Grammatčikova et al. 2013, Erofeeva 2005). Both studies suggest that the geographical opposition between centre and periphery is retained in modern urban Russian – with strong two-degree reduction in Central Russia, but a lesser difference between the two degrees in non-central areas of the country.

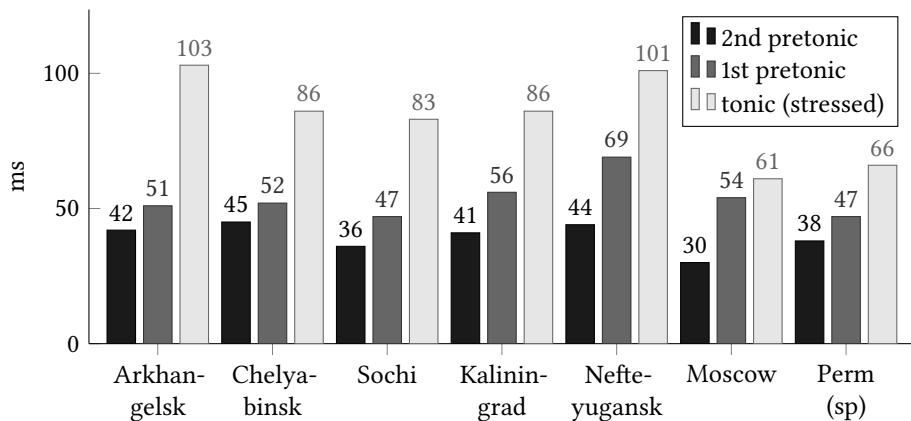
Grammatčikova et al. (2013) measured vowel durations in trisyllabic words with a  $C_{a-2}C\ a_{-1}Cá_0C$  structure in a text, read by students from different parts of Russia, and compared them to the pronunciation by a middle-aged Moscow speaker of Standard Russian. In the speech of the Moscow speaker, the first pretonic vowel was not uncommonly long in absolute terms, but it had the longest duration relative to the second pretonic and tonic vowels (Figure 2, page 248). The 5 students from other cities had smaller relative differences. One should take into account that these data are preliminary, based on a small number of vowel tokens by only 6 speakers.

The small-scale study by Erofeeva (2005), of relative vowel duration of unstressed /a/ and /o/ in spontaneous speech by four speakers from Perm with an audible local accent, showed hardly any reduction in two degrees: The vowels in first pretonic position were only slightly longer than those in second pretonic position (Erofeeva 2005, data from Appendix 1). In these data, /a, o/ have mean durations of 38.2 : 46.8 : 65.7 ms (for  $a_{-2} : a_{-1} : á_0$ ;  $n = 308$ , cf. Figure 2).

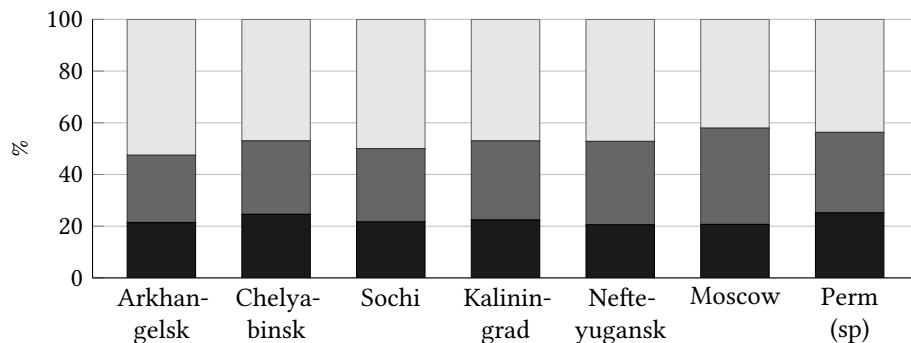
Since these two studies of regional Russian analysed only a few speakers, their data need to be confirmed using larger data sets.

#### 1.2.4 Sociolinguistic variables

Previous literature suggests that a lack of reduction in two degrees is not a stigmatised dialectal feature. Bondarko (1998) considers the pronunciation of /o/ as unreduced [o] in unstressed position as *dialectal* (and as violating the *orthoepic*, phonological rules of standard Russian), whereas pronunciation of *golová* as



(a) Mean vowel durations in CVCV'CV(CV) words (from Grammatčikova et al. 2013; Perm (sp) from Erofeeva 2005)



(b) Vowel durations relative to the total vowel duration in CVCV'CV(CV) words (from Grammatčikova et al. 2013; Perm (sp) from Erofeeva 2005)

Figure 2: Durations of /o, a/ after non-palatalised plosives and fricatives in words with the structure CV-CV'-CVC(CV), in read speech by students from 5 cities and one adult from Moscow (from Grammatčikova et al. 2013) and in spontaneous speech (sp) in Perm (from Appendix 1 to Erofeeva 2005).

[gʌlʌ'va] instead of normative [gəlʌ'va] – i.e. with reduction and neutralisation, but without differentiating between two qualitative degrees of reduction – is considered as pronunciation with an *accent*, violating only *orthophonic* rules, which concern the phonetic realisation of sounds (Bondarko 1998: 249). Speaking with dialectal features has low social status, for in Russian, the term *dialects* and *dialectal* refer to traditional rural dialects, which are associated with the speech of poor, elderly villagers with little education, living in backward regions. So, avoidance of unreduced unstressed /o/ in formal speech has higher priority than to distinguish two degrees. Furthermore, the literature only mentions the status of two-degree reduction in quality, not in quantity, suggesting that lack of a distinction in duration is even less stigmatised. The low socio-indexical status of unreduced unstressed [o] has been confirmed in attitudinal studies (e.g., in Andrews 1995). Empirical work on vowel reduction in Perm speech shows that unreduced [o] is indeed avoided in formal speaking styles, but much less so in spontaneous speech. Erofeeva (1993) found a frequency of 64% of unreduced [o] in spontaneous speech, but only 4% in read speech (Erofeeva 1993). The same low number – 4% – was also found in speech read by people from Perm in Verbickaja et al. (1984).

Virtually no empirical studies of unstressed vowel duration appear to have been done that take sociolinguistic variables into account, but the literature suggests that the first pretonic prominence varies not only with the region, but also with age, gender, level of education and with speaking style:

- *age*: The extra long and open first pretonic vowels in Moscow speech were common in older Moscow pronunciation (Lapteva 1999: 354; cf. Vysotskij 1973, group III in Figure 1).
- *educational level*: They are also associated more generally with vernacular Moscow speech (Vysotskij 1973, group II; Rozanova 1988, *inter alia*).
- *gender*: Women are claimed to use extra long and prominent (in tone) first pretonic open vowels to a larger degree than men, specifically as a means for expressing emphasis, where men would typically prefer longer consonants (Zemskaja et al. 1987, Rozanova 1988, Kasatkina 2003, all based on speakers from Moscow), though Kasatkina 2005 found an equal amount of prominent first pretonic vowels among women and men in her small empirical study. In other languages, women tend to have longer vowel durations than men and in many studies, they also speak at a slower rate (e.g. Simpson & Ericsdotter 2003). Urban women also tend to use standard and other prestige forms more often than men (Labov 2001).

- *speaking style*: We can expect regional differences to be lowest in a formal speaking style, such as in the reading tasks most experimental studies have been based on. In formal speech, people are more likely to adapt to standard language and suppress low-status socio-indexical features; cf. the differences in vowel quality relative to speaking style in Erofeeva (1993).

### 1.3 Research questions: Moscow vs. Perm

Our overarching research question was whether we will find regional variation in prosody in today's Russian urban speech. We chose to compare speakers from two large cities, Moscow and Perm. Moscow is Russia's capital, and Moscow's speech, especially vernacular speech, is known for long and open realisations of /o/ and /a/ in first pretonic position, as in [ma:'skva]. The city of Perm is situated in the Ural region on the border between European Russia and Siberia. Its speech is known for a relatively strong local accent with northern Russian traits. It is well-studied by the sociolinguists of Perm University, using probabilistic measures to study the social stratification of dialect features across speaker groups in Perm (Erofeeva 1995, Erofeeva 2005, *inter alia*).

More specifically, I wanted to answer the following questions:

1. Two-degree reduction: How is the durational two-degree reduction expressed in the two cities? In other words, how different in duration are the two pretonic vowels from each other, and from the tonic (stressed) vowel? We expect the first pretonic vowel to be relatively longer in Moscow speech.

This is our major question, but we also addressed possible gender effects:

2. Gender effect on overall vowel duration: Do the female speakers have longer vowel durations than the male speakers, in both cities?
3. Gender effect on duration of the first pretonic vowel: Do the girls use extra prominence on the first pretonic vowels, measured in duration, more often than the boys?
4. Gender effect on relative distance in vowel duration between the two cities: Do the Perm girls have a reduction pattern closer to that of the Moscow speakers, with stronger two-degree reduction, than the boys? If it is true that girls tend to converge more to standard language – or to another non-local norm with high status – also in Russian, then we can expect to see smaller differences between the female speakers than between their male peers from Moscow and Perm.

## 2 Data and methods

### 2.1 Materials

We asked participants in both cities to read aloud an utterance list containing trisyllabic words with final stress. The three target words in the reading task were /pota'kat<sup>j</sup>/ 'to connive', /topo'tat<sup>j</sup>/ 'to stamp feet' and /poko'pat<sup>j</sup>/ 'to dig a little', three words that were also analysed in Vysotskij (1973). They contain /a/ and /o/ after the non-palatalised voiceless plosives /p, t, k/, which facilitate vowel segmentation, and have the rhythmic structure CV-CV'-CVC. I will call the vowels in second pretonic position a<sub>-2</sub>, the first pretonics a<sub>-1</sub> and the tonic (stressed) vowels a<sub>0</sub>. The symbols a<sub>-2</sub> and a<sub>-1</sub> stand for the phonemes /a/ and /o/, which merge in unstressed position in Standard Russian.

We analysed 6 sentences, covering each word in two prosodic conditions: in utterance-medial position carrying the nuclear (final) accent, represented by (1) below, and in utterance-medial prenuclear position (2):<sup>9</sup>

- (1) utterance-medial, nuclear position:  
*Ja pokopát' pošla.*  
 I.NOM dig.INF go-F.SG.PST  
 'I went digging.'
- (2) utterance-medial, prenuclear position:  
*Ja topotát' uže ne budu.*  
 I.NOM patter.INF already NEG 1SG.IPFV.FUT  
 'I won't patter anymore.'

We left out utterance-initial and utterance-final positions in order to avoid boundary phenomena, which typically affect duration – utterance-initial strengthening and final lengthening.

---

<sup>9</sup>We recorded a larger number of utterances and more speakers – 10 utterances by 33 speakers, but we had gaps in the data due to misreadings, some clear cases of unnatural speech, long hesitations or pauses, or creaky voice. For the Moscow speakers, discarded renderings could be replaced by their second renderings. This could not be done for the Perm speakers, who read the utterances only once. We also discarded all data from one male speaker from Perm, because of disfluent speech, and from one female speaker from Perm, because of creaky voice. In order to avoid gaps in the data, we excluded several more speakers and the first renderings of the words (in citation form), which were sometimes read with hesitation, because some of the speakers might not have been familiar with the uncommon target words *topotat'* and *potakat'*. This left us with 26 speakers reading 6 utterances each (Table 1).

The reading task was performed in 2015 by 15–17-year-old pupils at a school in Perm and a school in Moscow, who were born in these cities or had moved there before the age of 6. Almost all speakers have parents with higher education. We have not screened our participants and assessed to which degree they speak with a local accent. Therefore, one cannot equate the speech of our Moscow participants with Central Standard Russian, but since the recordings are made in a formal setting and most pupils in both cities have parents with higher education, a strong degree of local vernacular accent is unlikely.

Six utterances each by 13 female and 13 male speakers were analysed for the present study (Table 1). The same speakers performed a range of other tasks (reading, semi-spontaneous and spontaneous speech and interviews), collected for, or in association with, Benedikte Vardøys PhD project on young Russians' perception of regional variation in Russian (cf. Vardøy 2021, Post & Andreeva 2023). The sound recordings were made in the school library or in a small room, using digital recorders and unidirectional head-mounted microphones (a Zoom H5 recorder with a Shure WH20 microphone in Perm, a Zoom H2 with a Samson QV mic in Moscow, set at 44.1 kHz, 16-bit, .wav). The participants read the utterances from paper, in the same order, a single time in Perm, two times in Moscow.

Table 1: Number of speakers and tokens analysed for the present study

	speakers	F	M	word tokens	vowel tokens
Moscow	13	7	6	78	234
Perm	13	6	7	78	234
Total	26	13	13	156	468

## 2.2 Acoustic and statistical analyses

The target vowels in the speech samples were manually segmented in Praat (Boersma & Weenink 1992–2024) via visual inspection of the waveform and spectrogram according to standard segmentation criteria; for examples, see Figure 3 and Figure 4. The durations of the target vowels per speaker and location were extracted using Praat scripts.

For statistical validation, we used the software JMP 16.2.0. Vowel durations were log-transformed because of positive skewness. As a first step towards determining the differences, linear mixed models (LMM) were fitted with the respective log-transformed measure as dependent variable and DISTANCE-TO-STRESS

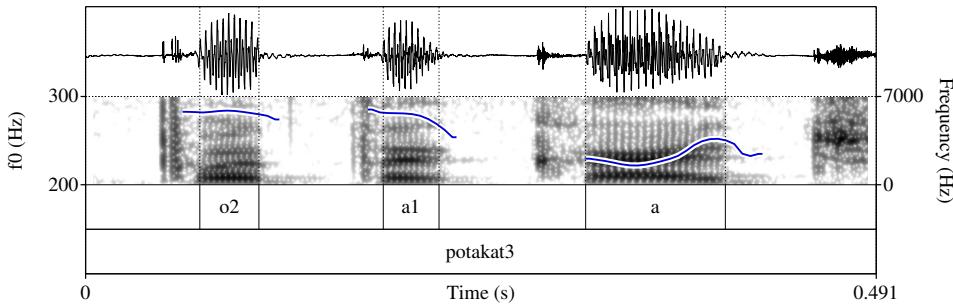


Figure 3: Waveform, spectrogram and F0 contour, with segmented vowels, of the target word *potakát'*, produced by a female speaker from Perm (ID 390F), made in *Praat* Boersma & Weenink 1992–2024

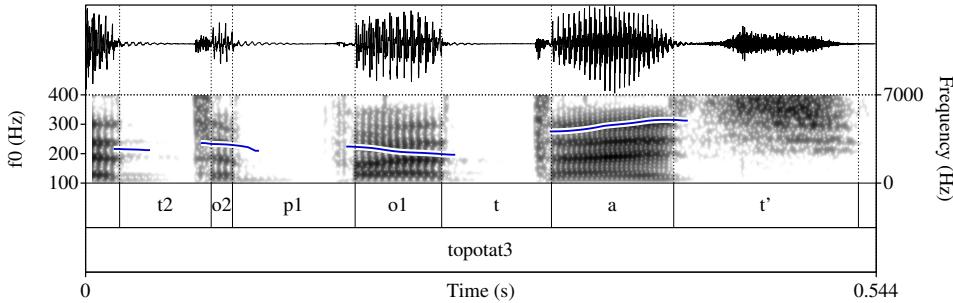


Figure 4: Waveform, spectrogram and F0 contour, with segmented vowels and consonants, of the target word *pokopát'* in utterance (2), produced by a female speaker from Moscow (ID 10F)

with three factor levels (0/1/2), LOCATION with two factor levels (Moscow/Perm) and GENDER with two factor levels (male/female) as fixed factors, as well as all their possible interactions. SPEAKER, WORD (*topotát'*/*pokopát'*/*pokatát'*) and POSITION-IN-UTTERANCE (utterance-medial prenuclear/utterance-medial nuclear accent) were taken as random factors. Separate post-hoc tests were carried out per variable, if appropriate. The confidence level was set at  $\alpha = 0.05$ .

### 3 Results

Predictably, we found a main effect of GENDER ( $F [1, 22] = 4.1169, p < 0.05$ ) and of DISTANCE-TO-STRESS ( $F [2, 431] = 815.4856, p < 0.001$ ) on the target vowel duration, with female speakers having significantly longer durations (cf. Figure 5, female voices in blue), and stressed vowels ( $\acute{a}_0$ ) being longer than the vowels in the first pretonic syllable ( $a_{-1}$ ), which in turn are longer than the vowels in the

second pretonic syllable ( $a_{-2}$ ), when taking both cities together (Post forthcoming). The variable LOCATION ( $F [1, 22] = 4.1169, p = 0.055$ ) was not significant.

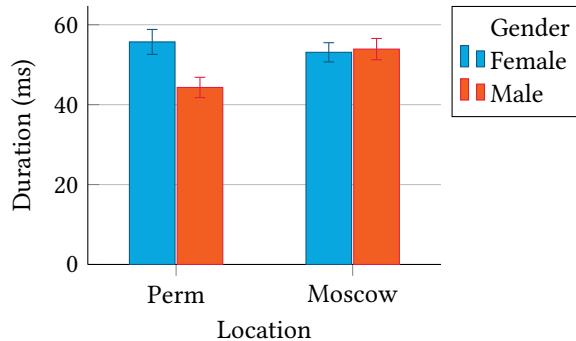


Figure 5: LOCATION vs. GENDER: Mean duration of vowels (all three positions together), for Perm (left) and Moscow (right), with female speakers in blue and male speakers in red. Error bars represent standard errors.

The statistical analysis also revealed a significant interaction between LOCATION and GENDER ( $F [1, 22] = 8.4120, p < 0.01$ ). Post-hoc tests revealed that the girls from Perm have significantly longer vowels than their male classmates, but in Moscow, the girls and boys have similar vowel durations (see Figure 5). This means that the earlier mentioned gender effect is only found in Perm.

More relevant for our main research question is the highly significant interaction between LOCATION and DISTANCE-TO-STRESS ( $F [2, 431] = 65.0217, p < 0.001$ ). In the realisations of the speakers from Moscow, vowel durations become significantly shorter with increasing distance from the stressed syllable, whereas in the realisations of the speakers from Perm, the vowels in the stressed syllable are significantly longer than the vowels in both first and second pretonic syllable (Figure 6). When comparing the mean values of actual durations according to positions to stress in Figure 6, we see that in Moscow, the vowels in the first pretonic position are, on average, twice as long as the vowels in the second pretonics, for both male and female speakers. In Moscow, the relative durations of the three consecutive vowels ( $a_{-2} : a_{-1} : á_0$ ) are almost 1 : 2 : 3 (with actual values 24.1 ms : 49.5 ms : 82.8 ms). In Perm, the difference between the two prestressed vowels is surprisingly small, the ratio being almost 1 : 1 : 3 (with the actual values 28.3 ms : 33.1 ms : 89.3 ms). The letter report from the post-hoc pairwise Tukey HSD test of the mean durations (log-transformed; Table 2) shows that the small difference in duration between the pretonic vowels in Perm is not statistically

significant, since  $a_{-2}$  and  $a_{-1}$  in Perm share the letter *C*, whereas  $a_{-2}$  and  $a_{-1}$  in Moscow end up with the letters *B* and *D*, respectively (Table 2).

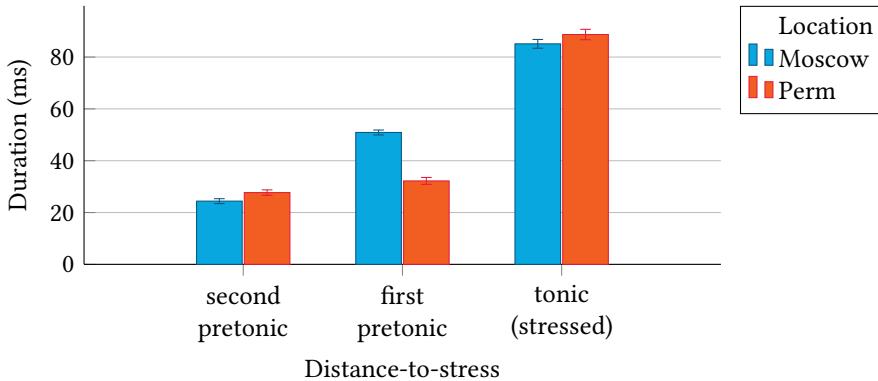


Figure 6: LOCATION VS. DISTANCE-TO-STRESS: Mean duration according to distance-to-stress and location (blue = Moscow; red = Perm; both genders). Error bars represent standard errors.

Table 2: Connected Letter report from pairwise Tukey HSD test of mean durations (log-transformed), according to location and distance-to-stress. Levels not connected by the same letter are significantly different.

Level	Least Sq Mean	
P, tonic	A	4.4715085
M, tonic	A	4.4284267
M, 1st pretonic	B	3.9156951
P, 1st pretonic	C	3.4001192
P, 2nd pretonic	C	3.2752833
M, 2nd pretonic	D	3.1268467

We saw that the female speakers use longer vowels than the men (in Perm), but the relative durations are not affected by gender: There was no significant interaction between GENDER and DISTANCE-TO-STRESS ( $F [2, 431] = 1.3204, p > 0.05$ ), nor between all three variables LOCATION, DISTANCE-TO-STRESS and GENDER together ( $F [2, 431] = 1.4552, p > 0.05$ ). The girls use the same vowel rhythm as the boys in each city: In Moscow, the relative durations of  $a_{-2} : a_{-1} : \acute{a}_0$  are 15 : 32 : 53% (of the total vowel duration in the word) for the girls and 15 : 31 : 53% for the

boys; in Perm, the numbers are 19 : 22 : 58% for the girls and 18 : 21 : 61% for the boys.

One must take into account that these numbers are mean values and that the variation between individual tokens is large. However, even with this large degree of variation, the difference between Moscow and Perm is highly significant.

When comparing the mean durational values of each individual speaker, one can see that all 13 speakers in Moscow had a larger durational difference between the two prestressed vowels than all 13 speakers in Perm. This result suggests that not a single speaker in Perm uses a Moscow word rhythm, or the other way round.

When we look even closer at all individual word tokens in Moscow, the second pretonic vowel is shorter than the first pretonic in all but a single case – in 77 out of 78 word tokens. In Perm, it is shorter in only 51 out of 78 renditions. This suggests that the two-degree reduction is very stable in Moscow.

## 4 Discussion

### 4.1 Large difference between pretonic vowels in Moscow, not in Perm

Our main research question concerns the expression of reduction in two degrees, that is, the relative durations of the second and first pretonic vowels (research question 1). As expected, we found a larger difference in duration between the two pretonic positions in Moscow than in Perm. This correlation between location and distance to stress is stronger than expected. The difference between the cities is not only large, but also stable, since all Moscow speakers have a rhythm different from all Perm speakers.

In Moscow, the distinction between the two vowels is very robust: not only is  $a_{-1}$  on average twice as long as  $a_{-2}$  (Figure 6), it is also longer in all but one word production in Moscow. Mark that  $a_{-1}$  would have been even longer in the position preceding a high vowel, since Moscow has clear vowel dissimilation in duration (Kasatkina 2005, cf. Section 1.2.3 above). Obviously, the reduction in two degrees is categorical for the Moscow speakers. A study of the formant values of our data shows that the large difference in duration between the two pretonic positions is paralleled by a large difference in first formant values (Post forthcoming), confirming that our Moscow speakers use reduction in two degrees in both quantity and quality. In Perm, on the other hand, we see huge variation and no statistically significant average durational difference between  $a_{-1}$  and  $a_{-2}$ , suggesting it is not a categorical distinction in Perm speech.

When comparing our data with previous research, we find that the Moscow durational pattern of almost 1 : 2 : 3 for  $a_{-2} : a_{-1} : á_0$  is actually very close to Potebnja's formula for standard Russian vowel strength, and the relative durations Bondarko (1998) refers to, although other studies of CSR found smaller (e.g. Vysotskij 1973, group I) or larger differences (e.g. Barnes 2006) between the two pretonic positions. Our results from Moscow of almost 1 : 2 : 3 range between Vysotskij's group I for standard pronunciation (58 : 80 : 118 ms) and group II for Moscow vernacular speech (35 : 92 : 113 ms; Figure 1, groups I and II).<sup>10</sup>

As a possible reference example of CSR pronunciation, I also recorded a teacher of Russian as a foreign language, who was born in Leningrad (Saint Petersburg) in the 1950s but has lived most of her life in Moscow. Her mean durations of  $a_{-2} : a_{-1} : á_0$  were 35.8 : 83.6 : 126.5 ms (cf. Figure 7 below), so her values are closer to Vysotskij's measurements from Moscow vernacular speech than to his values for CSR and to our data from Moscow adolescents.

We have not assessed to which degree our adolescent speakers in Moscow and Perm speak standard Russian, with or without local colouring. We can conclude that the Moscow boys and girls have weaker local colouring in their word rhythmic pattern than the older teacher, but their difference between degree 1 and degree 2 reduction is larger than in most existing data for Central Standard Russian.

Our Perm pattern of almost 1 : 1 : 3, with only a small durational difference between the unstressed vowels ( $a_{-2}$  is, on average, 86% of  $a_{-1}$ ), is not the closest to Vysotskij's dialect group represented in the Ural region (Figure 1, group IX), as one could expect, since Perm is situated there. This group has a larger difference between the two pretonic vowels (with 33 : 53 : 78 ms for  $a_{-2} : a_{-1} : á_0$ ). Our Perm rhythm is actually closest to dialect group VIII, the rural dialect system with the smallest difference between  $a_{-2}$  and  $a_{-1}$ , found in a peripheral area in the North of European Russia (Figure 1, group VIII, where  $a_{-2}$  is 90% of  $a_{-1}$ ). The two-degree reduction in Perm is also less pronounced than in the data from Arkhangelsk and Chelyabinsk (Figure 2), the two students with the smallest difference in Grammatčikova et al.'s (2013) study of modern regional standard speech. In our data from Perm, the  $a_{-2} : a_{-1}$  ratio is similar to the ratio in Erofeeva's (2005) data from 4 speakers with a Perm accent, with 0.87 : 1 resp. 0.81 : 1. An important difference between our and Erofeeva's data is that the latter are from spontaneous speech; cf. Section 4.4.3 below. One of the pupils we recorded

---

<sup>10</sup>Our data are not fully comparable to the results from previous studies, since not all parameters influencing vowel duration always coincide. For instance, Vysotskij (1973) included words in utterance-final position. Still, most previous research is based on words with very similar segmental and rhythmic structure as the current study.

in Moscow happened to be from the Northern Vologda region. He, too, made hardly any difference between  $a_{-2}$  and  $a_{-1}$ , although all his vowels had longer durations than his peers in Moscow and Perm (40 : 44 : 94 ms; represented as V in Figure 7 in the next section). One should be aware that these single speakers from our own data, as well as those from earlier literature, need not be representative of the word rhythm of the city where they grew up, but they probably give some indication of it, since our data from Moscow and Perm suggest that word rhythm is a relatively stable regional trait.

When comparing the first pretonic vowel  $a_{-1}$  with the stressed vowel  $a_0$ , the stressed vowel is much longer than both prestressed vowels in our data, both in Perm and in Moscow. The difference between  $a_{-1}$  and  $a_0$  is smaller in Erofeeva's (2005) data from Perm and considerably smaller in most studies of CSR (e.g., Knjažev 2006, Grammatčikova et al. 2013).<sup>11</sup> This is mainly due to the uncommonly short pretonic vowels among our adolescent speakers. Perm and Moscow show little difference in their relative duration of the stressed vowel. This might be a general feature of varieties of Russian. The data in Figure 1 from Vysotskij (1973) show considerable variation between the varieties in the first and second pretonic vowel, but a remarkably stable relative duration of the stressed vowel, of almost 50% of the total duration of all three vowels in all 6 represented rhythmic patterns. In our data, the relative duration of the tonic vowel is not very different, though somewhat longer: 51% in Moscow and 58% in Perm. Vysotskij included target words in utterance-final position, which may have led to relatively longer stressed vowels and thus a larger proportion of the total than in our data with only non-utterance-final positions. Still, the last vowel had a relatively longer duration in our data.

#### 4.2 First pretonic prominence is relative: Vowels vs. consonants

Although the first pretonic vowels are twice as long as the second pretonic vowels in Moscow, they are, with 51 ms, still relatively short, compared both to earlier studies and to the surrounding consonants. We segmented and measured the consonants in *topotat'* in utterance (2) *Ja topotat' uže ne budu*, an example of which is shown in Figure 4. The pretonic consonants are much longer in this word than the pretonic vowels, even in Moscow speech, with its relatively long first pretonic vowels (Figure 7).

---

<sup>11</sup>An exception, with a large difference between  $a_{-1}$  and  $a_0$ , is Duryagin's (2018) speaker of CSR, whose vowel durations were 37 : 60 : 106 ms (Duryagin 2018: 327), but here, the difference is partly caused by the words' utterance-final position, which can cause final lengthening.

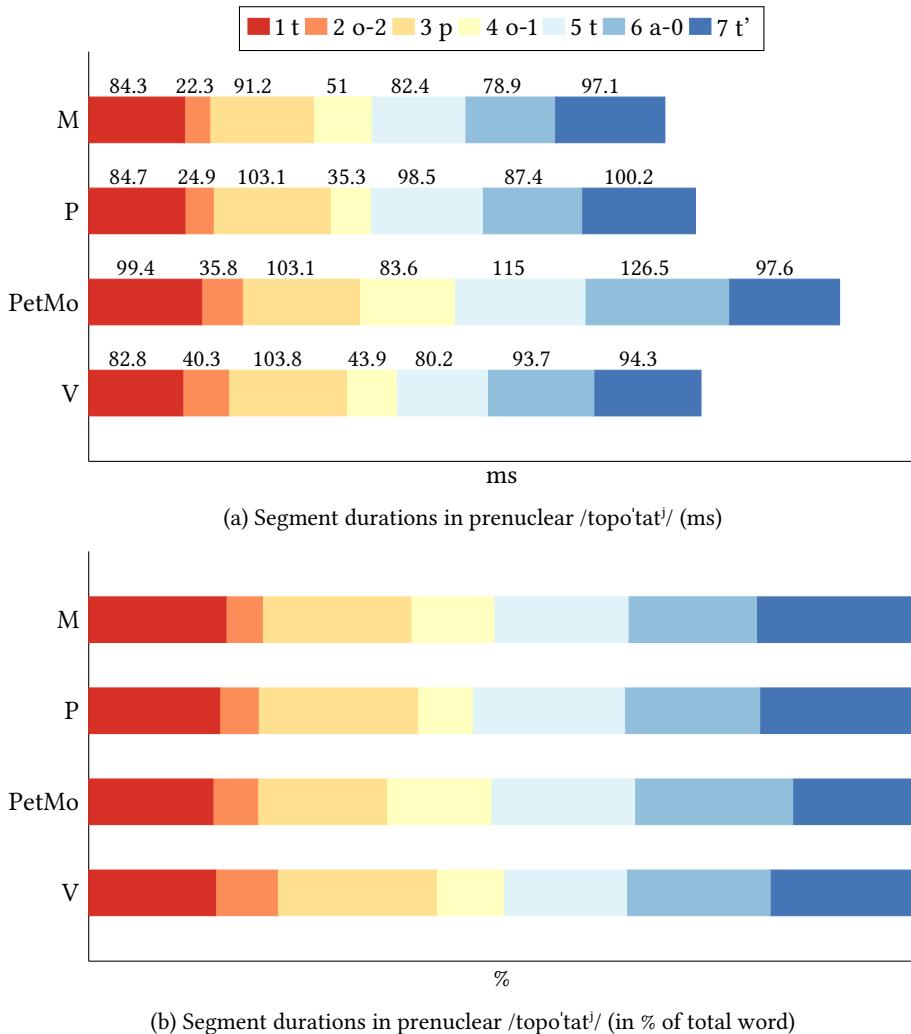


Figure 7: Durations of all consonants and vowels in *topotát'* in utterance (2), in milliseconds (upper graph) and relative to the duration of the whole word (lower graph). The upper row M shows the mean values for all 13 speakers from Moscow, the second row P for the 13 speakers from Perm. Row 3 (PetMo) gives the values for a teacher of Russian, and row 4 (V) for an adolescent from the Vologda region (Northern Russia). The numbers in the upper graph give the mean durations in milliseconds.

The numbers on which this figure is based are limited (one word production per speaker, so  $n = 13$  for row 1 (Moscow) and  $n = 13$  for row 2 for Perm), but there can be no doubt about the general picture that consonants are longer than vowels in unstressed syllables in both cities. The teacher I recorded (PetMo, third row), who is more than 40 years older than the other participants, has much longer vowels, both in milliseconds and relative to the other segments in the word (and not only in this utterance), but even in her speech, the unstressed consonants are longer than the vowels. In Vysotskij's (1973) data,  $a_{-1}$  has the same duration as the preceding consonant in CSR, but it is much longer in groups II and V.<sup>12</sup> The patterns in groups VIII and IX are again more like the Perm pattern.

Our Muscovite speakers obviously did not use the extremely prominent first pretonic vowels which the old Moscow dialect and Moscow vernacular speech are known for (Vysotskij 1973, Lapteva 1999), but they are still long compared to the second pretonic vowels.

#### 4.3 Two-degree reduction in East-Slavic: Strong centre and weak periphery

The durational reduction in two degrees is claimed to be a feature shared by all three East Slavic languages (Dubina 2012), but the difference between second and first pretonic vowels is small among speakers of Ukrainian from the west of Ukraine, at a long distance from Central Russia (Łukaszewicz et al. 2022).<sup>13</sup> Earlier literature on rural dialects in Russia (cf. Section 1.2.2) suggested that the relative difference between the two pretonic vowels is strongest in a central region of the East Slavic dialect continuum, and gets weaker when moving from the centre, and that this difference is retained in modern urban Russian (cf. Section 1.2.3). Our data, although representing only two cities, is in line with these earlier, preliminary observations.

#### 4.4 Sociolinguistic dimension

##### 4.4.1 Gender

Our remaining research questions concern possible gender effects, first on overall vowel duration: Do the girls have longer vowel durations than the boys, as

<sup>12</sup>Vysotskij (1973) measured not only vowel durations in 15 different dialect groups, but also the durations of the surrounding consonants (1973: 39–40, Figs. 2 and 3), but he did not discuss the consonantal durations.

<sup>13</sup>The word prosodic pattern in Ukrainian is complicated by iterative secondary stresses (Łukaszewicz & Molczanow 2018).

female speakers tend to have in many other languages? This is true for the female speakers in Perm, but not for the female speakers in Moscow, who use almost the same vowel durations as their male classmates, for all vowels in general, and also for each individual position (cf. Section 3).

The scarce literature on gender effects on phonetics in Russian suggests that the girls might have longer pretonic vowels than the boys, and that the Perm girls might have a vowel reduction pattern that is more similar to the Moscow speakers and to the Central Standard Russian norm than the Perm boys, in case it is true that they speak closer to a non-local norm (see Section 1.2.4 above). Neither can be found in our data. The girls do not use stronger first pretonic lengthening than the boys, nor do the girls in either city show less local colouring than their male peers, as young urban women often do in other countries (Labov 2001), at least not in their word prosodic structure, which might be less conscious than other, more salient features, such as the stigmatised pronunciation of [o] in unstressed position (Andrews 1995), which is avoided in formal speech (Erofeeva 1993). In their question intonation, however, the same girls in Perm did indeed speak with less colouring than the boys, as we found in a recent study of the same speakers (Post & Andreeva 2023).

These results should preferably be confirmed with larger data sets from more speakers. We have no guarantee that our 13 speakers from a single school in Moscow and their peers in Perm are representative for todays' youth from the two cities.

#### 4.4.2 Other sociolinguistic variables

As to other sociolinguistic parameters, we did not study the role of education, since almost all adolescents had parents with a high education level and they were still at school age themselves. We did not focus on age either, but one of our participants – the teacher from Saint Petersburg and Moscow – was more than 40 years older than the pupils from Moscow and Perm. It is no surprise that her vowels were much longer, since young people tend to have a faster articulation rate. More interesting for our research question is that the teacher also had a much larger difference between the pretonic vowels, due to her relatively long first pretonic vowels. Her two-degree reduction is closer to the vernacular Moscow speech and the older Moscow norm that Vysotskij recorded more than fifty years ago (Figure 1, groups II and III). It is stronger not only than in today's adolescent speech, but even stronger than in Vysotskij's (1973) pattern for Central Standard Russian (Figure 1, group I), so she has relatively strong local colouring in her word rhythm pattern.

#### **4.4.3 Even in read speech**

The results show a remarkable difference between the cities, even though the data are obtained from read speech. Speaking style has a major effect on the frequency of unreduced, [o]-like pronunciations of unstressed /o/ among speakers from Perm, ranging from 4% in read speech to 69% in spontaneous speech (Verbickaja et al. 1984, Erofeeva 1993, cf. Section 1.2.4).

Both Vysotskij (1973) and Erofeeva (2005) used data from spontaneous speech, which tends to show a higher degree of local colouring than read speech, and Erofeeva's speakers were even known to have a local accent, for which our speakers have not been screened. Therefore, it is remarkable that our data show the same high degree of local colouring as Erofeeva's spontaneous data, even though our recordings are from formal, read speech. This supports the suggestion that reduction in only one degree is not considered a serious violation of standard Russian, only of orthoepic rules (Bondarko 1998, cf. Section 1.2.4 above), unlike the production of unstressed /o/ as [o], and thus of lack of reduction, which is actively avoided in read speech. It is highly probable that the rhythmic feature we studied is less stigmatised than more salient dialect traits (cf. Grammatčikova et al. 2013: 72).

## **5 Conclusions**

Our study of relative vowel durations in prestressed position among young urban Russians from Moscow and from Perm shows that we do find regional variation in modern-day Russian urban speech, at least in prosody: All Moscow speakers make a clear distinction between two degrees of reduction, whereas the speakers in Perm hardly differentiate the second and first pretonic vowels. In speech by speakers from Perm we see very little first pretonic vowel prominence. In this regard, the Perm variety is in fact close to the most extreme rural dialects, which do not have, or hardly have, reduction in two degrees. This regional distinction between Moscow and Perm speech is very stable. There is no effect of gender, and it is maintained even in read speech, where the tendency to suppress regional dialect traits is strongest. Apparently, whereas the production of unreduced [o] is actively avoided, and/or on its way out of Perm speech, vowel reduction in only one degree is not.

The two-degree reduction in quantity appears to be categorical for our Moscow speakers. Our data do not suggest that this is the case in Perm. Although we do find a small difference in quantity in Perm, this difference is not statistically significant, and the degree of variation is very high. The small difference we see

in average values need not be a categorical distinction, only a gradual tendency to be somewhat longer.

It would be no surprise if today's Moscow youth speak with less local accent than previous generations. Still, although the first pretonic is never that prominent as in the extreme cases Moscow speech is known for, they retain a very distinct two-degree reduction. The relative difference between first and second pretonic is large, and larger than in many previous recordings of Central Standard Russian speech, due to the extremely short second pretonics among our Moscow speakers. It will be interesting to observe the further development and the social status of a Moscow accent, both in and outside Russia, in the current, changing world.

Although only covering two cities, our data are in line with previous observations of an opposition between centre and periphery in the East Slavic dialect continuum in the expression of the word prosodic structure. A difference between strong pretonic prominence in the central area and weaker prominence in non-central areas is still found in modern urban Russian.

In future research, data from more cities should be added. Moscow and Perm might be extremes on a scale, since both cities are known for a relatively strong local accent. When Grammatčikova et al. (2013) claimed that Russians often can distinguish speakers from different cities, they mentioned three cities as examples: Bryansk (Western Russia, not far from Belarus and Ukraine), Perm (Ural region) and Moscow. I suspect that these cities are not chosen randomly, but because they are known for their local features.

## Acknowledgements

The author is obliged to the following people for support in various stages of the research: First of all, to Benedikte Fjellanger Vardøy for doing the recordings in Perm and for fruitful cooperation in an earlier analysis of the vowels, further to Svetlana Djačenko for doing most of the segmentations and to Bistra Andreeva for statistical analyses and advice. I would also like to thank Alexander Krasovitsky, Sergej Knjazev, Elena Erofeeva, Brechtje Post, Elaine Schmidt and Dirk Jan Vet for advice on phonetics at different stages of this research, the members of the Phonetics group at Saarland University for their hospitality and advice during my research stay in 2022, and the two anonymous reviewers for valuable response. I remain responsible for all remaining shortcomings. Last but not least, I thank the speakers and the secondary schools in Moscow and Perm they attended for their participation.

This research was supported by various grants from the University of Bergen and by the Meltzer Foundation in Bergen, Norway. The data collection has been approved by the Norwegian Social Science Data Services (NSD).

## References

- Al'muxamedova, Zel'fa (ed.). 1985. *Gradacionnaja fonologija jazyka i prosodija slova russkoj dialektnoj reči*. Kazan': Kazan' University Press.
- Andrews, David R. 1995. Subjective reactions to two regional pronunciations of great Russian: A matched-guise study. *Canadian Slavonic Papers* 37(1–2). 89–106. DOI: 10.1080/00085006.1995.11092083.
- Andrews, David R. 2006. The role of Émigré Russian in redefining the “Standard”. *Journal of Slavic Linguistics* 14(2). 169–189.
- Auer, Peter. 2005. Europe's sociolinguistic unity, or: A typology of European dialect/standard constellations. In Nicole Delbecque, Johan van der Auwera & Dirk Geeraerts (eds.), *Perspectives on variation: Sociolinguistic, historical, comparative*, 7–42. Berlin, New York: de Gruyter Mouton.
- Avanesov, Ruben. 1984. *Russkoe literaturnoe proiznošenie*. 6th edn. Moscow: Prosveščenie.
- Barnes, Jonathan. 2006. *Strength and weakness at the interface: Positional neutralization in phonetics and phonology*. Berlin, New York: de Gruyter Mouton. DOI: 10.1515/9783110197617.
- Bethin, Christina Y. 2006. Stress and tone in East Slavic dialects. *Phonology* 23(2). 125–156. DOI: 10.1017/S0952675706000868.
- Boersma, Paul & David Weenink. 1992–2024. *Praat: Doing phonetics by computer*. <https://www.fon.hum.uva.nl/praat/>.
- Bondarko, Lija. 1998. *Fonetika sovremennoj russkogo jazyka*. Saint Petersburg: Saint Petersburg University Press.
- Bondarko, Lija, Ljudmila Verbickaja & Lev Zinder. 1966. Akustičeskie xarakteristiki bezudarnosti. In Vyacheslav Ivanov (ed.), *Strukturnaja tipologija jazykov*, 56–64. Moscow: Nauka.
- Čekmonas, Valerij. 2001. K izučeniju vokalizma govorov pskovščiny. Ritmičeskaja struktura slova i akustičeskie osobennosti realizacii glasnyx v govore d. Teševicy pskovskogo rajona. In Klavdija Gorškova & Marina Remnëva (eds.), *Dialektnaja fonetika russkogo jazyka v diachronnom i sinchronnom aspektakh*, 43–85. Moscow: Moscow State University Press.
- Comrie, Bernard & Greville G. Corbett. 1993. *The Slavonic languages*. London: Routledge.

- Crosswhite, Katherine M. 2000. Vowel reduction in Russian: A unified account of standard, dialectal, and “dissimilative” patterns. *University of Rochester Working Papers in the Language Sciences*, Spring 2020. 107–171.
- D’jačenko, Svetlana. 2015. Glasnye pervogo predudarnogo i udarnogo slogov v arxaičeskom južnorusskom govore. Količestvennyj analiz. In Ljudmila Kalnyn’ (ed.), *Issledovaniya po slavjanskoj dialektologii XVII*, vol. 17, 244–297. Moscow: Russian Academy of Sciences, Institut slavjanovedenija.
- Dubina, Andrei. 2012. *Towards a tonal analysis of free stress*. Radboud Universiteit Nijmegen. (Doctoral dissertation).
- Duryagin, Pavel V. 2018. Kačestvo i dilitel’nost’ bezudarnyx glasnyx russkogo jazyka kak akustičeskie ključi dlja opredelenija slovesnoj granicy: Perceptivnyj èksperiment na materiale psevdoslov [Duration and formant values of unstressed vowels in Russian as acoustic cues for segmentation: A perceptive experiment based on nonce words]. *Rusistika/Russian Language Studies* 16(3). 322–343. DOI: 10.22363/2618-8163-2018-16-3.
- Erofeeva, Elena. 1993. *Fonetičeskaja variativnost’ fonologičeskoj sistemy glasnyx (sopostavitel’nyj analiz peterburgskogo i permskogo vokalizma)*. Saint Petersburg: Saint Petersburg State University. (Candidate dissertation). [dlib.rsl.ru/viewer/01000128725#?page=8](http://dlib.rsl.ru/viewer/01000128725#?page=8).
- Erofeeva, Elena. 2005. *Idiomy kak verojatnostnaja struktura idiomov: Sociolinguističeskij aspekt: Na materiale fonetičeskogo urovnja*. Saint Petersburg State University. (Doctoral dissertation).
- Erofeeva, Tamara. 1995. *Sociolect: Stratifikacionnoe issledovanie*. Saint Petersburg State University. (Doctoral dissertation).
- Erofeeva, Tamara, Elena Erofeeva & I. I. Gračeva. 2000. *Gorodskie sociolekty: Permskaja gorodskaja reč’*. Zvučaščaja xrestomatija (book + cd). Perm’ – Bochum: Permskij gosudarstvennyj universitet – Ruhr-Universität Bochum.
- Grammatčikova, E. V., S. V. Knjazev, L. V. Luk’janova & S. K. Požarickaja. 2013. Ritmičeskaja struktura slova i mesto realizacii tonal’nogo akcenta v regional’nyx variantax sovremennogo russkogo literaturnogo jazyka. In A. V. Archipova & I. M. Kobozeva (eds.), *Aktual’nye voprosy teoretičeskoj i prikladnoj fonetiki*, 69–90. Moscow: Buki-Vedi.
- Hertz, Birgitte, Hanne Leervad, Henrik Møller & Peter Schousboe. 2005. *Svidanie v Peterburge. Møde i Petersborg*. Copenhagen: Gyldendal.
- Iosad, Pavel. 2012. Vowel reduction in Russian: No phonetics in phonology. *Journal of Linguistics* 48(3). 521–571. DOI: 10.1017/S002226712000102.
- Kalenčuk, Marija. 2021. Uzual’nye i kodificirovannye proiznositel’nye normy. In Marija Kalenčuk & Dmitrij Savinov (eds.), *Norma proiznošenija v uzuse i kodiranii*.

- fikacii, 4–25. Moscow: Russian Language Institute of the Russian Academy of Sciences.
- Kasatkina, Rozalija. 2003. Vocalism of the Russian language: Gender differences. In *15th International Congress of Phonetic Sciences (ICPhS-15), 1839–1842*. Barcelona. [www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2003/p15\\_1839.html](http://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2003/p15_1839.html).
- Kasatkina, Rozalija. 2005. Moskovskoje akan'e v svete nekotoryx dialektnyx dannyx. *Voprosy jazykoznanija* 2. 29–45.
- Knjazev, Sergej. 2006. *Struktura foneticheskogo slova v russkom jazyke: Sinxronija i diaxronija*. Moscow: MAKS-Press.
- Kocharov, Daniil, Tatiana Kachkovskaia & Pavel Skrelin. 2015. Position-dependent vowel reduction in Russian. In *18th International Congress of Phonetic Sciences (ICPhS-18)*. Glasgow. [www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0402.pdf](http://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0402.pdf).
- Kochetov, Alexei. 2006. The role of social factors in the dynamics of sound change: A case study of a Russian dialect. *Language Variation and Change* 18(1). 99–119. DOI: 10.1017/S0954394506060030.
- Kodzasov, Sandro. 1999. Russian. In Harry van der Hulst (ed.), *Word prosodic systems in the languages of Europe*, 852–870. Berlin: de Gruyter.
- Krause, Marion. 2010. Zur Typologie von Sprachsituationen: Binnensprachliche Variation zwischen Standard und Dialekt im heutigen Russland. *Wiener Slawistischer Almanach* 65. 53–81.
- Krysin, Leonid. 2007. Russkaja literaturnaja norma i sovremennaja rečevaja praktika. *Russkij jazyk v naučnom osveščenii* 14(2). 5–17.
- Kuznecov, Vladimir. 1997. *Vokalizm svjaznoj reči: Èksperimental'noe issledovanie na materiale russkogo jazyka*. Saint Petersburg: Saint Petersburg University Press.
- Labov, William. 2001. *Principles of linguistic change*, vol. 2: Social factors. Oxford: Blackwell.
- Lapteva, Ol'ga. 1999. *Živaja russkaja reč' s teleekrana: Razgovornyj plast televizionnoj reči v normativnom aspekte*. 2nd edn. Moscow: URSS.
- Łukaszewicz, Beata & Janina Mołczanow. 2018. The role of vowel parameters in defining lexical and subsidiary stress in Ukrainian. *Poznan Studies in Contemporary Linguistics* 54(3). 355–375. DOI: 10.1515/pscl-2018-0014.
- Łukaszewicz, Beata, Janina Molczanow & Anna Łukaszewicz. 2022. *Pretonic lengthening as the lexical stress domain extension*. Lisbon: Universidade de Lisboa. 372–376. DOI: 10.21437/SpeechProsody.2022-76.

- Molczanow, Janina. 2015. The interaction of tone and vowel quality in optimality theory: A study of Moscow Russian vowel reduction. *Lingua* 163. 108–137. DOI: 10.1016/j.lingua.2015.05.007.
- Nikolaeva, Tatjana. 1977. *Frazovaja intonacija slavjanskix jazykov*. Moscow: Nauka.
- Padgett, Jaye & Marija Tabain. 2005. Adaptive dispersion theory and phonological vowel reduction in Russian. *Phonetica* 62(1). 14–54. DOI: 10.1159/000087223.
- Panov, Mixail. 1967. *Russkaja fonetika*. Moscow: Prosveščenie.
- Paufošima, Rozalija. 1978. Perestrojka sistemy predudarnogo vokalizma v odnom vologodskom govore. In Sergej Vysotskij (ed.), *Fizičeskie osnovy sovremennych fonetičeskix processov v russkix govorax*, 18–66. Moscow: Nauka.
- Post, Margje. 2017. Why regional prosodic variation is worth studying: An example from Russian. In Margrete Dyvik Cardona, Randi Koppen & Ingunn Lunde (eds.), *Remaining relevant: Modern language studies today* (Bergen Language and Linguistics Studies (BeLLS) 7), 164–182. Bergen: University of Bergen. DOI: 10.15845/bells.v7i0.1138.
- Post, Margje. Forthcoming. Word prosodic structure and vowel reduction in Moscow and Perm Russian. In Berit Gehrke, Denisa Lehertová, Roland Meyer, Daria Seres, Luka Szucsich & Joanna Zaleska (eds.), *Advances in Formal Slavic Linguistics 2022* (Open Slavic Linguistics). Berlin: Language Science Press.
- Post, Margje & Bistra Andreeva. 2023. Polar question intonation in Russian speech from Moscow and Perm. In *Proceedings of the 13th International Conference of Nordic Prosody*, 147–154. DOI: 10.2478/9788366675728-012.
- Potebnja, Aleksandr. 1866. *Dva issledovanija o zvukax russkogo jazyka: I.O polnoglasii. II. O zvukovyx osobennostjax russkix narečij*. Voronež: Tipografija V. Gol'dstejna.
- Rozanova, Nina. 1988. Ob odnoj osobennosti staromoskovskogo proiznošenija v sovremennoj reči moskvičej. In D. N. Šmelev & E. A. Zemskaja (eds.), *Raznovidnosti gorodskoj ustnoj reči: Sbornik naučnyx trudov*, 208–223. Moscow: Nauka.
- Savinov, Dmitrij. 2013. *Èvoljucija sistem vokalizma v južnorusskix govorax*. Moscow: Russian Language Institute of the Russian Academy of Sciences. (Doctoral dissertation).
- Simpson, Adrian P. & Christine Ericksdotter. 2003. Sex-specific durational differences in English and Swedish. In Maria-Josep Solé, Daniel Recasens & Joaquín Romero (eds.), *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS-15)*, 1113–1116. Barcelona: The International Phonetic Association. [https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2003/p15\\_1113.html](https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2003/p15_1113.html).

- Vardøy, Benedikte Fjellanger. 2021. Mapping young Russians' perceptions of regional variation in Russian. *Journal of Linguistic Geography* 9(1). 50–69. DOI: 10.1017/jlg.2021.5.
- Verbickaja, Ljudmila. 1977. Variantnost' sovremennoj proiznositel'noj normy russkogo literaturnogo jazyka. *Vestnik Leningradskogo universiteta. Serija istorija, jazyka i literatury* 2(8). 133–137.
- Verbickaja, Ljudmila, L. V. Ignatkina, N. F. Livačuk, T. A. Sergeeva, M. V. Cvetkova & V. G. Ščukin. 1984. Regional'nye osobennosti realizacii russkoj reči (na fonetičeskem urovne). *Vestnik Leningradskogo universiteta. Serija istorija, jazyka i literatury* 2(8). 71–80.
- Vysotskij, Sergej. 1973. O zvukovoj strukture slova v russkix govorax. In Sof'ja Bromlej (ed.), *Issledovaniya po russkoj dialektologii*, 17–41. Moscow: Nauka.
- Zemskaja, Elena, Margarita Kitajgorodskaja & Nina Rozanova. 1987. Osobennosti mužskoj i ženskoj reči v sovremennom russkom jazyke. In *Proceedings of the 11th International Congress of Phonetic Sciences (ICPhS-11)*, vol. 1, 191–194. Tallinn: International Phonetic Association. [https://www.coli.uni-saarland.de/groups/BM/phonetics/icphs/ICPhS1987/11\\_ICPhS\\_1987\\_Vol\\_1/p11.1\\_191.pdf](https://www.coli.uni-saarland.de/groups/BM/phonetics/icphs/ICPhS1987/11_ICPhS_1987_Vol_1/p11.1_191.pdf).
- Zlatoustova, Ljubov'. 1981. *Fonetičeskie edinicy russkoj reči*. Moscow: Moscow State University Press.

# Chapter 12

## How important is information about grandparents when selecting a dialect speaker?

Akiko Takemura

Université d'Orléans & Kobe University

This paper aims at setting a new criterion for selecting a local dialect speaker by inquiring about the origin of their parents, which accommodates to the reality in large cities. In Japanese dialectology, it is widely accepted that a dialect speaker should ideally be *haenuki* ('native-born'), i.e. a person from a family who has lived in the survey area for three generations. However, this criterion has been used hitherto without checking the assumptions that it is both justified and useful. After reexamining the data collected from three families by the National Language Research Institute (1965) and conducting one survey, I conclude that the *haenuki* criterion is impractical, and not fully justified and that the most important criterion when selecting a representative speaker of a dialect should be having both parents native born.

### 1 Introduction

Dialect surveys crucially rely on a sample of speakers taken to be representative of a dialect. The criteria used for selecting a representative speaker and the very definition of who can be considered to be a speaker of a dialect are important methodological issues. In Japanese dialectology, it is widely accepted that a dialect speaker should ideally be *haenuki* ('native born'), i.e., a person from a family who has lived in the survey area for three generations. However, this criterion has been used hitherto without checking whether it is both justified and useful.



In this paper, I would like to discuss whether it is useful to take into account information about a speaker's grandparents in dialect surveys.

After reviewing in detail the existing literature on the *haenuki* criterion and its motivation, I present the results of a reexamination of data obtained by the National Language Research Institute (1965) and of a survey, conducted by the author, of Japanese dialects intended to clarify whether it is indeed justified: the former is on the transmission of lexical accent across three generations, and the latter is on the transmission of lexical accent and phonological rules across two generations. In view of these results, I conclude that the *haenuki* criterion is unpractical and not fully justified, especially in large cities, and that the most important criterion for selecting a representative speaker of a dialect should be having both parents native born.

## 2 Criteria for a representative speaker of a dialect

Table 1 details the different criteria proposed in the Japanese literature to determine whether a speaker can be taken or not as a representative of a dialect.<sup>1</sup> In Japanese dialectology, a speaker is considered representative of a dialect if they, along with their parents and grandparents, were born and raised in the survey area. However, this widely accepted criterion has been adopted without much justification and its usefulness remains to be shown.

The strict application of this criterion poses the practical problem that in some cases such speakers are rare or even nonexistent (Sibata 1984, Sibata 1988), especially nowadays due to the increased mobility of populations for reasons of marriage, education, work, etc. For example, Sibata (1984: 64, my translation) reports that "in 1965, no one met the criteria of being over 65 years old and *haenuki* in the surroundings of the Urawa Station in Saitama prefecture", and similarly Sibata (1988: 566, my translation) reports that:

We have been able to find many natives who were born in the area and have lived there for a long time, but from now on there will be fewer people who meet these conditions. [...] Even if the conditions are relaxed, I do not think it will be possible to find anyone who meets the conditions in the future.

---

<sup>1</sup>Yoshida (1984) says that the influence of parents and spouses is very limited, but also that the origins of parents and spouses should be taken into account when researching accents. It is somewhat contradictory, but I assume that the former concerns the influence on the acquisition of grammar or lexical items.

Table 1: Criteria for dialect speakers

Source	Criterion
National Language Research Institute (1981)	lexical accent is influenced by where a person spent their language formation period, and not by any afterward relocation
Uwano (1984), Uwano (1997)	<ul style="list-style-type: none"> <li>speakers who have lived in the area throughout their so-called “language formation period”</li> <li>it is ideal if both the parents and grandparents are from the location; parental influence is weak, especially the father’s influence, as it is seen in the vocabulary used only at home</li> <li>“people who have lived somewhere else than their birth place at least once” because they are more sensitive to variation (assuming that those people often have a better sense of language)</li> </ul>
Yoshida (1984)	<ul style="list-style-type: none"> <li>ideally, <i>haenuki</i></li> <li>people who have lived in the area until the age of language formation (around 13 years old) and have little history of living outside</li> <li>influence of parents and spouse is very limited</li> <li>note that the influence of parents’ language is more important in surveys of young people</li> <li>origins of parents and spouses should be taken into account when it comes to accentual research</li> </ul>

Outside the circles of Japanese dialectology, the NORM (Non-mobile, Old, Rural, Men) criterion is widely used in dialect research, but it is difficult to apply in the case of large cities. To capture the linguistic dynamics in large cities, Chambers & Heisler (1999: 40–46) proposed the Regionality Index (RI):

In the Dialect Topography sample, we aim for a demographic cross section of the survey area rather than a population of indigènes as in traditional dialect surveys. Our reasoning is straightforward: some proportion of urban speech communities is made up of people who were born outside the community, and the variants they use in their speech are heard in that speech community and have some status in it. We want to know what those variants are and the extent of their use. Our quantitative model allows us to distinguish indigènes and interloper variation by correlating variants with the provenance (via RI) of the respondents who use them.

The RI is thus a measure of the degree of representativeness of a speaker for a given dialect based on the speaker's background factors such as birthplace, current residence during the survey, upbringing between the ages of 8 to 18, and the birthplace of their parents. The RI is computed using this information and yields values ranging from 1 to 7. The lower the RI value, the greater the speaker's nativeness. To illustrate, consider a speaker with an RI value of 1. This individual is not only born and raised in the same location as this individual's parents, but also lives there currently, reflecting a strong local connection. Conversely, an RI value of 7 denotes an individual who resides within the surveyed region but was born and raised outside the area, with parents also coming from the other area. The RI aids in distinguishing speech variants employed by natives versus newcomers, making it particularly suitable for extensive investigations within large cities. However, it may not prove optimal for smaller-scale studies where the selection of a limited number of representative speakers is unavoidable.

As noted previously, the *haenuki* criterion has been used in Japan in dialect research without much consideration. Yet, its applicability in contemporary times, specifically in conducting dialect research in large cities, raises considerable challenges. The scarcity of people who satisfy such strict criteria makes its feasibility questionable. Consequently, inquiries into the credibility of this criterion are well-founded. In an effort to assess its validity, this study aims to address two key research questions:

1. How important is information about grandparents?
2. Do parents' origins matter in dialect acquisition?

To answer the first question, I reexamined the data collected in the 1960s (National Language Research Institute 1965) on lexical accent transmission across three generations in three different locations in Hokkaido (Section 3). I then investigated the influence of parents on the acquisition of lexical accent and phonological rules in young speakers in Kagoshima (Section 4) (Takemura 2012) to answer the second question.

### 3 Transmission of lexical accent across three generations in Hokkaido

The National Language Research Institute (1965) presents data collected in the 1960s on three different families located in three locations on Hokkaido Island in order to study the process of language standardization. Hokkaido was traditionally inhabited by the indigenous Ainu people, and it was only during the Meiji era (1868–1912) that Japanese speakers began to settle in Hokkaido. The settlers came from all over Japan bringing with them their own dialect. Therefore, Hokkaido at that time was a linguistically heterogeneous place without any characteristic dialect, and a standard language emerged.

The National Language Research Institute (1965) researchers collected data from three locations in Hokkaido: Bibai, Ikeda and Kutchan (Figure 1). At each location, a family of three generations participated in the survey. They were asked to pronounce each word on wordlists, and researchers wrote down the accent patterns observed, and I compared to what extent the accent patterns for a given word match across the different generations.

There are gaps in the data for the National Language Research Institute (1965), so I focused on those lexical items for which we have information on their accentuation in at least two generations and compared the agreement ratio as follows. Agreement ratio is calculated: the denominator corresponds to the total number of lexical accents obtained between the given two generations. Meanwhile, the numerator hinges on the tally of lexical items articulated in the same manner across the two generations. To illustrate, consider the lexical item *hinomaru* ('the rising-sun flag'). For the first generation of the family from the town of Ikeda, the pronunciation follows the pattern Low-High-High-Low (abbreviated as LHHL hereafter). However, the second and third generations pronounce it as LHHH. Consequently, there is no concurrence between the first generation and the subsequent two generations.

Certain lexical items lack accent information for specific generations. For instance, for the word *tomodachi* ('friend'), the first generation's recorded accent

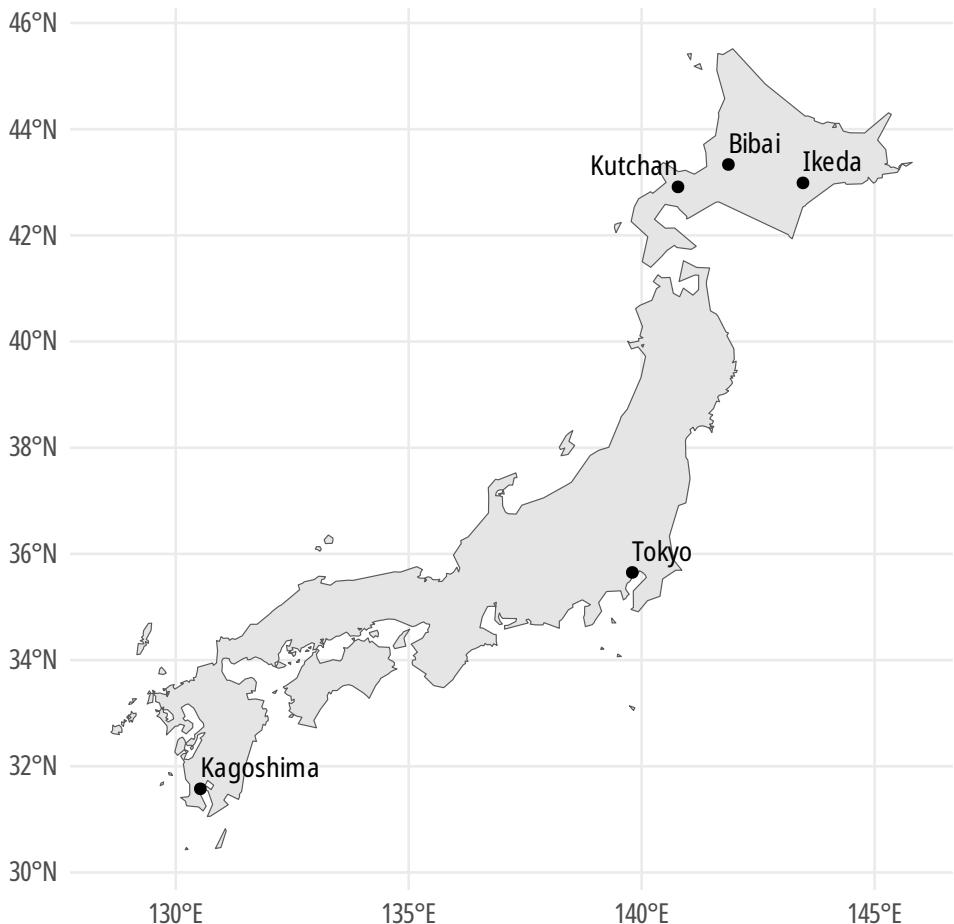


Figure 1: Map of Japan showing Tokyo and the locations of the surveys

is LHHL, while the third generation's is LHHH. However, there is no available information for the second generation. In such cases, the agreement ratio cannot be computed between the first and second generations, or between the second and third generations. Nonetheless, the calculation of the agreement ratio between the first and third generations remains possible. Figure 2 summarizes the results, and more details can be found in the appendix.

The transmission of dialect accentuation from grandparents to grandchildren differs from family to family, but we can still notice some trend in the data. There is a larger ratio disagreement than agreement ratio between the first and third generations in all three families, which shows that accent transmission is imperfect between the first and third generations. Interestingly, a family from Ikeda

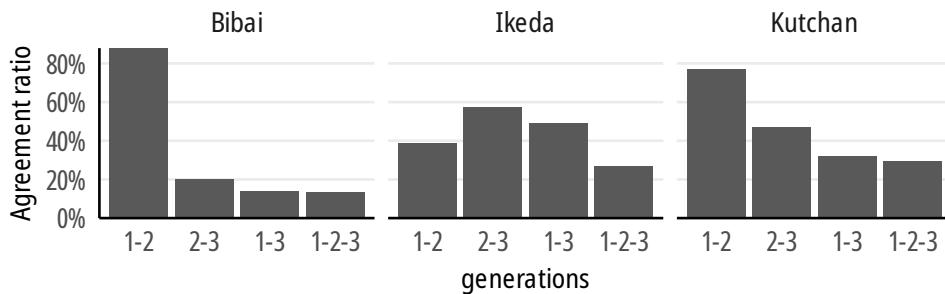


Figure 2: Agreement ratio in accentuation between generations in Hokkaido; generation 1 is that of the grandparents, 2 that of the parents, and 3 that of the children

town shows a reverse trend from other families for the agreement between the first and second generations. Whereas the agreement ratio between the first and second generations is usually higher than between the first and third generations, in the family from Ikeda it is the opposite, in accord with the findings of Inoue & Hanzawa (2022) that the speech of children raised with their grandparents is influenced by that of their grandparents.

Revisiting the initial research question posited in the previous section – the significance of information about grandparents – a definitive answer remains elusive. However, an overarching trend emerges indicating that accentuation transmission encounters imperfections between the first and third generations. This implies that intergenerational influence is more restricted to adjacent generations.

#### 4 Parental influence in the acquisition of lexical accent and phonological rule in Kagoshima

If, as seen in the previous section, there is an influence between adjacent generations, we expect that parental origins play an important role in dialect acquisition. On the other hand, if a child acquires its local dialect from his peers, there should be no difference between the children with locally born parents and those with non-locally born parents as to their dialect acquisition. I can thus categorize children into four groups depending on their parental origins (Table 2). Group 1 in Table 2 has both locally born parents. Group 2 has a local mother but non-local father. Group 3 has a non-local mother but a local father. Lastly, Group 4 has neither a local mother nor a local father.

Table 2: Child categorization according to their parental origins

		Father	
		Local	Non-local
Mother	Local	Group 1	Group 2
	Non-local	Group 3	Group 4

Here I present the results of the survey by Takemura (2012) on accent acquisition of the Kagoshima dialect in Japan (Figure 1). It focuses on two aspects: acquisition of the lexical accents of words, which is stored in the mental lexicon as a part of lexical representations, and acquisition of productive accent rules, which must be inferred from the linguistic input. The survey aims at answering the following questions:

1. Do parental origins influence accent acquisition?
2. If parental origins matter in dialect accent acquisition, is there a difference between the acquisition of lexical accent and the acquisition of phonological rules?

Whereas Tokyo Japanese is mora-based and has a free pitch accent system, the Kagoshima dialect is syllable-based and has only two possible pitch patterns: Pattern A with a high pitch on the penultimate syllable and Pattern B with a high pitch on the final syllable. Contrary to Tokyo Japanese, the location of the high pitch shifts when a particle is added to a word since the domain of pattern is the prosodic word (Tables 3 and 4).<sup>2</sup>

For example, *tabemóno* ('food'), in Table 3, is traditionally a Pattern A word. Within this word, the penultimate syllable *mó* is pronounced with high pitch, as indicated by the acute accent on the vowel. When we add a nominative particle *ga* after *tabemóno*, then the high pitch shifts to *nó* because it is now the penultimate syllable. Pattern B behaves similarly, with the high pitch shifting to the final syllable depending on word length, e.g. *nomimonó* ('beverage') but *nomimono=gá* in Table 4.

<sup>2</sup>Takemura (2012) also presents data on the acquisition of rule-based accentuation of compounds. However, the acquisition of compound accentuation is also affected by the confounding factor of ongoing influence of Tokyo Japanese (Kubozono 2006), which makes those data more problematic than useful to be included in the present study.

Table 3: Kagoshima Pattern A: High pitch on the penultimate syllable

	Form	Gloss
in isolation	<i>tabemóno</i>	‘food’
with a particle	<i>tabemonó=gá</i>	‘food=NOM’

Table 4: Kagoshima Pattern B: High pitch on the final syllable

	Form	Gloss
in isolation	<i>nomimonó</i>	‘beverage’
with a particle	<i>nomimono=gá</i>	‘beverage=NOM’

Tables 5 and 6 illustrate how I checked the acquisition of phonological rule. When a speaker pronounces ‘food’ as *tabemóno* in isolation and as *tabemonó=gá* with a particle, I thus conclude that the speaker has acquired the phonological rule since the location of the high pitch shifts according to the length of the prosodic word. However, if the pronunciation is *tabemóno=gá* with no shift, then I conclude that even though the speaker has acquired the traditional lexical accent, the speaker has not acquired the phonological rule. In the case of a speaker pronouncing *tabemonó* and *tabemono=gá*, it signifies the attainment of the phonological rule but not the acquisition of lexical accent. Conversely, if the pronunciation is *tabemonó* and *tabemonó=gá*, it implies the speaker has not acquired either the lexical accent or the phonological rule. I use the same criteria for Pattern B words such as *nomimonó* (‘beverage’) (Table 6).

There were 55 participants aged between 15 and 28 years old, whose family background is summarized in Table 7. Figure 3 summarizes the results of isolated forms, which reflect the acquisition of lexical accent. It indicates the agreement ratio of productions that matches the traditional pattern. We observe that Group 1 (both parents locally born) is the most conservative and that Group 4 (no parent locally born) is the most innovative. A Tukey test shows that Group 1 differs significantly from Groups 3 and 4 ( $p < 0.001$ ).<sup>3</sup> Thus, even though all informants grew up in the Kagoshima area, the acquisition of lexical accent differs between those with non-local parents.

<sup>3</sup>A Tukey HSD test was used to correct for multiple comparisons. There was also a statistically significant difference between Group 2 and Group 4 ( $p < 0.05$ ), but no statistically significant differences were observed between any other two groups..

Table 5: Criteria for process of the acquisition of Pattern A

In isolation	With a particle	Lexical accent	Phonological rule
<i>tabemonó</i>	<i>tabemonó=gá</i>	YES	YES
<i>tabemonó</i>	<i>tabemonó=gá</i>	YES	NO
<i>tabemonó</i>	<i>tabemono=gá</i>	NO	YES
<i>tabemonó</i>	<i>tabemonó=gá</i>	NO	NO

Table 6: Criteria for process of the acquisition of Pattern B

In isolation	With a particle	Lexical accent	Phonological rule
<i>nomimonó</i>	<i>nomimono=gá</i>	YES	YES
<i>nomimonó</i>	<i>nomimonó=gá</i>	YES	NO
<i>nomimóno</i>	<i>nomimonó=gá</i>	NO	YES
<i>nomimóno</i>	<i>nomimono=gá</i>	NO	NO

Table 7: Participants in the Kagoshima survey

		Father		Total
		Local	Nonlocal	
Mother	Local	Group 1: 18	Group 2: 13	31
		Male: 4	Male: 2	Male: 6
		Female: 14	Female: 11	Female: 25
	Nonlocal	Group 3: 16	Group 4: 8	24
		Male: 4	Male: 3	Male: 7
		Female: 12	Female: 5	Female: 17
Total		34	21	55
		Male: 8	Male: 5	Male: 13
		Female: 26	Female: 16	Female: 42

Concerning the acquisition of the phonological rule of pitch shift in prosodic words with a particle, Group 4 (non-local parents) is again the most innovative and Group 1 is the most conservative (Figure 3), and the difference between the two is statistically significant (Tukey test,  $p < 0.001$ ).

Figure 4 compares for each group the ratio of realizations matching the traditional patterns. Group 1 behaves differently compared to the other groups, since there is a better match with the traditional pattern for the phonological rule than for the lexical accent.

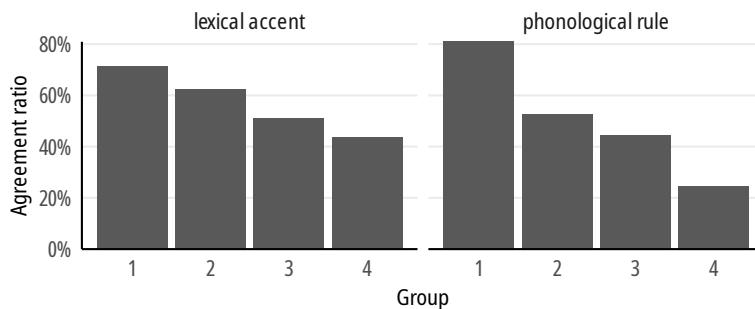


Figure 3: Acquisition of lexical accent and phonological rule

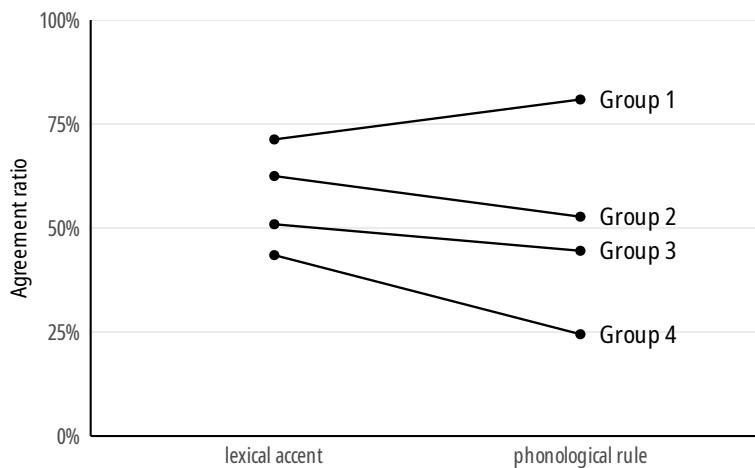


Figure 4: Acquisition of lexical tone vs. acquisition of phonological rule in Kagoshima

The Kagoshima survey therefore shows evidence that parental origins do influence dialect acquisition for both lexical accent and phonological rule. These results agree with the conclusions of Trudgill (1986) on Norwich English and Sugitō & Okumura (1984) on Kansai Japanese. Moreover, the acquisition of phonological rules seems to pose problems for speakers with non-local parents, in agreement with the results of Payne (1976) on Philadelphia English and of Takemura (2010) on Kansai Japanese. It can be hypothesized that those speakers with locally born parents were raised in an environment relatively free of influence from other dialects, which fosters a relatively faithful acquisition of the traditional dialect. On the other hand, speakers with at least one non-local parent were raised in a dialect contact environment, which challenges faithful transmission. Speakers with both native parents nevertheless also show some effects of internal changes in the accent category membership of lexical items, but transmission of phonological rules appears to be stable. Speakers with at least one non-native parent seem to have not fully acquired the traditional tone rule. This trend fits the common knowledge in historical linguistics (Hock 1991) that rules are more resistant to both internal diachronic change (people with locally born parents) and to borrowing (people with at least one non-local parent).

## 5 Conclusions

The present study questioned whether the *haenuki* criterion widely held in Japanese dialectology, i.e., that in order to be considered as representative of a given dialect, a speaker should have native-born parents and grand-parents, is truly justified and thus whether information about a speaker's grandparents is really important and useful for dialect research. We first saw that the *haenuki* criterion poses problems in the context of dialect surveys in large cities since it is often the case that few, if any, speakers satisfy this criterion, so that it is not practically applicable in many real-life situations.

The reexamination of data on lexical accent acquisition across three generations in Hokkaido shows limited transmission between the first generation (grandparents) and the third generation (children) in two families out of three, though there seems to be some influence between adjacent generations, i.e. between grandparents and parents and between parents and children. It is likely that the origin of grandparents is not important in itself, but whether a child was raised with his grandparents (Inoue & Hanzawa 2022). In this case, the information regarding the grandparents' origins would not offer substantial assistance on its own. However, in all the three families included in the Hokkaido survey,

the three generations of grandparents, parents, and children lived together, and the reasons for the limited transmission of accentuation between grandparents and children remain to be investigated. We should also keep in mind that Hokkaido constitutes a peculiar case since it was settled by Japanese speakers only in the second half of the 19th century, and for a long period there was no Hokkaido dialect proper. Data from a region with a traditional dialect might bring different results. However, this does not alter the conclusion that grandparents might not always have a significant impact on dialect transmission. It suggests that incorporating details about grandparents might not always be essential or beneficial when selecting a representative speaker.

The results of the survey on the acquisition of lexical accent and phonological rule in young speakers of Kagoshima Japanese show the determining influence of the parents' origins. Speakers with both native parents show the most faithful acquisition of both the lexical accent and phonological rule of the traditional dialect, and those without a native parent show the least faithful acquisition. Moreover, there is a stark distinction between speakers with both native parents and all the other groups: for all the other groups, speakers show a less faithful acquisition of the phonological rule than that of the lexical accent of words, but the situation is reversed for those speakers whose both parents are native-born. This suggests that the acquisition of phonological rules, i.e., grammar, is strongly influenced by the origins of the parents, much more than lexical acquisition.

Both of the two surveys examined above highlight the role of intergenerational transmission in dialect acquisition, and indicate that we cannot consider a speaker as representative of a dialect solely based on whether they are themselves native-born. When looking at the result of the reexamination of data in Hokkaido, it becomes apparent that a speaker's grandparents' origins have a smaller impact on dialect accent acquisition, more specifically lexical accent acquisition, compared to their parents' origins. Moreover, from the result of Kagoshima survey I observed differences in dialect acquisition between speakers with both native-born parents and all others, even those with one native-born parent and a non-native one, especially in the acquisition of phonological rules. We can conclude that the *haenuki* criterion is not only unpractical but also not fully justified, and that the most important criterion when selecting a representative speaker of a dialect should be to have both parents native-born.

## Appendix

Table 8: Data of the family in Bibai

Generations	Agreement	Disagreement	Total
1-2	443 (87.9%)	61 (12.1%)	504 (100.0%)
2-3	96 (19.9%)	386 (80.1%)	482 (100.0%)
1-3	67 (13.9%)	415 (86.1%)	482 (100.0%)
1-2-3	63 (13.1%)	413 (85.7%)	476 (100.0%)

Table 9: Data of the family in Ikeda

Generations	Agreement	Disagreement	Total
1-2	56 (38.6%)	89 (61.4%)	145 (100.0%)
2-3	113 (57.4%)	84 (42.6%)	197 (100.0%)
1-3	89 (49.2%)	92 (50.8%)	181 (100.0%)
1-2-3	34 (26.8%)	93 (73.2%)	127 (100.0%)

Table 10: Data of the family in Kutchan

Generations	Agreement	Disagreement	Total
1-2	390 (76.9%)	117 (23.1%)	507 (100.0%)
2-3	242 (47.0%)	273 (53.0%)	515 (100.0%)
1-3	161 (32.1%)	340 (67.9%)	501 (100.0%)
1-2-3	150 (29.6%)	357 (70.4%)	507 (100.0%)

Table 11: Agreement for traditional lexical accent in Kagoshima

	Agreement	Total number	Percentage
Group 1	616	864	71.3%
Group 2	390	624	62.5%
Group 3	391	768	50.9%
Group 4	167	384	43.4%

Table 12: Agreement for phonological rule in Kagoshima

	Agreement	Total number	Percentage
Group 1	699	864	80.9%
Group 2	329	624	52.7%
Group 3	342	768	44.5%
Group 4	94	384	24.4%

## References

- Chambers, Jack K. & Troy Heisler. 1999. Dialect topography of Québec City English. *Canadian Journal of Linguistics/Revue canadienne de linguistique* 44(1). 23–48. DOI: 10.1017/S0008413100020703.
- Hock, Hans Henrich. 1991. *Principles of historical linguistics*. 2nd edn. Berlin: Mouton de Gruyter. DOI: 10.1515/9783110219135.
- Inoue, Fumio & Yasushi Hanzawa. 2022. Högen no chiikisa kara neneisa e: Shōnai Hamaogi tsuiseki chōsa no tajū taiō bunseki [From regional differences to age differences in dialect: Multiple correspondence analysis on Shōnai Hamaogi follow-up survey]. *Gengo Kenkyū* 162. 63–89. DOI: 10.11435/gengo.162.0\_63.
- Kubozono, Haruo. 2006. *Akusento no hōsoku* [The laws of accent]. Tokyo: Iwanami Shoten.
- National Language Research Institute (ed.). 1965. *Kyōtsūgo ka no katei: Hokkaidō ni okeru oyako san sedai no kotoba* [From regional differences to age differences in dialect: Linguistic data of three generations in Hokkaidō]. Tokyo: Shūei Shuppan. DOI: 10.15084/00001238.
- National Language Research Institute (ed.). 1981. *Daitoshi no gengo seikatsu: Bunseki-hen* [Linguistic usage in large cities: Analysis]. Tokyo: Sanseidō. DOI: 10.15084/00001261.
- Payne, Arvilla Chapin. 1976. *The acquisition of the phonological system of a second dialect*. Philadelphia: University of Pennsylvania. (Doctoral dissertation).
- Sibata, Takeshi. 1984. Högen chirigaku [Geography of dialects]. In Kiichi Iitoyo, Sukezumi Hino & Ryōichi Satō (eds.), *Högen kenkyūhō* [Dialect study methods], vol. 2, 61–90. Tokyo: Kokusho Kankōkai.
- Sibata, Takeshi. 1988. *Högenron* [Dialectology]. Tokyo: Heibonsha.
- Sugitō, Miyoko & Ayako Okumura. 1984. Oya no högen akusento ga kodomo no akusento-gata no hatsuwa ni ataeru eikyō [Effect of parental dialect accent on children's speech]. *Ōsaka Shōin Joshi Daigaku Ronshū* 21. 1–11.

- Takemura, Akiko. 2010. Dialect acquisition in the view of parental origins: The case of Kansai dialect of Japanese. *Kobe Papers in Linguistics* 7. 78–90. DOI: 10.24546/81001859.
- Takemura, Akiko. 2012. Parental influence on dialect acquisition: The case of the tone system of Kagoshima Japanese. *NINJAL Research Papers* 3. 103–116. DOI: 10.15084/00000492.
- Trudgill, Peter. 1986. *Dialects in contact*. Oxford: Basil Blackwell.
- Uwano, Zendō. 1984. Akusento chōsahō [Methods for accent research]. In Kiichi Iitoyo, Sukezumi Hino & Ryōichi Satō (eds.), *Hōgen kenkyūhō* [Dialect study methods], vol. 2, 229–273. Tokyo: Kokusho Kankōkai.
- Uwano, Zendō. 1997. Fukugō meishi kara mita Nihongo shohōgen no akusento [Accents of Japanese dialects from the viewpoint of compound nouns]. In Tetsuya Kunihiro, Hajime Hirose & Morio Kōno (eds.), *Akusento, intonēshon, rizumu to pōzu* [Accent, intonation, rhythm and pause], 231–270. Tokyo: Sanseidō.
- Yoshida, Norio. 1984. Hōgen chōsahō [Dialect research methods]. In Kiichi Iitoyo, Sukezumi Hino & Ryōichi Satō (eds.), *Hōgen kenkyūhō* [Dialect study methods], vol. 2, 275–299. Tokyo: Kokusho Kankōkai.

## Part IV

# Theoretical approaches and innovations in dialectology



# Chapter 13

## Dialectology as “language making”: Hegemonic disciplinary discourse and the One Standard German Axiom (OSGA)

Stefan Dollinger

UBC Vancouver

This paper problematizes the current anti-pluricentric perspectives in German dialectology in the context of “language making” (Krämer et al. 2022). Disciplinary history and cross-linguistic comparison shed light on what appears to be discipline-internal theoretical hegemony on what makes a language and what a dialect. The paper proposes the existence of a long-standing, discipline-defining One Standard German Axiom (OSGA) to be operative, an axiom that “unmakes” non-dominant standard varieties. It will be shown that, given the unbroken chain of tradition in German dialectology (via, e.g. Kranzmayer or Mitzka) based on Germanic *Stämme* (‘tribes’), the concept of “German language” is *a priori* defined as a stand-alone single entity. A comparison between *Stämme* in German literature – now obsolete – and *Stämme* in German dialectology – still strong – illustrates the far-reaching ramifications of OSGA. Three fail-safes are suggested to move the debate onto an epistemologically sounder footing and to allow for the dynamic, in part predictable, development of multiple linguistic standards in German via Pluricentric Theory (Multi-Standard Theory). Pluricentric Theory remains, it is argued, the theory of choice, though the present paper extends Clyne’s (1995) model with transnational cross-linguistic influence.

### 1 Dialectology as language making: Past and present

While the problem of objectivity in academic inquiry is part and parcel of daily appraisals, it is somewhat of a paradox that questions of a more epistemological

Stefan Dollinger. 2025. Dialectology as “language making”: Hegemonic disciplinary discourse and the One Standard German Axiom (OSGA). in Susanne Wagner & Ulrike Stange-Hundsdörfer (eds.), *(Dia)lects in the 21st century: Selected papers from Methods in Dialectology XVII*, 287–317. Berlin: Language Science Press. DOI: 10.5281/zenodo.15006617



nature are generally considered as lying outside the purview of dialectology. The present paper is atypical in this sense only. While linguistic data is referred to, I focus on a meta-level to explore presuppositions of dialectological fields. *Language making* is a cover term “for processes in situations which largely operate independently of each other but which all contribute to the same effect, to the creation of imagined linguistic units with clear-cut boundaries as ‘a language’” (Krämer et al. 2022: 2); language unmaking, then, is the undoing of such linguistic units (all standard varieties/languages are considered as imagined, human-made units in the present approach, cf. “Regel Nummer 1”, Dollinger 2021: 42).

The concept of “standard” is treated as a conceptual prerequisite for the nameable linguistic standard varieties that are at the core of this paper. The precise elements that such a standard is comprised of, very much in focus in German dialectology, are backgrounded in this paper to allow a bird’s eye view, as it were. The focus of attention is instead bestowed on disciplinary modes of thought, which are *a priori* assumptions, and speakers’ perceptions and language attitudes. Presuppositions such as these, whether expressed or not, are directly related to researchers’ construals of language, variety and dialect and, dependent on one’s viewpoint, one’s (inadvertent) engagement in acts of language making or language unmaking. The present contribution is in this regard no different from any other dialectological contribution on the German language.

Language making is a broad concept that encompasses phenomena from highly diverse yet related areas of research. These include prescriptivism (e.g. Beal et al. 2023, Chapman & Rawlins 2020), standardization (e.g. Ayres-Bennett & Bellamy 2021, Hickey 2012), historical language change (e.g. Watts 2011, Wright 2020), contemporary language change (e.g. Maegaard et al. 2020, Grondelaers & van Hout 2011), language planning (e.g. Joseph et al. 2020, Fishman 2006), linguistic purism (e.g. Langer & Davies 2005, Deumert & Vandenbussche 2003), language attitude and perception studies (e.g. Kircher 2012, De Cillia & Ransmayr 2019) and multilingualism (e.g. Ayres-Bennett & Fisher 2022, Blommaert 2008). In recent years, the effects of language making have received increased attention in sociolinguistics, historical sociolinguistics and dialectology to a degree that the umbrella term *language making* seems useful and indeed warranted.

The basic question, however, is not new and cuts right to the core of dialectology: what is a dialect and what is a language? The relationship of the terms language and dialect, their developments, their near-equivalents and their histories in other languages represents a highly complicated array of meanings and uses (e.g. Van Rooy 2020, Moschonas 2004, Maxwell 2022). Chambers & Trudgill (1998: 4) summarize the problem succinctly when they write that “we need to recognize that, paradoxically enough, a ‘language’ is not a particularly linguistic notion at

all”. In Chambers & Trudgill’s (1998) paradox, I argue, lies a problem with present-day data-based, “bottom-up” computationally-heavy approaches (see Dollinger 2019c: 72–76). While Chambers and Trudgill concede that “linguistic features obviously come into it”, ‘language’ is a notion that is socio-politically formed and, in many contexts, a notion that was formed before the onset of the scientific study of languages and dialects in the early 19<sup>th</sup> century. Considering that the notions in the early 1800s were at best driven by romantic notions of language, at worst by supremacist ideas of nation-making (Hutton 1999), the establishment of the academic fields in modern language from the mid-19<sup>th</sup> century is therefore problematic regarding potentially lingering hegemonic perspectives, which need to be problematized upfront today, if we wish to avoid carrying them forward.

In some disciplines this process of upfront dehegemonization has already begun (Hudley et al. 2024, Costa et al. 2024). For English, Watts (2011), for example, dissects a number of founding myths in the creation of “the language English” that were – and still are – crucial in the construction of that language. The *myth of the longevity of English*, in which the Beowulf manuscript takes central stage, and its interpretation as an “old” text (Watts 2011: 28–52), or the *myth of the unbroken tradition of English*, linking it with the classical languages (Watts 2011: 53–82), and other myths (the *myth of the polite language English*, the *myth of greatness* and the like), are critically assessed and their constructive nature is highlighted. Another example of a discipline-internal presupposition would be that English is considered a Germanic language, and not, for instance, a neo-Romance one, while the linguistic material is ambiguous. The Germanic-family interpretation can be best upheld either by a strict genealogical view of language (once Germanic, always Germanic) or the foregrounding of grammar over vocabulary and a select focus on (the few) grammatical features that were carried over to the present from Proto-Germanic times (e.g. past tense dental suffix, a pronoun case system), not the many that were discarded along the way (e.g. adjectival declension in two ways, active ablaut series, common noun declension, free word order). At this point, data meets interpretation and it is the interpretational frame that decides how the data is being treated and classified. Another way of putting it: It is one’s theory (whether expressed or unexpressed) that frames data interpretations. Present-day textbook knowledge often has its roots in these myths that may have been passed on field-internally as *sine-qua-nons* between generations of researchers.

I argue that contemporary approaches of dialectological disciplines appear to be confined by analogous disciplinary presets and that these perspectives, often inadvertently, work against the interests of some minority groups of speakers. I intend to show with examples from German dialectology that a pluricentric

view of German as a language (variety) with equivalent standard varieties (at least Standard German German, Standard Austrian German, Standard Swiss German, with Standard South Tyrolean German a possibility today) is at a systemic disadvantage in the current German Studies framework. Historically it is a particularly noteworthy and clear example, though the problem extends to a large array of philologies. This disadvantage derives from the potential to contradict directly one of the founding assumptions of German dialectology, which I call the “One Standard German Axiom” (OSGA) (Dollinger 2019b, 2019c: 14, 2024). As such, the present paper goes beyond the recent, vibrant and most welcome focus on standard varieties and their creation (e.g. Beal et al. 2023, Ayres-Bennett & Bellamy 2021) in that it reflects on an epistemological weakness of present-day dialectological approaches to German – approaches that are data-driven and bottom-up – as it seeks to identify unintended continuities of past language making practices with today’s dialectological, generally computational, methods.

In the light of experiences in the early 20<sup>th</sup> century it seems tempting to treat national frames as obsolete. In a globalized world we are indeed faced with transnational phenomena (e.g. Pennycook 2007, Schneider 2023), which need to be incorporated into existing models of language. Any reassessment of philosophical traditions and epistemological underpinnings would need to consider the problematic history of German dialectology in the racial-ethnographic enterprise of the late 19<sup>th</sup> and early 20<sup>th</sup> centuries, and open the modelling of German to allow for cross-language influences and global influences that may begin to resolve a debate over (Standard) Austrian German that has been ongoing since Lewi’s 1875 complaints about the variety (Muhr 2020a).

This paper is organized in three parts. In Section 2, I will briefly sketch the historical genesis of the field of *Deutsche Philologie* (‘German philology’) and *Deutsche Dialektologie* (‘German dialectology’) in relation to Austria in broad historical strokes from Jacob (1785–1863) and Wilhelm Grimm’s (1786–1859) time, via Eberhard Kranzmayer’s (1897–1975) pan-German influence in 20<sup>th</sup>-century Austria, to the present day. This will be accomplished by contrasting a now-obsolete ethnographic approach to German literature – based on *Stamm* ‘tribe’ – with ethnographic interpretations in German linguistics. In Section 3, I will draw attention to epistemological inconsistencies in the dialectological modelling of German. I suggest that Pluricentric Theory (Clyne 1995) is still the model of choice for today’s settings, as it is open for transcultural and transnational innovations. I show that a “pluri-areal” notion is no theoretical concept, but a conceptually empty term whose sole purpose is to uphold the One Standard German Axiom. In Section 4, I will offer three sociolinguistically-inspired principles

as “fail-safes” that would mitigate lingering hegemonic perspectives and undetected legacies deriving from a kind of philology that was once rooted in ideas of supremacy and linguistic superiority.

## 2 Dialectology and German nationalism: The One Standard German Axiom (OSGA)

Chambers & Trudgill (1998) use the example of German to highlight that distinguishing a language from a dialect cannot be carried out by the criterion of mutual intelligibility: “While we would normally consider German to be a single language”, they write from the view of the mid-1990s, “there are some types of German which are not intelligible to speakers of other types”. By contrast, Danish, Swedish and Norwegian, which are “languages with names” (Piller 2017: 51) and are generally granted some form of linguistic autonomy, are often mutually intelligible. These social and political angles of languages are given their due share in “Haugen’s Sequence” (steps to create standard varieties, Haugen 1972 [1964]) and “Weinreich’s Dictum” of a language being “a dialect with an army and a navy” (Weinreich 1945, Dollinger 2023c: 125–126).

The question of “Was ist d/Deutsch?” (e.g. Maas 2018) is therefore a question of prime importance. Nation-making by means of language has been practiced in modern times since the French Revolution (Coulmas 1985) and has become a template for a number of European states. Ideas for German unification in the 18<sup>th</sup> century involved a fierce linguistic debate between an unequal pair of scholars. On the one hand was Johann Siegmund Valentin Popowitzsch [Janez Žiga Valentin Popovič] (1705–1774), a Slovene-German bilingual farmer’s son. Popovič was largely an autodidact and, against the odds, was appointed by Empress Maria Theresia in 1753 as the first university professor of German in Austria (Faninger 1996). On the other hand was the Leipzig University professor, theatre theorist and critic Johann Christoph Gottsched (1700–1766). While in the 1740s and 1750s Popovič proposed a koiné of all German-speaking varieties, including all Austrian ones, Gottsched sought to impose the East Central German (Saxon-based Luther German) as the sole standard in a battle that Gottsched would win with his Viennese followers (e.g. von Sonnenfels, Waldner 2024). As a result, the southern (Austrian) standard was largely surrendered for the northern (East Central German) standard as of the 1770s (Havinga 2018). As a German empress, Maria Theresia’s “Prussification” of the Austrian standard was a logical step towards the political integration of the many scattered Germanies.

However, after Germany was unified in 1871 without the long-term, traditional leader Austria, a situation that was not reverted in 1918 or 1945 – with the temporary exception of the *Anschluss* in 1938–1945 – a certain degree of linguistic autonomy in Austria has always been maintained. While Lewi (1875) considered Austrian Standard German (*Österreichisches Hochdeutsch*) a negative category, by 1926 a philosopher who earned his living as an elementary school teacher in Austria at the time would engage in language making by writing an Austrian dictionary. This philosopher, Ludwig Wittgenstein, intuitively recognized the need for local-made resources and described the scope of his Austrian Dictionary (Wittgenstein 1926) as follows:

In das Wörterbuch sollen nur solche, aber alle solche Wörter aufgenommen werden, die österreichischen Volksschülern geläufig sind. Also auch viele gute deutsche Wörter nicht, die in Oesterreich ungebräuchlich sind. (Wittgenstein 1925: 3)

‘The dictionary should include only those words, but all such words, that are known to Austrian elementary students. Therefore it excludes many a good German word that are unusual in Austria.’ (translation SD)

After World War II the approach Wittgenstein had used was carried forward in the 1951 *Österreichisches Wörterbuch* (ÖWB, ‘Austrian Dictionary’ 1951), which is today available in its 44<sup>th</sup> edition (2025) and has been the best-selling dictionary in Austria for decades. As predicted by Weinreich’s Dictum, newfound national drive – in opposition to Germany – demanded the codification of a new standard. In Austria, as a new standard variety of German, in Luxembourg, as a new language with its own name. From the viewpoint of German dialectology, such decentralization may seem to threaten the integrity of the field. It is therefore not surprising that resistance against the ÖWB has been fierce from the start. In 1958, for instance, the editor of the ÖWB summarizes the critique against ÖWB “as a wholly superfluous, spirit-of-the-times, hostile project that is against the idea of an all-encompassing pan-Germanism”<sup>1</sup> (Krassnigg 1958: 156). It is bizarre, however, to have the *raison-d’être* for a new democratic Austria as independent from Germany be linguistically criticized as not being “German enough”.

To this day, the ÖWB has had no significant academic support from German dialectology and linguistics, which is consistent with one of the founding presuppositions of Germanistik, in which the standard functions as an indispensable unifying force. At the height of Deutschkunde following World War I, German was considered by academic Germanists, with the war lost, as the “last

---

<sup>1</sup> “als ein gänzlich überflüssiges, zeitbedingtes, feindseliges Unternehmen gegen die Sache des Gesamtdeutschstums”

joint property of all Germans” (“unsere Sprache, dem letzten Gemeinbesitz der Deutschen”) (Petersen 1924: 415). This sentiment seems embedded in conceptions of German as a discipline-defining feature, the “One Standard German Axiom” – OSGA (Dollinger 2019c: 14). OSGA is found early and dominantly, e.g. in Jacob Grimm’s *Deutsche Grammatik*, whose title alone – dealing in fact with Germanic languages – illustrates the dominant role assigned to the German language and which stresses the “force” (*Kraft*) of that standard. That sentiment, the “force” of the standard, Grimm links to Luther:

Luthers Verdeutschung der Bibel, die für uns mit jedem Menschenalter kostlicher und zum heiligen Kirchenfest wird (woran geflissentlich kein Wörtchen geändert werden sollte) hat dem Hochdeutschen maennliche Haltung und Kraft gegeben. (Grimm 1819: 6)

‘Luther’s German translation of the Bible, which becomes ever more precious with each generation and has turned into holy communion itself (where under any circumstances even the littlest word must not be changed) has given High German [Hochdeutsch, standard German] masculine posture and strength.’ (translation SD)

It was Grimm’s “declared goal of his life’s work” to show the “unity of the German character in language, mythology, law and custom” (“Die Einheit des deutschen Wesens in Sprache, Mythos, Recht und Sitte”) (Lämmert 1967: 22). Without language there is no German people in Grimm’s view. OSGA underpins all of this.

OSGA is pervasive, for instance, in Eberhard Kranzmayer’s writings, Austria’s most important dialectologist from the early 1930s to the 1970s (Pohl 2006: 397), who reconfirmed conceptions of Austrian dialects as subsumed under the one German standard. Kranzmayer had a profound pan-German orientation throughout his lifetime, pre- and post-war. After he took on race-philological roles in the Third Reich, he lied in both of his denazification proceedings (see Dollinger 2023a) and was therefore later in the position to shape post-WWII German linguistics in Austria (Kronsteiner 2016, Jandl 2022). Kranzmayer considered the existence of a written standard variety as a powerful indicator of the independent status of a people. Until 1945 he performed linguistic land claims for the Third Reich, e.g. by denying the Slovene-speaking populations in Carinthia and Slovenia their own linguistic standard variety of Slovene, making them either “worthy” of Germanization or not (the latter ensuing all the consequences of the Third Reich murder machine) (Dollinger 2023b, 2024). Kranzmayer used the existence of a written standard as proof that the Friuli are a people: “Aus dem Empfinden

der völkischen Eigenständigkeit heraus besitzen die Friauler seit Jahrhunderten eine eigene Schriftsprache” (‘Out of their own racial [*völkisch*] independence the Friuli have had their own written language for centuries’) (Kranzmayer 1943: 2). What is standard and what not may not be the direct focus of German dialectology, but it is an important presupposition that makes or breaks “languages with names” and, in the German tradition, people.

While linguistic land claims of that kind stopped with the end of World War II, the conceptual underpinnings and the importance of the maintenance of OSGA were carried over into the post-war world. I argue that OSGA, as an unreflected default notion, has led to the rejection of pluricentric approaches (e.g. Elspaß et al. 2017: 72–73, Dürscheid & Elspaß 2015, Glauninger 2013: 564) and the recycling of the ad-hoc term “pluri-areal” (Wolf 1994), which has never been defined and which I have shown elsewhere to be synonymous with geographical variation and theoretically empty (Dollinger 2019c: 62–64). The current quest to describe the German standard more realistically (*Gebrauchsstandard*) further reinforces OSGA, as it assumes a priori that newer standards need to be categorically different with “absolute” variants that occur nowhere else (*Homogenitätsgedanke*, Elspaß et al. 2017: 72). The adoption of several areas under one standard, the “pluri-areal” view, I argue, is driven by OSGA. Such a perspective is in contradiction to the founding spirit of sociolinguistics, a discipline that has always argued for speakers of the “smaller” or disadvantaged communities – be they social, such as African American Vernacular English (e.g. Wolfram 1969), or regional, such as Panamanian Spanish (e.g. Cedergren 1973).

It is obvious that German dialectology has been struggling with the idea of several standards of German, which has been considered a “problem” (Ammon 1995). Generally, that standard is left unspecified as the standard of Germany. While in recent years conceptions of what precisely that single standard entails have been foregrounded (*Gebrauchsstandard*), this is done by disregarding the identity-confirming level of national constructs, thus considering international state borders to be irrelevant (e.g. Dürscheid & Elspaß 2015). It seems worthwhile to revisit in greater detail the underpinnings of the construct *German*, for which we start with literary approaches that have become outdated, which will be compared to the dominant dialectological approach today.

## 2.1 “German literature” as a tribal (*stamm-based*) expression

Josef Nadler was professor of German literature in Vienna from before WWI to 1945, practicing, like others at the time, a *völkisch* literary approach (Ranzmaier

2008). He is selected as an early example of a discriminatory approach in German Studies that would become the main – and sole – approach in the Third Reich and that would be, consequently, discarded after 1945. Nadler was part of *Deutschkunde*, a “science” that believed in the mental superiority of all things German – German thought, German poetry, German blood.

*Deutschkunde* gave a number of arguments into the hands of *völkisch* politicians and may be considered as having partially enabled their rise (Hutton 1999: 2–3). Nadler stands out for moulding the then-generally accepted approach into a multi-volume literary history of German (Nadler 1912–1928 [1<sup>st</sup> ed.], Nadler 1938–1941 [2<sup>nd</sup> ed.]). The central concept was German *Stämme* (‘tribes’), which was derived from race-ideological concepts (e.g. Nadler 1934). *Stämme* were considered the locus of all modes of being, expression and innovation. This approach was carried out with the help of

Siedlungsgeschichte, Ortsnamenkunde, Sprachentwicklung, Bodenfunde und Volkskunde [...]. [...] Der Stamm ist das, was es in der Wirklichkeit allein noch gibt, ein familiengeschichtlicher Blutsverband (Nadler 1934: 7–8)

‘Settlement history, toponomastics, language development, archeological finds and European ethnography (*Volkskunde*). The tribe is that alone which in reality still exists, a family-genealogical community of blood relationship.’ (translation SD)

This school claims that each *Stamm* develops, in association with its topography (*Landschaft*), particular characteristics and contributes its share to the *deutsche Volk*. The idea of *Stämme* working as *wechselnde Organe* – ‘complementary organs’ – in the *Volks*-body guarantees that the German people remains self-sufficient and does not need to rely on foreign – *blutsfremd* (‘alien blooded’) – input. Nadler’s express goal was to bestow a kind of “objectivity” on literary and cultural study that had been claimed for German dialectology:

Nicht weniger Philologie, sondern mehr, aber angewandte, Dialektforschung, Stammeskunde, Familiengeschichte, Anthropologie, eine Literaturgeographie, die die Erde nach unseren Bedürfnissen suchend abgeht. Was unsere letzte Sehnsucht sein soll, Anschluß der Geschichte des Schrifttums an die großen Ergebnisse verwandter, fördernder, vorausgesetzter Disziplinen. (Nadler 1912: Band 1: vii–viii)

‘Not less philology, but more, and applied dialectology, tribal (settlement) research, genealogy, anthropology and a literary geography that scours the

earth for our needs. Concerning what we consider our last desire, connecting the history of our literature to the great results of related, supporting and foundational disciplines.' (translation SD)

Dialectology was considered one such *verwandt* – ('related'), *fördernd* - ('supporting') and *vorausgesetzt* – ('foundational') – discipline of literary study in a *völkisch*-racist perspective. The prospect to "search the earth for our needs" reads in the Third Reich context as a threat of "scientifically" approved ethnic cleansing. It was with this approach that Eberhard Kranzmayer wrote studies as director of a "war-decisive" research institute (political think tank) in Klagenfurt (Dollinger 2023a). German dialectology, centered in Marburg, Vienna and Munich, willingly and often enthusiastically offered scholarly support from about WWI to the end of WWII for this purpose (Hutton 1999: 42, Burrell 2023: 66–67, 11, Dollinger 2024). Its concepts of language and standard were presented as inevitable, scientifically proven, and objective, with *Stämme* at their foundation. After WWII, such literary perspectives were corrected; in dialectology however, *Stämme* remained central.

## 2.2 “*Stämme*” in German dialectology: Past and present

The idea of *Stämme* still looms large in dialectology, as visualized in Figure 1. The key point in Figure 1 is that all subvarieties of *Hochdeutsch* – that one Standard German – are non-standard. These varieties are tied to particular German *Stämme*: the Franconians, the Bavarians, with Austrian and Swiss treated, if at all, as *Stämme* themselves, but, more often, in the case of Austria, as comprised of Bavarian and Allemanic *Stämme*, and in the case of Swiss mostly the latter.

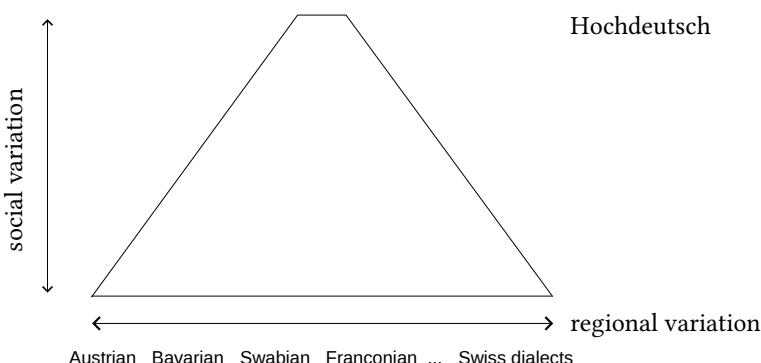


Figure 1: Visualizing the dominant concept of standard German "Hochdeutsch" (Dollinger 2019c: 40)

If Austrian German should be dealt with, Austria would be conceptualized as one such *Stamm* (e.g. Kranzmayer 1956: I: 106), a problem addressed in Muhr (1998).

While German literature was forced to reorient itself after WWII and replace *Stamm*-based perspectives, there was no complete restart and break in German dialectology between what was done before and after. Language study in German did, in particular, not just continue with pre-war personnel (of whom Kranzmayer in Austria – see below – or Mitzka in Marburg are just two prominent NSDAP members), it has not systematically dehegemonized its conceptions of language. Auer (2013: 18), for instance, considers the program of *Germanistik* that has been held together by a national perspective of *deutsche Wissenschaft*, as having been replaced after WWII with something else, something more objective. He considers Lämmert's (1967) half-century old account as “still unsurpassed” (*immer noch unübertroffen*). The trope of “bad” dialectology in the Third Reich and “good” dialectology since does not appreciate the problem, however. If Lämmert et al.'s (1967) book of four essays is now, more than half a century later, still unsurpassed it would mean that the desiderata, one of which being that “German national identity is based on the assumption of a natural linguistic unity” (Lämmert 1967: 34) have not been addressed. Lämmert continues:

Wann – fragt man sich – werden die Deutschen einer Sprachtheorie entsagen, die sie politisch bereits 1866 widerlegten, mit der sie aber danach noch ein Jahrhundert lang fortfuhrten, politische Grenzen je nach Bedarf zu zementieren oder zu negieren? (Lämmert 1967: 34)

‘When – one asks oneself – will the Germans disavow a theory of language, which was politically obsolete in 1866, yet they have continued to operate with for another century, as needed, to buttress or to quash political boundaries?’ (translation SD)

Lämmert (1967: 34) describes OSGA without naming it and sees its rationale in a “particularly rigorous foundation of German national identity in the assumption of a natural linguistic unit of the nation”<sup>2</sup>.

Today, of course, no one would adopt such wording. Many, however, take it for granted that the German language is unified, with one standard, or at least one standard that is more equal than all others, and therefore reify resistance against multiple standards. These assumptions, whether expressed or not, come with

---

<sup>2</sup>“besonders nachdrückliche Fundierung des deutschen Nationalbewußtseins auf der Annahme einer naturgegebenen Spracheinheit der Nation”

heavy historical baggage, as they were developed by philologists who subscribed to a pan-German perspective, in which Austria is a key “possession”, while Luxembourg was not (see Dollinger 2024 for a most extreme case).

The very long-reach of a *Stamm*-based concept in German dialectology is reflected in very modern perspectives. We can see it unintentionally expressed in approaches of current cross-border scenarios with German. One notices in Figure 2 a disparity in the treatment of borders between Netherlandic Dutch (NL) and German (D) in Figure 2a, and the treatment of the Austrian-German border in Bavaria in Figure 2b. Relying on Scheuringer’s (1990a, 1990b) non-quantitative traditional study of that border (see Dollinger 2019c: 42–47), the interpretation seems to unwittingly follow the spirit of OSGA, as Munich and Vienna are heteronomous to a joint, cross-border super-regional standard whose sole legitimacy is the Bavarian *Stamm* and dialect zone from a millennium ago.

Figure 2b therefore shows the long reach of what may be considered a discriminatory concept of language, as Austria and Switzerland appear under OSGA’s umbrella. Pluricentric Theory, however, predicts linguistic diversification across political borders (e.g. Kremer 1979, Boberg 2000), though Scheuringer’s (1990b) qualitative study is not the kind of study that would allow the detection of such diversification as it is too coarse. When, for instance, the geographical dialect continuum between The Netherlands and Germany was disrupted, the “cohesive social system between locations across the state border diverged after the establishment of the state border in the area in 1830, and as a result, so did the dialect variation” (de Vriend et al. 2009: 133). Any other outcome would, in the light of cross-border evidence, be surprising. The same process would be expected to apply at the Austrian-German border.

Historically, German dialectology operated with tropes of domination, which can be seen, for instance, in Anton Pfalz (1927: 58) – Kranzmayer’s superior from 1926 to 1937 at the Vienna dictionary project – who insisted that a difference must be made between dominant and non-dominant *Stämme* and that Saxons and Franks in Austria would be “fully Bavarianized in short periods of time” [“in kurzer Zeit vollständig baiwarisiert”] (Pfalz 1927: 60–61), the dominant *Stamm*. The theme of dominance was big in *völkisch* linguistics. The dominant *Stämme* win and dominate others and therefore further improve the German race, the argument goes. The idea of *Stämme* is central and Germanistik has “frequently improved” *deutsche Volkskunde* in the “spirit of the Grimm Brothers” (Pfalz 1927: 62). While this kind of discourse has stopped, the frame of OSGA carries over this trope of domination in which one standard “tames” all other varieties.

It needs to be noted that there is no structural difference between Nadler’s application of *Stämme* to literature and the application of *Stämme* to language,

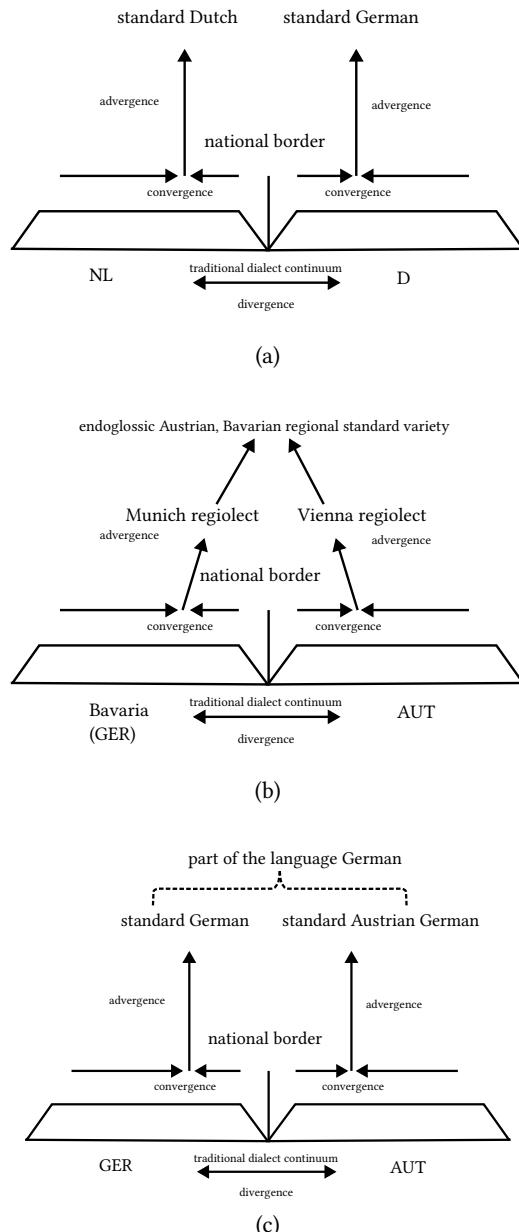


Figure 2: The NL-D (after Auer 2005: 21; 2a) and AUT-GER borders (Auer 2005: 27; 2b); Dollinger's (2019c: 54–58) reinterpretation (2c)

shown in Figures 1 and 2b. *Stämme* continue to be treated as relevant in linguistics and the varieties of the various *Stämme* are considered hyponyms to Standard German, under the OSGA umbrella. While historical accounts speak of Old High German as having had “no uniform linguistic entity”, “but a series of more or less clearly demarcated dialects that are lumped together under the umbrella term [Old High German]”<sup>3</sup> (Ernst 2021: 76), current conceptualizations of present-day German refer to these dialect zones and foreground them in interpretations: “It is frequently seen that [present-day] variants surprisingly clearly coincide with the old dialect zones; occasionally there are unexpected alignments”<sup>4</sup> (Elspaß et al. 2017: 73). Elspaß’ interpretations such as the one above leave the *Stamm*-based model intact and conceptually and constructively unchallenged. In Third Reich diction, Nadler considered *Stämme* as the immutable *sine qua non*: “There is no race history in the humanities that does not take note of the indisputable fact of *Stämme*”<sup>5</sup> (Nadler 1934: 18). While very different in application, the basic, field-defining construct of *Stämme* remains unchallenged and only the language historian occasionally refers to the arbitrary linguistic demarcation that has become disciplinarily, conceptually enshrined and has, as Elspaß’ interpretations illustrate, become the default expectation today.

Hutton (1999: 78) points out that treating *Stämme* as objective analytical units has advantages. It “preserves disciplinary history intact, for it relegates Nazism to the non-scientific realm, to that of the methodologically aberrant”. German dialectology has not been willing to engage in this disciplinary argument (see, e.g. Langer’s (2021) review of Dollinger 2019c) and clearly prefers a “pluri-areal”, non-pluricentric approach to German. When today someone argues that Austria is no single dialect zone (95% Bavarian dialects, but some Alemannic too), and therefore there cannot be an Austrian Standard German (e.g. Elspaß & Niehaus 2014, Herrgen 2015, Elspaß 2020: 50), they unwittingly engage in acts of language making that reconfirms the conservative focus on one standard of German under OSGA.

---

<sup>3</sup>“keine einheitliche Sprachform”, “sondern aus einer Reihe von mehr oder weniger deutlich voneinander abgegrenzten einzelnen Dialekten bestand, die unter diesem Oberbegriff zusammengefasst werden.”

<sup>4</sup>“Allerdings zeigt sich häufig, dass ... Varianten erstaunlich deutlich mit den alten Dialekträumen übereinstimm[en], hin und wieder ergeben sich unerwartete Allianzen.”

<sup>5</sup>“Es kann keine geistesgeschichtliche Rassenkunde geben, die nicht von dem unabweisbaren Tatbestand der Stämme Kenntnis nehme.”

### 3 Pluricentric theory plus three fail-safes

*Stamm*-based approaches are static conceptions of German that, while permitting the tweaking of the one standard, do not allow for the development of newer standards. In its guise as “pluri-areal” (Elspaß et al. 2017), they sound similar to pluricentric theory, yet as basically *Stamm*-based ideas of the German language they cannot effectively adapt to developments since the 1800s, e.g. such as transcultural influences. A single standard for German, often in the term *standard German*, remains the linguistically dominant perspective at the beginning of the 2020s.

Linguists may, for instance, document respondents’ doubt in Standard Austrian German rather than interpreting doubts as expressions of linguistic insecurity (e.g. Koppensteiner & Lenz 2020), a concept well established in English dialectology (Preston 2013), a concept essential in non-dominant varieties such as Austrian German, yet a concept not used in German dialectology. Researchers may unwittingly engage in acts of language unmaking and become non-consenting agents of OSGA via the disciplinary legacy of German dialectology. For instance, in the Swiss context Dürscheid (2009: 60) considers the standard of Germany as the usual “benchmark”, not Standard Swiss German. Senior Marburg researchers proclaim an entirely new era in German language history, without recourse to linguistic insecurity, and see “de-nationalization” as “a new phase for the standard norm” (Herrgen 2015: 156). Pluricentricity, however, is considered as “an entirely political concept” (Elspaß & Niehaus 2014: 50), while the political dimension of the notion “German” is not addressed. Instead, bottom-up aggregate approaches render the state border a-priori as irrelevant. Methodologically, a dozen pre-defined areas comprise the German-speaking lands, stripping the national identity level of its relevance by demanding categorical differences on each side of the national border (e.g. Elspaß 2020: 52).

What is methodologically reified is the construct German (it is what is fed the algorithm); what is voided are younger standard varieties such as Austrian or Swiss German, nipping Standard South Tyrolean German pre-emptively in the bud. An insensitivity to the social salience of variables leads to assessments such as “[l]ess than two per cent of variation in standard German ... does hardly make a ‘variety’” (Elspaß & Niehaus 2014: 50), while, in fact, it is up to the speakers to decide how to view their code. Such statements should be seen as blatant statements of language making and unmaking, not as disinterested analysis. For good reason, Ruck (2023: 383) pleads for the inclusion of a language political angle in German studies curricula that would, among other things, reveal such disciplinary biases. Approaches such as Muhr (2017) or De Cillia & Ransmayr

(2019), clearly in a minority today in the light of “Deutsch in Österreich” Special Research Clusters, are aware of language political imbalances and aim to mitigate linguistic insecurity in Austria.

As such, the 2020s are in stark contrast to a generation earlier, when Michael Clyne (1995), Peter von Polenz (1994) and Ulrich Ammon (1995) brought pluricentric concepts for a short time into the mainstream of German dialectology. A logical consequence of ignoring the autonomy of Standard Austrian German (cf. Chambers & Trudgill 1998: 9–12) and of supporting anti-pluricentric approaches in the 2010s is Muhr’s (2021) call for an “Austriazistik” – Austrian Studies – that is independent of *Germanistik* – German Studies. Which model of language would be most conducive to accommodate the changing nature of present-day varieties? It is likely not the monocentric model in Figure 1, unchanged in its conceptions for centuries, and stuck in a past where one German-speaking state seemed possible, indeed desirable. This is the “pluri-areal” model.

The model of language that seems open for expansion and adaptation is, however, the pluricentric model. It is the default setting in English linguistics (with particular focus on World Englishes) and the model that was applied by Clyne (1995) to German. The term *center* is to be understood in a metaphorical manner, not a precise location. Auer (2021) suggests the term “multiple-standard language” rather than pluricentric language in order to avoid this confusion. It is useful to visualize the pluricentricity – multiple standards – of German schematically in “Trudgill’s pyramids” as follows (Figure 3).

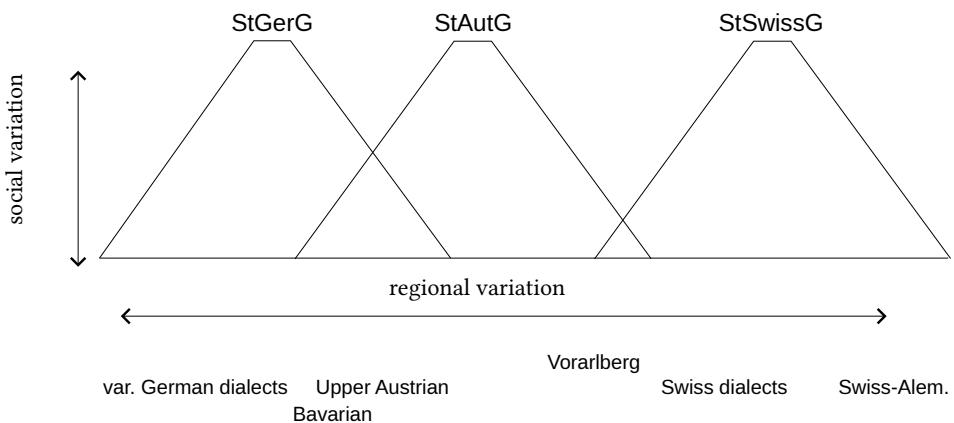


Figure 3: Visualizing multiple standards of German (modeled after Trudgill’s (2000) pyramid concept; Dollinger 2019c: 39)

On the horizontal plane, the model shows regional non-standard variation. This feature overlap – often used as an argument against Standard Austrian Ger-

man – is built into the pluricentric model. It predicts, for instance, that the regions of Bavaria in Germany and Upper Austria share more non-standard features with one another than Vorarlberg and Upper Austria. It does not employ, however, the millennium-old level of *Stamm*-based dialects and thus avoids the unmaking of younger standards, such as the Austrian or Swiss standards. Social variation is depicted on the model’s vertical axis, variation that may also be situationally dependent: Speakers may use Standard Austrian German where the context requires it (top of pyramid), dialects when appropriate (base of pyramid), or intermediate forms (*Umgangssprache* ‘colloquial language’) somewhere in between.

The pluricentric model is open to modifications. For instance, the influence of varieties cross-linguistically (e.g. Buschfeld & Kautzsch 2020, Schneider 2022), can be shown visually as a cross-bar that covers all three varieties. Transnational influences from additional contexts and varieties may be depicted by bars of different colours cutting across a particular pyramid, a pair, or all three. For instance, as global English terms spread via socially upper or intermediate strata, a “bar of Global English” (e.g. *die Compliance, das Tool* from business German, or *chillen, slayen, texten, SMSen* from youth language) cut across the upper middle of all three pyramids without, and this is key, dissolving each pyramid’s reference point of an autonomous standard (Figure 4).

Adaptations specific to a particular variety may also be visualized in different shapes of the pyramid. For instance, the wider functional range of Swiss dialects as used in TV talk shows or on radio compared to Austrian (and German) ones may be reflected in a broader x-dimensional box of the dialects on the y-axis, as shown in Figure 4. Influences from Germany on the smaller standards (e.g. <-ig> pronounced as /iç/ and not as /ig/, *Eimer* for *Kübel* ‘bucket’) may be depicted by arrows spreading out from Germany into Austria and Switzerland. Vice versa, reverse influence (e.g. *es geht sich (nicht) aus* – an AutG expression denoting ‘it suffices / does not suffice’) would be shown by, perhaps thinner because less forceful, arrows. These adaptations are possible without diminishing the identity-affirming frame of standard varieties for some (small) nations, as they maintain national priming effects, which may also serve as identity markers. They visualize Schneider’s (2022) suggestion of “epicenters” in one country that may influence other centers but do not hegemonize smaller standards and frames by virtue of differential speaker numbers (e.g. 83 million Germans vs. 9 million Austrians).

The model can be extended when new standards emerge, as may happen with South Tyrolean German once it features a codified dictionary (Hofer 2020). The pluricentric model also allows the maintenance of reference frames that function

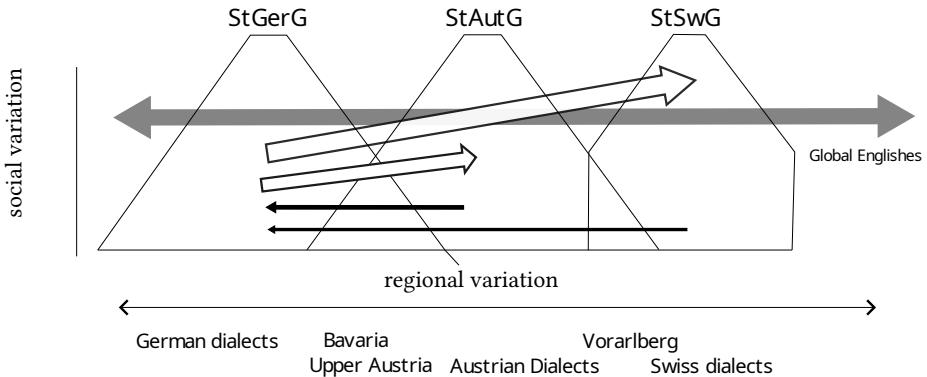


Figure 4: Pluricentric, multi-standard model of present-day German, with inter-language and global influences visualized by arrows

as identity markers in international contexts. While dialects and *Umgangssprache* (colloquial language) are identity markers domestically or in some cross-border contexts (e.g. Upper Austria and Bavaria), in the international or EU context the standard varieties take on this function (e.g. De Cillia et al. 2020 for language and identity in Austria).

## 4 Three fail-safes against linguistic hegemony

Three “fail-safes” are suggested to mitigate the inadvertent effects of linguistic hegemonic discourse in German dialectology. They would allow the detection of obsolete language conceptions for present-day speakers. They would support the implementation of Lämmert’s half-century-old suggestion to discard the traces of linguistic domination in the concept of German. They may be used in connection with the pluricentric model to see where further refinement is needed. The three fail-safes (Dollinger 2019b, 2019c: 107–114) are rooted in the goal of increasing the relevance of dialectology and linguistic study for as many speakers as possible while avoiding the sociolinguistic language unmaking of non-dominant (Austria, Switzerland), perhaps not fully developed (e.g. South Tyrol), standards, or their demotion to one of 15 all-German-language regions (e.g. Dürscheid & Elspaß 2015).

### 4.1 The speaker is always right (fail-safe #3)

In addition to linguistic production data, language attitudinal data needs to be considered, even foregrounded. De Cillia & Ransmayr (2019: 137, Abb. 36) show

that in Austria 80% of secondary school students, and 90% of their teachers, perceive “more than one standard in German”. This pluricentric perspective is currently not appropriately reflected in German dialectology. From a sociolinguistic perspective, the current anti-pluricentric approach violates fail-safe #3, which is “the speaker is always right” (Dollinger 2019c: 113–114). That is, the *informed* speaker, who has been briefed about the construction of standard varieties and their reifying nature, is always right. As De Cillia & Ransmayr’s (2019) results show, the need for Standard Austrian German is not diminishing. Such results are in clear contradiction to the Marburg School’s verdict of the “de-nationalization [*Entnationalisierung*] of German” (Herrgen 2015: 157), which is based on a methodologically questionable assessment of Austrian listeners ranking the standard speaker from Germany a tiny bit more “pure Hochdeutsch” than the Austrian one (Herrgen 2015: 154) (see for a detailed critique Dollinger 2019c: 78–84).

De Cillia & Ransmayr’s (2019) results, in contrast, may be considered a mandate for *Germanistik* to reconsider the relevance of its language models for speakers beyond Germany. It would imply that we need to, in Labovian ways, mitigate linguistic insecurity appropriately and inform speakers of the possibilities. For instance: “You don’t need to try to sound like someone from Hanover when you speak standard.” If after such information the speaker prefers a Northern German model, we can rest assured that our data is valid. The speaker is, after all, indeed always right, and the primary problem here is to assess which majority of speakers is given credence. Paramount is the sociohistorical embedding, which, in the Austrian context, is one of ridicule and linguistic insecurity.

Current practice is rather the opposite, however, as the bar for a newer Austrian standard is set unreasonably high by German dialectologists. When, for instance, Elspaß (2020: 66) states that a given feature is dominant but not “absolute” in Austria – meaning, in a crude dichotomy, that the feature is not *exclusively* used in Austria – and that thus there cannot be a Standard Austrian German, the bar for acceptance of Standard Austrian German is lifted so high that very few established languages on earth would ever clear it.

#### 4.2 Work with explicit, falsifiable theories (fail-safe #2)

Fail-safe #2 (Dollinger 2019c: 111–113) calls for explicit, falsifiable theories. Pluricentricity is such a theory, as, for instance, it makes predictions of how language variation across political borders develops. It predicts, among other things, that the varieties on either side of an arbitrary political boundary will diverge depending on a list of factors (e.g. Auer 2005: 29). It predicts that there will be,

often considerable, feature overlap (e.g. Bavaria and Upper Austria in Figure 2), yet the same or near-identical features will be framed differently, depending on which side of the border one is positioned (geographically, socially, attitudinally and cognitively).

“Pluri-areal” creed, by contrast, offers no theory; it does not make falsifiable predictions. It is synonymous with unspecified geographical variation and equivalent with the term “anti-pluricentric” (Dollinger 2019b: 103–106, 2019c: 62–176). This distinction is important, because the statement that German is a “pluri-areal” language, a language with geographical variation, cannot be falsified. As that view does not specify in which ways this variation occurs, it cannot make predictions; it cannot do more than to negate pluricentric theory. Paradoxically, “pluri-areal”, however, is presented as on a par with pluricentric theory and, in most cases, as superseding the pluricentric approach (e.g. Dürscheid & Elspaß 2015, Elspaß et al. 2017, Langer 2021, DiÖ 2021, Meer & Durgasingh 2025).

Pluricentric predictions include, for instance, that linguistic production will diversify with the duration of the political division at a given border. They are testable and falsifiable. In this context, I would like to correct the assessment that pluricentric approaches “lend themselves to nationalistic views and exploitations because they are interested predominantly in linguistic perception rather than production” (Schneider 2022: 470–471). While perception and the cognitive frame of speakers certainly play an important role in Pluricentric Theory, production is not unimportant, though it is not the sole criterion – as it is in most “pluri-areal” studies – for lumping varieties together as a language. The proof that anti-pluricentric perspectives are not theoretically anchored can be taken via Karl Popper’s epistemological concept (explored in Dollinger 2019c: 89–90), or with Lewin’s (1952: 169) *bon mot* that “there is nothing more practical than a good theory.” This practical angle of pluricentric theory – as applied to Standard Austrian German – can be seen, for instance, in German as a Second Language teaching in Austria, where German-made material is often irrelevant, or, at least, inappropriate. A first-hand report is from a German Beginner Textbook by Hueber (Dollinger 2021: 59–60), which in the first week was dealing with a difficult phonetic distinction – <ä> vs. <e> or /ɛ/ vs. /e/ – that is near-categorically absent in Standard Austrian German and even in Germany a stylistic and not a phonemic distinction. Likewise, professional translators and interpreters universally understand and appreciate the necessity of the concept of Standard Austrian German, as seen in recent movies such as *Rush* (2013) which portrays the life of Formula 1 star Niki Lauda, with German actor Daniel Brühl, playing the Austrian Lauda, getting lessons on SAutG, thus rendering the movie quite authentic.

### 4.3 Uniformitarian Hypothesis: Vertical and horizontal readings (fail-safe #1)

From a sociolinguistic angle perhaps the most powerful argument for the consequent adoption of a pluricentric perspective in mainstream German dialectology can be made via the Uniformitarian Hypothesis. Fail-safe #1 (Dollinger 2019c: 109–111) calls for theorizing to be in line with both the vertical and horizontal readings of the Uniformitarian Hypothesis (e.g. Labov 1994: 21). What I call the *vertical* reading of the Uniformitarian Hypothesis is common practice in historical linguistics. Linguists assume, all other things being equal, that the language change of the past must have proceeded in a similar manner to the change of the present. The vertical, the diachronic, reading is universally accepted as one of the theoretical anchors of historical linguistics.

In addition to this well-known vertical (diachronic) reading, I would like to formalize a synchronic, a horizontal reading. The horizontal reading of the Uniformitarian Hypothesis (Dollinger 2019c: 110–111) can be described as follows: If two languages have comparable cross-border situations, codifications and social scenarios, they *must* be modeled the same way unless there are salient differences. Such equivalent situations exist, for instance, for American/Canadian English, Dutch/German or Austrian German/German German. Recent models of Netherlandic Standard Dutch and Belgian Standard Dutch (e.g. De Ridder 2020) underscore via horizontal uniformitarian comparison the distinction between Austrian Standard German and German Standard German. In other words: What applies to the sociolinguistic goose, must also apply to the sociolinguistic gander. If we model younger standard varieties, such as Canadian English or Belgian Dutch, in one way, with two autonomous standard varieties, we *must* model the Austrian-German context in like manner.

Table 1 is a comparison of basic benchmarks in Austria and Canada regarding the non-dominant status of German and English. With only some key benchmarks showing, it gives an example of the horizontal reading of the Uniformitarian Hypothesis. The social conditions for codification were fully present after World War II in both countries (Row A). At this point, Canada became de-facto independent from the UK (issuing their own passports as of 1947). Austria had by that time overcome the Third-Reich unification with Germany which resulted in a renewed focus on Austrian dimensions. Row (B), how many speakers recognize a distinct variety, is important, with clear majorities in both countries around the 2010 polling dates. This language awareness is a process and in 1945, the answers would have looked different.

Table 1: Sociolinguistic benchmarks for assessment with the Uniformitarian Hypothesis, horizontal reading

		CanE	AutG
A)	Conditions for codification	After WWII	After WWII
B)	How many speakers see a distinct variety?	80% (Dollinger 2019a: 234)	80–90% (De Cillia & Ransmayr 2019: Abb. 36)
C)	Linguistic insecurity today	Somewhat; decreasing	Considerable
D)	Codification	First general dictionary 1937, full-size dictionaries 1962–1967	Lewi (1875), Wittgenstein (1926) <i>Österreichisches Wörterbuch</i>
E)	Academic interest	Increasing	Decreasing, but substantial for “German in Austria”
F)	History	Canada as colony until 1866	Austria leader of German states until 1866

In light of Table 1, an argument would need to be brought forth *against* the younger standard varieties of Austrian German and Canadian English today and not *for* them. Figure 2b is therefore remarkable as it breaks, without a discernible rationale, with the horizontal reading of the Uniformitarian Hypothesis and suggests that Austrian and German represent some sort of special case. A *Sonderweg* seems to be invoked by those German linguists who address the issue (e.g. Glauninger 2013), while most seem to unwittingly apply a treatment in violation of the horizontal reading of the Uniformitarian Hypothesis and in compliance with OSGA.

## 5 Conclusion

The suggested three fail-safes combined would considerably reduce lingering disciplinary hegemonic perspectives that constitute some form of language unmaking. They would curb cyclical interpretations of data – studying German means that any approach confirms the existence of one German (OSGA). Since 2010, the rhetoric against pluricentricity has gotten markedly more aggressive (see for a summary Dollinger 2021: 139–159). In comparison to other dialectological fields such as Portuguese, followed by English, Spanish, and even French (where skepticism to multiple standards has a long tradition, see Oakes 2001), German dialectology is today the *most* conservative in its language modeling (see

Dollinger 2023a). Multiple standards of German are, for instance, downplayed as irrelevant (e.g. Besch & Wolf 2009: 245), denied their existence (Stephan Elspaß, quoted in Muhr 2020b: 10, Elspaß & Niehaus 2014: 50), or at least put into question (Koppensteiner & Lenz 2020: 48,74). I have attempted to show that such questions are in the last instance, harking back to Lämmert (1967), the unintended result of undetected field-internal pan-German frames of reference. The question is, therefore, to what degree German dialectology is still influenced by underlying, pan-German presuppositions today. If the *Germanistik*-internal resistance against Standard Austrian German, a trope in the field since at least 1951 (e.g. Muhr 2020a: 500–502), is taken as a measure, the influence is considerable and warrants its own name, for which I suggested the “One Standard German Axiom” (OSGA). OSGA has been an underlying assumption, a driving force in German dialectology and linguistics since Grimm’s day, whether explicit (until about 1960), or implied (since then).

From such a backdrop, it is clear why language teachers of German outside of Germany understand the *sine qua non* of Pluricentric Theory (e.g. Callies & Hehner 2023). Ruck (2021), in German didactics, is a positive review in obvious contradiction to Langer’s (2021) utmost dismissal of Dollinger (2019c). Pluricentricity is, using Lewin’s phrasing, a most practical theory – a theory that applies, as any linguistic theory should, cross-linguistically. The long-standing resistance against Standard Austrian German is deep-rooted in German linguistics. It applies a threshold that keeps being raised, in its most recent reiteration, by Elspaß, for example. If in 1800 German had had that bar as high as it is now, there would be no such language today. Standard varieties do not just emerge bottom-up, they are also shaped top-down. German dialectology has been one top-down player in the language making process, as are national education boards, national broadcasters and a range of other agents.

I have argued that a One Standard German Axiom has been at the core of work in German dialectology – a presupposition that goes back to pan-German thought. While no present-day linguist has anything to do with this period or thinking, it was a different matter for some of their teachers and teachers’ teachers (e.g. the cases of Kranzmayer or Mitzka). This continuity in German linguistics has already been reconstructed from the disciplinary record, e.g. Hutton (1999: 1–2), who looked at the connection of German linguistics and anthropology. While today there are clear differences to the past in a number of important ways, the answer to lingering conceptual continuities seems to lie in field-internal presuppositions that have been passed down through generations of researchers, from the Humboldtian tradition (which “puts prime emphasis on language as a defining national identity”, Hutton 1999: 263), via Weisgerber (the

post-WWII “redemption of the German people through their language”, Hutton 1999: 143), in the form of a priori assumptions concerning one standard in German. As such, the One Standard German Axiom is an expression of a process of disciplinary language (un)making that has real-life consequences for the speakers of Austrian German and other non-dominant varieties (Dollinger 2021: 155–159). The case for Standard Austrian German is strong. If German dialectology continues to fail to register the relevance of attitudinal findings and speakers’ cognitive identity constructions, it remains conceptually stuck in its pan-German past. De-hegemonization might start with a reformulation of the a priori assumption of how many standards of German there should be presently. The suggested answer is: more than one.

## References

- Ammon, Ulrich. 1995. *Die deutsche Sprache in Deutschland, Österreich und der Schweiz: Das Problem der nationalen Varietäten*. Berlin: de Gruyter.
- Auer, Peter. 2005. The construction of linguistic borders and the linguistic construction of borders. In Markku Filppula, Juhani Klemola, Marjatta Palander & Esa Penttilä (eds.), *Dialects across borders: Selected papers from the 11th International Conference on Methods in Dialectology (Methods XI), Joensuu, August 2002, 3–30*. Amsterdam: John Benjamins. DOI: 10.1075/cilt.273.03aue.
- Auer, Peter. 2013. Über den *Topos* der verlorenen Einheit der Germanistik. *Zeitschrift für Literatur und Linguistik* 43(172). 16–28.
- Auer, Peter. 2021. Reflections on linguistic pluricentricity. *Sociolinguistica* 35(1). 30–47.
- Ayres-Bennett, Wendy & John Bellamy (eds.). 2021. *The Cambridge handbook of language standardization*. Cambridge: Cambridge University Press.
- Ayres-Bennett, Wendy & Linda Fisher (eds.). 2022. *Multilingualism and identity*. Cambridge: Cambridge University Press.
- Beal, Joan, Morana Lukač & Robin Straaijer (eds.). 2023. *The Routledge handbook of linguistic prescriptivism*. Abingdon: Routledge.
- Besch, Werner & Norbert Wolf. 2009. *Geschichte der deutschen Sprache: Längsschnitte – Zeitstufen – Linguistische Studien*. Berlin: Erich Schmidt Verlag.
- Blommaert, Jan. 2008. *Grassroots literacy: Writing, identity and voice in Central Africa*. London: Routledge.
- Boberg, Charles. 2000. Geolinguistic diffusion and the U.S.-Canada border. *Language Variation and Change* 12. 1–24.
- Burrell, Courtney M. 2023. *Otto Höfler’s characterisation of the Germanic peoples: From sacred men’s bands to social daemonism*. Berlin: de Gruyter.

- Buschfeld, Sarah & Alexander Kautzsch (eds.). 2020. *Modelling world Englishes: A joint approach to postcolonial and non-postcolonial Englishes*. Berlin: de Gruyter.
- Callies, Marcus & Stefanie Hehner (eds.). 2023. *Pluricentric languages and language education: Pedagogical implications and innovative approaches to language teaching*. Abingdon: Routledge.
- Cedergren, Henrietta. 1973. *The interplay of social and linguistic factors in Panama*. Cornell University. (Doctoral dissertation).
- Chambers, Jack K. & Peter Trudgill. 1998. *Dialectology*. 2nd edn. Cambridge: Cambridge University Press.
- Chapman, Don & Jacob D. Rawlins (eds.). 2020. *Language prescription: Values, ideologies and identity*. Bristol: Multilingual Matters.
- Clyne, Michael G. 1995. *The German language in a changing Europe*. Revised edn. Cambridge: Cambridge University Press.
- Costa, James, Daniela Lauria, Beatriz Lorente & Zorana Sokolovska. 2024. Historical foundations: Some threads for integrating and interrogating historiography in Critical Sociolinguistics. In Alfonso del Percio & Mi-Cha Flubacher (eds.), *Critical Sociolinguistics: Dialogues, dissonances, developments*, 37–56. London: Bloomsbury Academic.
- Coulmas, Florian. 1985. *Sprache und Staat*. Berlin: de Gruyter.
- De Cillia, Rudolf & Jutta Ransmayr. 2019. *Österreichisches Deutsch macht Schule?* Wien: Böhlau. <https://austria-forum.org/web-books/en/osterreichdeutsch00de2019isds>.
- De Cillia, Rudolf, Ruth Wodak, Markus Rheindorf & Sabine Lehner. 2020. *Österreichische Identitäten im Wandel: Empirische Untersuchungen zu ihrer diskursiven Konstruktion 1995-2015*. Wiesbaden: Springer VS.
- De Ridder, Reglindis. 2020. Dutch national varieties in contact and in conflict. In Rudolf Muhr, Josep Àngel Mas Castellas & Jack Rueter (eds.), *European pluricentric languages in contact and conflict*, 65–79. Berlin: Peter Lang.
- de Vriend, Folkert, Charlotte Giesbers, Roeland van Hout & Louis ten Bosch. 2009. The Dutch-German border: Relating linguistic, geographic and social distances. *International Journal of Humanities and Arts Computing* 2: *Computing and language variation*. 119–134. DOI: 10.1515/9780748641642-009.
- Deumert, Ana & Wim Vandenbussche (eds.). 2003. *Germanic standardizations*. Amsterdam: John Benjamins.
- DiÖ. 2021. *Statement regarding the lecture “Österreichisches Deutsch oder Deutsch in Österreich? Über einen Problemfall, seit 1848, der Wissenschaftsgeschichte” held on August 25, 2021 by Prof. Dr. Stefan Dollinger*. Dollinger’s lecture can be found at [https://www.youtube.com/watch?v=IMf6pji\\_LfQ](https://www.youtube.com/watch?v=IMf6pji_LfQ); Dollinger’s re-

- sponse to the statement can be found at <https://www.academia.edu/62390449/>.  
<https://www.dioe.at/artikel/3034>.
- Dollinger, Stefan. 2019a. *Creating Canadian English: The professor, the mountaineer, and a national variety of English*. Cambridge: Cambridge University Press.
- Dollinger, Stefan. 2019b. Debunking “pluri-areality”: On the pluricentric perspective of national varieties. *Journal of Linguistic Geography* 7(2). 98–112.
- Dollinger, Stefan. 2019c. *The pluricentricity debate: On Austrian German and other Germanic Standard varieties*. London: Routledge.
- Dollinger, Stefan. 2021. *Österreichisches Deutsch oder Deutsch in Österreich? Identitäten im 21. Jahrhundert*. Wien, Hamburg: new academic press.
- Dollinger, Stefan. 2023a. Afterword: Who is afraid of pluricentric perspectives? In Marcus Callies & Stefanie Hehner (eds.), *Pluricentric languages and language education: Pedagogical implications and innovative approaches to language teaching*, 217–221. Abingdon: Routledge.
- Dollinger, Stefan. 2023b. Eberhard Kranzmayer’s Deutschtum: On the Austrian dialectologist’s pan-German frame of reference. *Journal of Austrian Studies* 56(3). 63–89.
- Dollinger, Stefan. 2023c. Prescriptivism and national identity: Sociohistorical constructionism, disciplinary blindspots, and Standard Austrian German. In Joan Beal, Morana Lukač & Robin Straaije (eds.), *The Routledge handbook of linguistic prescriptivism, Vol. 1*, 121–139. Abingdon: Routledge.
- Dollinger, Stefan. 2024. Eberhard Kranzmayer’s dovetailing with Nazism: His fascist years and the “One Standard German Axiom (OSGA)”. *Discourse & Society* 36(2). 147–179. DOI: 10.1177/09579265241259094.
- Dürscheid, Christa & Stephan Elspaß. 2015. Variantengrammatik des Standarddeutschen. In Roland Kehrein, Alfred Lameli & Stefan Rabanus (eds.), *Regionale Variation des Deutschen: Projekte und Perspektiven*, 563–584. Berlin: de Gruyter.
- Dürscheid, Christa. 2009. Variatio delectat? Die Plurizentrität des Deutschen als Unterrichtsgegenstand. In Monika Clalüna & Barbara Etterich (eds.), *Deutsch unterrichten zwischen DaF, DaZ und DaM*, 59–69. Stallikon: AkDaF.
- Elspaß, Stephan. 2020. Areal variation and change in the phraseology of contemporary German. In Elisabeth Piirainen, Natalia Filatkina, Sören Stumpf & Christian Pfeiffer (eds.), *Formulaic language and new data: Theoretical and methodological implications*, 43–78. Berlin, Boston: de Gruyter.
- Elspaß, Stephan, Christa Dürscheid & Arne Ziegler. 2017. Zur grammatischen Pluriarealität der deutschen Gebrauchsstandards – oder: Über die Grenzen des Plurizentritätsbegriffs. *Zeitschrift für deutsche Philologie* 136. 69–91.

- Elspaß, Stephan & Konstantin Niehaus. 2014. The standardization of a modern pluriareal language: Concepts and corpus designs for German and beyond. *Orð og Tunga* 16. 47–67.
- Ernst, Peter. 2021. *Deutsche Sprachgeschichte*. 3rd edn. Wien: Facultas.
- Faninger, Kurt. 1996. *J. S. V. Popowitz: Ein österreichischer Grammatiker des 18. Jahrhunderts*. Bern: Peter Lang.
- Fishman, Joshua A. 2006. *Language loyalty, language planning and language revitalization: Recent writings from Joshua A. Fishman*. Nancy H. Hornberger & Martin Pütz (eds.). Clevedon: Multilingual Matters.
- Glauninger, Manfred. 2013. Deutsch im 21. Jahrhundert: »pluri«-, »supra«- oder »postnational«? In Ingeborg Fiala-Fürst, Jürgen Joachimsthaler & Walter Schmitz (eds.), *Mitteleuropa: Kontakte und Kontroversen*, 459–468. Dresden: Thelem.
- Grimm, Jacob. 1819. *Deutsche Grammatik, Erster Theil*. Göttingen: Dieterichsche Buchhandlung.
- Grondelaers, Stefan & Roeland van Hout. 2011. The Standard language situation in the low countries: Top-down and bottom-up variations on a diaglossic theme. *Journal of Germanic Linguistics* 23(3). 199–243.
- Haugen, Einar. 1972. Dialect, language, nation [1964]. In Evelyn S. Firchow (ed.), *Studies by Einar Haugen*, 496–509. Berlin, Boston: de Gruyter Mouton.
- Havinga, Anna D. 2018. *Invisibilising Austrian German*. Berlin: de Gruyter.
- Herrgen, Joachim. 2015. Entnationalisierung des Standards: Eine perzeptions-linguistische Untersuchung zur deutschen Standardsprache in Deutschland, Österreich und der Schweiz. In Alexandra N. Lenz & Manfred M. Glauninger (eds.), *Standarddeutsch im 21. Jahrhundert: Theoretische und empirische Ansätze mit einem Fokus auf Österreich*, 139–164. Göttingen: Vandenhoeck und Ruprecht.
- Hickey, Raymond (ed.). 2012. *Standards of English: Codified varieties around the world*. Cambridge: Cambridge University Press.
- Hofer, Silvia. 2020. *Deutsch ist nicht gleich Deutsch: Zum Umgang mit der pluri-zentrischen Sprache Deutsch und standardsprachlicher Variation an Südtiroler Oberschulen*. University of Vienna. (Doctoral dissertation).
- Howard, Ron. 2013. *Rush*. Produced by Andrew Eaton, Eric Fellner, Brian Oliver, Peter Morgan, Brian Grazer and Ron Howard. Published by Universal Pictures (United States), StudioCanal (United Kingdom), Universum Film (Germany).
- Hudley, Anne H. Charity, Ignacio L. Montoya, Christine Mallinson & Mary Bucholtz. 2024. Conclusion: Decolonizing linguistics. In Anne H. Charity Hudley, Christine Mallinson & Mary Bucholtz (eds.), *Decolonizing linguistics*, 445–463. Oxford: Oxford University Press.

- Hutton, Christopher M. 1999. *Linguistics and the Third Reich*. London, New York: Routledge.
- Jandl, Marco. 2022. Die Germanistik in Graz in der Nachkriegszeit. In Heimo Halbrainer, Susanne Korbel & Gerald Lamprecht (eds.), *Der „schwierige“ Umgang mit dem Nationalsozialismus an österreichischen Universitäten: Die Karl-Franzens-Universität Graz im Vergleich*, 157–194. Graz: Clio.
- Joseph, John E., Gijsbert Rutten & Rick Vosters. 2020. Dialect, language, nation: 50 years on. *Language Policy* 19. 161–182.
- Kircher, Ruth. 2012. How pluricentric is the French language? An investigation of attitudes towards Quebec French compared to European French. *Journal of French Language Studies* 22(3). 345–370.
- Koppensteiner, Wolfgang & Alexandra N. Lenz. 2020. Tracing a standard language in Austria using methodological microvariations of verbal and matched guise technique. *Linguistik Online* 2/20(102). 47–82.
- Krämer, Philipp, Ulrike Vogl & Leena Kolehmainen. 2022. What is “language making”? *International Journal of the Sociology of Language* 274. 1–27.
- Kranzmayer, Eberhard. 1943. *Das Volk der Friuler: Kleine Schriften des Instituts für Kärntner Landesforschung*. Klagenfurt: Reichspropagandaamt Kärnten.
- Kranzmayer, Eberhard. 1956. *Historische Lautgeographie des gesamtbairischen Dialektraumes: Mit 27 Laut- und 4 Hilfskarten in besonderer Mappe*. Wien: Böhlau.
- Krassnigg, Albert. 1958. Das Österreichische Wörterbuch. *Muttersprache* 68. 155–157.
- Kremer, Ludger. 1979. *Grenzmundarten und Mundartgrenzen*. Köln: Böhlau.
- Kronsteiner, Otto. 2016. Kranzmayer, Eberhard. In Katja Sturm-Schnabl & Bojan-Ilija Schnabl (eds.), *Enzyklopädie der slowenischen Kulturgeschichte in Kärnten/Koroška. Von den Anfängen bis 1942 : J – Pl*, vol. 2, 693–694. Wien: Böhlau. <https://austria-forum.org/web-books/koroska02de2016isds>.
- Labov, William. 1994. *Principles of linguistic change*, vol. 1: Internal Factors. Malden: Blackwell.
- Lämmert, Eberhard. 1967. Germanistik: Eine deutsche Wissenschaft. In Eberhard Lämmert, Walther Killy, Karl Otto Conrady & Peter von Polenz (eds.), *Germanistik: Eine deutsche Wissenschaft. Beiträge von Eberhard Lämmert, Walther Killy, Karl Otto Conrady und Peter v. Polenz*, 7–42. Frankfurt am Main: Suhrkamp.
- Lämmert, Eberhard, Walther Killy, Karl Otto Conrady & Peter von Polenz (eds.). 1967. *Germanistik – eine deutsche Wissenschaft: Beiträge von Eberhard Lämmert, Walther Killy, Karl Otto Conrady und Peter v. Polenz*. Frankfurt am Main: Suhrkamp.

- Langer, Nils. 2021. Stefan Dollinger. 2019. The pluricentricity debate. On Austrian German and other Germanic Standard varieties (Routledge Focus). Abingdon: Routledge. 137 S. *Zeitschrift für Rezensionen zur Germanistischen Sprachwissenschaft* 13(1–2). Response at <https://www.academia.edu/45577124/>, 2–9. DOI: 10.1515/zrs-2020-2060.
- Langer, Nils & Winifred V. Davies (eds.). 2005. *Linguistic purism in the Germanic languages*. Berlin: de Gruyter.
- Lewi, Hermann. 1875. *Das Österreichische Hochdeutsch: Versuch einer Darstellung seiner hervorstechendsten Fehler und fehlerhaften Eingenthümlichkeiten*. Wien: Bermann & Altmann.
- Lewin, Kurt. 1952. *Field theory in social science: Selected theoretical papers*. Dorwin Cartwright (ed.). London: Tavistock Publications.
- Maas, Utz. 2018. Was ist *deutsch* oder was ist *Deutsch*? Fragen an die Sprachgeschichte. *Zeitschrift für Deutsche Philologie* 137(1). 73–128.
- Maegaard, Marie, Malene Monka, Kristine Køhler Mortensen & Andreas Candefors Stæhr (eds.). 2020. *Standardization as sociolinguistic change: A transversal study of three traditional dialect areas*. Abingdon: Routledge.
- Moschonas, Spiros A. 2004. Relativism in language ideology: On Greece's latest language issues. *Journal of Modern Greek Studies* 22(2). 173–206.
- Maxwell, Alexander. 2022. Noam Chomsky and the language/dialect dichotomy. *Beiträge zur Geschichte der Sprachwissenschaft* 32. 72–98.
- Meer, Philipp & Ryan Durgasingh. 2025. *Modeling variation: Pluricentricity and pluriareality – The debate surrounding both models, and potentials for their complementarity*. Vol. 2. Amsterdam: Benjamins.
- Muhr, Rudolf. 1998. Die Wiederkehr der Stämme – Gemeinschaftlichkeitsentwürfe via Sprache im Europa der neuen sozialen Ungleichheit – Dargestellt am Beispiel des Österreichischen Deutsch. In Bernhard Kettemann & Rudolf de Cillia (eds.), *Sprache und Politik: Verbal-Werkstattgespräche*, 30–55. Frankfurt am Main: Peter Lang.
- Muhr, Rudolf. 2017. Das Österreichische Deutsch. *Zeitschrift für deutsche Philologie* 136. 23–41.
- Muhr, Rudolf. 2020a. Eine kurze politische Geschichte des Österreichischen Deutsch. In Thomas Walter Köhler, Christian Mertens & Anton Pelinka (eds.), *Ein Hauch von Welt: Österreich vor und nach Saint Germain*, 499–524. Wien: Braumüller.
- Muhr, Rudolf. 2020b. Pluriareality in sociolinguistics: A comprehensive overview of key ideas and a critique of linguistic data used. In Rudolf Muhr & Juan Thomas (eds.), *Pluricentric theory beyond dominance and non-dominance: A critical view*, 9–78. Graz: PCL-Press.

- Muhr, Rudolf. 2021. Überlegungen zur Errichtung einer eigenständigen Austriazistik. In Hans-Joachim Solms & Jörn Weinert (eds.), *Deutsche Philologie? Nationalphilologien heute: Sonderheft zum Band 139* (Sonderhefte der Zeitschrift für deutsche Philologie), 125–146. Berlin: Erich Schmidt Verlag.
- Nadler, Josef. 1912. *Literaturgeschichte der deutschen Stämme und Landschaften*, vol. I. Band: Die Altstämme (800–1600). Regensburg: J. Habbel.
- Nadler, Josef. 1934. Rassenkunde, Volkskunde, Stammeskunde. *Dichtung und Volkstum [Euphorion]* 35. 1–18.
- Nadler, Josef. 1938–1941. *Literaturgeschichte des deutschen Volkes: Dichtung und Schrifttum der deutschen Stämme und Landschaften*. 2nd edn., vol. 1-4. Berlin: Propyläen-Verlag.
- Oakes, Leigh. 2001. *Language and national identity: Comparing France and Sweden*. Amsterdam: Benjamins.
- ÖBV im Auftrag des Bundesministeriums für Bildung (ed.). 1951. *Österreichisches Wörterbuch*. Wien: Österreichischer Bundesverlag.
- Pabst, Christiane M. (ed.). 2025. *Österreichisches Wörterbuch: Vollständige Ausgabe mit dem amtlichen Regelwerk*. 44th edn. Wien: Österreichischer Bundesverlag.
- Pennycook, Alastair. 2007. *Global Englishes and transcultural flows*. London: Routledge.
- Petersen, Julius. 1924. Literaturwissenschaft und Deutschkunde. *Zeitschrift für Deutschkunde* 38(6). 404–415.
- Pfälz, Anton. 1927. Angeblich fränkische Mundarten in Österreich. *Oberdeutsche Zeitschrift für Volkskunde* 1(1). 54–62.
- Piller, Ingrid. 2017. *Intercultural communication: A critical introduction*. 2nd edn. Cambridge: Cambridge University Press.
- Pohl, Heinz-Dieter. 2006. Hornung, Maria (b. 1920). In Keith Brown (ed.), *Encyclopedia of language & linguistics*, 2nd edn., 397–398. Amsterdam: Elsevier.
- Preston, Dennis R. 2013. Linguistic insecurity forty years later. *Journal of English Linguistics* 41(4). 304–331.
- Ranzmaier, Irene. 2008. *Stamm und Landschaft: Josef Nadlers Konzeption der deutschen Literaturgeschichte* (Quellen und Forschungen zur Literatur- und Kulturgeschichte 48 (282)). Berlin, New York: de Gruyter.
- Ruck, Julia. 2021. Dollinger, Stefan: The pluricentricity debate. On Austrian German and other Germanic Standard varieties. London, UK: Routledge 2019. *ÖDaF-Mitteilungen: Fachzeitschrift für Deutsch als Fremd- und Zweitsprache* 37(1). 150–153.
- Ruck, Julia. 2023. How can German Studies become a site of resistance to linguistic indifference in language policy regimes? *German Quarterly* 96. 383–385.

- Scheuringer, Hermann. 1990a. Bayerisches Bairisch und österreichisches Bairisch: Die deutsch-österreichische Staatsgrenze als Sprachgrenze? In Ludger Kremer & Hermann Niebaum (eds.), *Grenzdialekte*, 361–381. Hildesheim: Georg Olms Verlag.
- Scheuringer, Hermann. 1990b. *Sprachentwicklung in Bayern und Österreich: Eine Analyse des Substandardverhaltens der Städte Braunau am Inn (Österreich) und Simbach am Inn (Bayern) und ihres Umlandes*. Hamburg: Buske.
- Schneider, Edgar W. 2022. Parameters of epicentral status. *World Englishes* 41. 462–474.
- Schneider, Edgar W. 2023. All things new in Singapore: On creativity, complexity, and usage associations in English. *English Today* 39(1). 24–34.
- Trudgill, Peter. 2000. *Sociolinguistics: An introduction to language and society*. 4th edn. Harmondsworth: Penguin.
- Van Rooy, Raf. 2020. *Language or dialect? The history of a conceptual pair*. Oxford: Oxford University Press.
- von Polenz, Peter. 1994. *Deutsche Sprachgeschichte vom Spätmittelalter bis zur Gegenwart. Band III. 19. und 20. Jahrhundert*. Frankfurt am Main: de Gruyter.
- Waldner, Gernot. 2024. Hier spricht die Aufklärung: Zur Sprachreform des Joseph von Sonnenfels. In Lydia Rammerstorfer, Gernot Waldner & Christian Wolf (eds.), *In Wien um 1800*, 189–205. Vienna: Böhlau.
- Watts, Richard. 2011. *Language myths and the history of English*. New York: Oxford University Press.
- Weinreich, Max. 1945. Der yivo und di problemen fun undzer tsayt. *Yivo Bleter* 1. 3–18.
- Wittgenstein, Ludwig. 1925. *Geleitwort zum Wörterbuch für Volksschulen*. [http://wittgensteinsource.org/WFV/Ts-205\\_f](http://wittgensteinsource.org/WFV/Ts-205_f). Unpublished preface, dated 22.04.1925.
- Wittgenstein, Ludwig. 1926. *Wörterbuch für Volks- und Bürgerschulen*. Wien: Hölder-Pichler-Tempsky.
- Wolf, Norbert. 1994. Österreichisches zum Österreichischen Deutsch. *Zeitschrift für Dialektologie und Linguistik* 61(1). 66–76.
- Wolfram, Walt. 1969. Linguistic correlates of social differences in the negro community. In James Alatis (ed.), *20th annual round table: Linguistics and the teaching of Standard English To speakers of other languages and dialects* (Georgetown Monograph Series on Languages and Linguistics 22), 249–257. Washington, D.C.: Georgetown University Press.
- Wright, Laura (ed.). 2020. *The multilingual origins of Standard English*. Berlin: de Gruyter Mouton.



# Chapter 14

## „Es werden im wesentlichen [sic!] nur Wörter aufgenommen, welche deutlich unterschiedlich zum Hochdeutschen sind.“: On the verticality of lay dialect collections and the attempt to measure it

Yvonne Kathrein

University of Innsbruck, Department of German Studies, Tyrolean Dialect Archive

Dialect collections created by laypersons not only provide information about the different conditions under which they were created and the associated intentions, but also about how laypersons conceptualize their dialect and try to delimit it “upwards”. Using the example of three selected collections from the Austrian province of Tyrol, the present study shows not only which entries are generally present there, but also how dialectal they are – taking into account the underlying dialect – in comparison to the standard language. In doing so, the individual linguistic levels serve as a template according to which the deviations are classified.

It turns out that the collections have about the same amount of special dialect lexis (i.e., lexis that occurs only in the dialects) and about the same amount of entries that differ from the standard language only at the phonological and/or morphological level. However, differences appear in the number of stylistically marked elements and in the calculated  $d$ -values.

Furthermore, the study shows that the evaluations of the collections can not only be refined by perceptual dialectological methods, but also that the collections can be a good complement to existing perceptual dialectological studies.



## 1 Introduction

In recent years we have noticed an increasing number of online dialect collections created by laypersons. These word lists with “translation”, often compiled in years of work by several dialect-speaking and -interested persons, mostly have the purpose to save dialectal words and phrases from being forgotten. But also – from the laypeople’s point of view – the special characteristics of the respective dialect are documented in this way.

Such collections have received little scholarly attention, not least because their compilation does not stand up to scientific criteria: The expertise of the collectors is based primarily on the fact that they are dialect speakers of the dialect being documented. This also means that the entries in the collections were selected mainly by self-observation rather than by word research based on a fixed questionnaire: The collectors themselves considered which words of a particular dialect might be worth recording.

I am not talking about dialect dictionaries that were compiled by experts such as the “Wörterbuch der Bairischen Mundarten in Österreich (WBÖ)”, the “Bayrisches Wörterbuch (BWB)” or the “Schweizerisches Idiotikon”; all of them are long-term projects with a duration of more than a hundred years (cf. Stöckle 2020, Schnabel et al. 2020, Landolt & Roth 2020). Of course, laymen were and are significantly involved in their compilation, e.g. by responding to the calls of the respective dictionary commissions and answering the questionnaires sent to them, sometimes even over a period of years. (cf. Dollmayr & Kranzmayer 1963: vii–viii, Landolt & Roth 2020: 144, Denz et al. 2002: v, xi–xiv) A more recent method of involving the public is to respond to online questionnaires. (cf. Retti 1999, Hofer & Meier 2015). In any case, however, standardized questionnaires developed by scientists are used, among other instruments, to collect material.

This does not apply to those collections that will be discussed here: I am talking about collections (and not dictionaries) that have been created in bottom-up processes involving (almost) exclusively linguistic laypersons, i.e., people interested in local history or people enthusiastic about their dialect, either as individuals or as a group. And they have collected their material, as I said, not by (standardized) questionnaires, but by self-observation, which has led to more or less mature collections without any claim to completeness (on the qualitative spectrum cf. Section 2 as well as Eickmans 1980).

But it is precisely this incompleteness, this non-systematicity, this non-scientificity, this lay idea of one’s own dialect that is of interest. And I see this in the interest of perceptual dialectology. Although the question of how laypeople conceptualize dialect(s) is central to this relatively new field of dialectology, lay

dialect collections have not been considered so far. This is surprising because such collections involve issues of salience as well as the spatial extension of dialects. And these, in turn, are central concepts of perceptual dialectology (for the spatial, horizontal dimension cf. e.g. Diercks 1988, Auer 2004, Lameli et al. 2008, Anders 2008, Preston 2010, Hundt 2010, Schröder 2019; for salience cf. e.g. Lenz 2010, Purschke 2011, 2014, Auer 2014, Anders et al. 2014, Elmentaler et al. 2010, Palliwoda & Schröder 2016, Hettler 2018).

However, the spatial dimension to which perceptual dialectology has been devoted so far needs to be expanded, in my opinion: It is not only about the spatial, namely the horizontal, extension of dialect(s), but also about how laypeople understand the vertical dimension. And here, layperson's dialect collections are an almost predestined source: They have to deal with the question whether individual entries are still dialect at all or whether they already belong to a more widely used, supra-regional, colloquial variety – perhaps even a standard variety. And thus, they also touch on a question that has existed since the first lay linguistic studies, namely the question of the interweaving of “subjective” and “objective” material, i.e. the material obtained by laypeople and the material processed by scientists (cf. Preston 1999).

The present study is thus a first attempt to approach these lay collections from what I understand, with reference to the above remarks, to be a perceptual dialectological standpoint. It aims at evaluating a sample of three inner-Alpine online lay collections with regard to a) their content and b) their dialectality, or in other words, it addresses the question to what extent the entries differ from the standard language and how one could methodically proceed to measure these deviations.

## 2 Dialect collections of laypersons

As outlined above, the dialect collections of laypersons are currently experiencing a renaissance, at least it seems so, as we can observe a large number of them appearing on the Internet (but also in printed form) presenting different types of data, namely:

- the local vocabulary of geographic units such as
  - a single place (e.g. *Außervillgraten*<sup>1</sup> or *Lustenau*<sup>2</sup>),

---

<sup>1</sup><https://ausservillgraten.tirol.gv.at/kultur/villgraterisch> (Accessed 21 November, 2022)

<sup>2</sup><https://www.lustenau.at/media/209/download/mundartdatei.pdf?v=1> (Accessed 21 November, 2022)

- a valley (e.g. Zillertal<sup>3</sup> or Vinschgau<sup>4</sup>)
- or a larger historically developed region (e.g. Burgenland<sup>5</sup>, Bavaria<sup>6</sup> or the former region Baden<sup>7</sup>),

or they present

- the vocabulary of a linguistic unit (e.g. “Fränkisch”<sup>8</sup>, “Bodenseealemannisch”<sup>9</sup>)

or

- a combination of both (e.g. “Hegau-Alemannisch”<sup>10</sup>)

as well as

- the vocabulary of (former) groups (e.g. Walser<sup>11</sup>).

The collections are accessible on the Internet in different ways: They exist either as a stand-alone website or are integrated as a subpage in a parent page. These higher-level pages can in turn be very different. They range from official community websites to sites of clubs and customs groups, tourism associations, music groups, hotels and guesthouses, etc., up to personal sites. They can be found as a separate element in the sitemap or – somewhat hidden – in a blog post or as a pdf. Although search engines react differently to searches for such collections, in most cases the search leads quickly to a result, even if the name of the page reflects the dialectal pronunciation of the respective dialect (e.g. “Sainihánserisch” for the dialect of St. Johann in Tyrol (Austria), “Muntafuner Dialekt” for the dialect of the valley Montafon in Vorarlberg (Austria) or “Soorser Wöörterbüechli” for the dialect of Sursee (Switzerland)).

---

<sup>3</sup><https://www.zillertaler-woerterbuch.com/> (Accessed 21 November, 2022)

<sup>4</sup>[http://www.chrizia.com/c\\_ling\\_wbvd\\_abc.htm](http://www.chrizia.com/c_ling_wbvd_abc.htm) (Accessed 21 November, 2022)

<sup>5</sup><https://mundart-burgenland.at/woerterbuch.php?lang=hianzisch&char=A> (Accessed 21 November, 2022)

<sup>6</sup><https://www.bavariastore.de/bayerische-mundart/> (Accessed 21 November, 2022)

<sup>7</sup><https://freiburg-schwarzwald.de/alemannisch/badisch-deutsch.htm#Badisch%20-%20Deutsch%20-%20%C3%9Cbersetzer> (Accessed 21 November, 2022)

<sup>8</sup><https://aus-meinem-kochtopf.de/fraengisch-werdderbuch-woerterbuch-fraenkisch-deutsch/> (Accessed 21 November, 2022)

<sup>9</sup><https://www.alemannisch.de/eip/pages/seealemannisch.php> (Accessed 21 November, 2022)

<sup>10</sup><https://www.staff.uni-mainz.de/pommeren/Miszellen/Alemannisch.html> (Accessed 21 November, 2022)

<sup>11</sup><http://www.walser-alps.eu/mundart/woerterbuch> (Accessed 05 December, 2022)

As can be concluded from the various operators of the websites, the target audience is equally diverse. It can be locals who speak the respective dialect themselves, it can be people who are interested in the respective dialect in general, but it can also be (temporary) immigrants or guests. On the one hand, these collections are intended to inform them about the respective dialect and its peculiarities. This also includes enabling the smoothest possible communication between “ingroup” and “outgroup” (whether it is really possible or not<sup>12</sup>). On the other hand, self-expression, the presentation of the curious, the incomprehensible, the unpronounceable, the comical is also more or less present; this is especially the case with those collections that are specifically aimed at tourists.

As for the history of their creation, the collections are the work of a single person or, more often, the work of several people who compile idioms/dialectalisms over a certain period of time (which can be years). They usually have predecessor collections as a basis on which new words and expressions are added. Publication on the Internet sometimes makes it possible to participate in the collection if one knows the dialect in question. This also means that the authorship is not always traceable.

It goes without saying that due to this heterogeneous process, the collections vary in length, quality, content and presentation. At the macro level, the vocabulary is usually presented in two columns (as in a vocabulary book), with a bar at the top or bottom of the page that allows navigation between letters (see Figure 1). Furthermore, navigation can be enabled by a search function (see Figure 2). However, the content can also be organized into topics, similar to an onomasiological dictionary (see Figure 3).

At the micro level, entries may simply consist of the dialect word or phrase with a more or less strict transcription – usually using at least some diacritics – and the translation without any other additional information (see Figure 1). They may contain grammatical information and/or syntactic context if this seems necessary, as well as audio and/or visual material (see Figure 2), sometimes even information on (presumed) etymology.<sup>13</sup> In addition, the collections are often supplemented by dialectal idioms, rhymes and phrases, and sometimes by local names.

---

<sup>12</sup>For example, there is an extensive collection of words used in (Upper) Austria to facilitate communication between foreign and local students: [http://www.fim.uni-linz.ac.at/Woerterbuch\\_oesterr\\_deut\\_englisch.htm](http://www.fim.uni-linz.ac.at/Woerterbuch_oesterr_deut_englisch.htm) (Accessed 5 December, 2022)

<sup>13</sup>For example in a collection called “Allgäuer Wörterbuch” ([https://www.dein-allgaeu.de/regionen/regionen\\_woerterbuch\\_a.html](https://www.dein-allgaeu.de/regionen/regionen_woerterbuch_a.html)) or the “Wörterbuch ‘Schriftdeutsch – Alemannisch’” ([http://hausen.pcom.de/kultur\\_bildung/olschowka/w%C3%B6bu\\_schrda\\_alem\\_start.htm](http://hausen.pcom.de/kultur_bildung/olschowka/w%C3%B6bu_schrda_alem_start.htm)) (both accessed 5 December, 2022)

## GRIAS ENK BEIN ORIGINAL ZILLACHTOLER WEACHTABUACH!

STARTSEITE DIALEKT BRÄUCHE **WÖRTERBUCH** ESSEN & TRINKEN VON A BIS Z WÖRTERBUCH SPECIALS  
GALERIE FREUNDE IM ZILLERTAL KONTAKT



Wörterbuch » A

### A

a	ein; in
alluane	allein(e)
ameacht	schon einmal, früher
appotegg'n	Apotheke
auffa	herauf, hinauf
aufgeig'n	Geige spielen; geigen
auf'n	hinauf
außaklauraum	(heraus-) nehmen
auss'n	draußen
Auss(a)rgab	Retourgeld

A B D E F G H I J K L M N O P R S T U V W Z

Essen & Trinken

Figure 1: The collections may be quite rudimentary such as the “Original Zillachtoler Weachtabuach” [“Original dictionary of the Ziller valley”, author’s translation]

The screenshot shows a website for a dialect collection. At the top, there is a navigation bar with links to 'Lexikon', 'Sprüche', 'Reime', 'Lieder', and 'Anekdoten'. Below this is a search bar with the placeholder 'Suchbegriff eingeben:' and a 'Suchen' button. To the right of the search bar is a video player with a play button and a timestamp of '0:00'. Below the video player is a section titled 'Sonderzeichen' with definitions for 'á', 'é', and 'ö'. The main content area is titled 'Lexikon' and contains an alphabetical bar from A to Z. Under the bar, there are two columns: 'dialekt' and 'hochdeutsch'. The 'dialekt' column lists words like 'a', 'a da', 'a nan - a nan ánan', 'Aa', 'Aachzeit', 'Abort', 'acha - kim acha', and 'Achéta - áchétzn'. The 'hochdeutsch' column provides their standard German translations. Below the table, there is a link to a quiz about Mundartausdrücke. The website has a background image of a mountain range and a logo in the bottom left corner.

Figure 2: A search function that allows searching for dialect words as well as the Standard German translation, an alphabetical bar, a two-column outline, audio examples and information on transcription: The St. Johann in Tirol collection is in many parts more sophisticated than others (<https://www.sainihanserisch.at/lexikon.html>)

This is not an exhaustive description of the different forms of presentation and contents of the individual collections. The description merely gives a small insight into the variety and diversity of amateur dialect collections.

### 3 Methods: Preparation of a corpus and evaluation procedure

#### 3.1 Choosing the collections

For this paper, three lay collections from the Austrian province of Tyrol were selected (see Figure 4). All of them are accessible via the Internet.

They were selected according to their geographical distribution and their information content: The dialects should be clearly distinguishable from each other (see Figure 5 for the dialect regions) and the collections should allow at least a broad phonetic transcription of the individual entries by providing audio material.

Mundart

## VOSCHTEICKILATS UND DOWISCHILATS

7. April 2016 durch HPV Sexten 0 Kommentare 618 Views

Home / Mundart / Voschteickilats und Dowischilats

Ein Beitrag von *Regina Senfter Stauder*

Endlich **Langis!** Als der harte Winter die Kinder wieder „freigab“ und man allmählich auf die wärmere Zeit zusteuerte, konnte das kleine Volk nichts mehr im Haus halten. Draußen war wieder überall fröhliches Jauchzen zu vernehmen. Mit den Kindern der Nachbarhäuser trafen wir uns zu sozialen Spielen. „**Spiel mo Voschteickilats odo Dowischilats?**“, lautete die Frage fast täglich. „**Haint tiomo Voschteickilats.**“ Rund ums Bauernhaus gab's genug Verstecke und wer als Letzter **ogiprettli wourdn isch**, musste als Nächster suchen. Besonders spannend war das Versteckspiel auf den Feldern, wenn viele **Roggla, Schwednraita** oder **Herflan** dastanden. So manches Heumännchen wurde dabei von uns Kindern im Elfer des Gefechtes zu Boden gerissen. Wenn die Lust sich auszutoben uns regelrecht in den Beinen kitzelte, dann wurde **Dowischilats / Leischtl** gespielt. Durch Auszählverse wurde bestimmt, wer als erster fängt. Der bekannteste war wohl: **Ellile, sellile, siggile, sa, rippile, rappile, knoll**. Frei wie die Vöglein waren wir. Ungezwungen konnten wir springen, uns entspannen und selber Gefahren ausloten. Körperliches und seelisches Gleichgewicht trainierten wir in höchstem Maße. Wir wussten, wie man auf Bäume klettert, auf Zäunen balanciert und Bäume an gefährlichen Stellen überquert. Der Spannungsbogen der Natur war uns vertraut – vom völlig Harmlosen bis zum äußerst Gefährlichen – von der köstlich schmeckenden Walderdbeere bis zur giftigen Kreuzotter. Lernen mit Erdung und unter dem Himmelszelt war die Devise. Theoretische Kenntnisse nutzten wir mit Kopf und Fuß. Sind nicht gerade das die heute so viel zitierten und mundstrapazierte Kompetenzen, welche es zu erreichen gilt? Wir hatten keinen dicht gedrängten Terminkalender wie so manches Kind von heute, waren an den Nachmittagen nicht auch noch in Institutionen gepflicht und kannten keine Konzentrationsprobleme oder Hyperaktivität. Niemand hat uns „fast zu Tode gefördert“. Frei waren wir, frei zu denken und zu handeln – ja, einfach frei, ohne ständige Kontrolle von irgendeiner Seite. Selbst mussten wir Verantwortung übernehmen, selbst durften wir bestimmen, wann wir müde waren.

**Wir hatten fast nichts (an Spielzeug), wir brauchten auch nichts und doch hatten wir alles. Es lebe die Kindheit, das Fundament fürs Leben!**

Spiele aus Großmutter's Zeit:

Spiele im Haus	
Guggamuine	Blinde Kuh
Gurflschpiel: Fohr af Grazz	Spiel mit Murmeln
Di Soge, do Knouche, do Gotto...	Fadenspiele mit den Händen
Spiele im Freien	
Dowischilats /Leischtl	Fangen
Voschteickilats	Verstecken
Kaschtl hupfn/ Amaslam	Tempel hüpfen
hutschn	schaukeln
Schällilie paudn	mit Zapfen und Reisig Ställchen bauen
Afn Schrällilan mit an Brette au und o raitn	wippen

Figure 3: The local vocabulary presented in subject areas, here “Voschteickilats und Dowischilats” [“hide and catch”, author's translation], introduced by a text about children's games on the website of the monument protection association of Sexten (South Tyrol/Italy) (<http://heimatpflege-sexen.eu/voschteickilats-und-dowischilats/>)

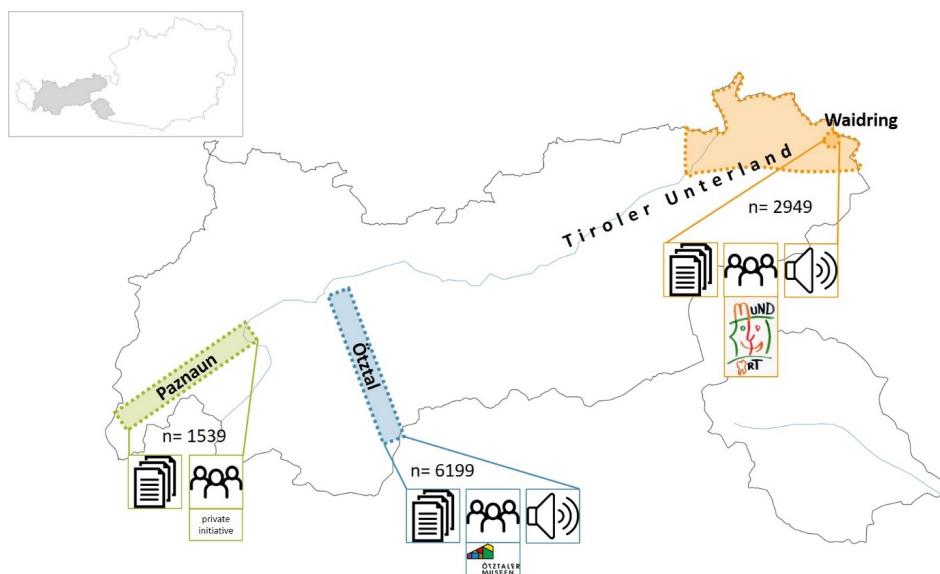


Figure 4: Three collections from three different regions in Tyrol were chosen. (Underlying map created by Y. Kathrein with SprachGIS – [regionalssprache.de](http://regionalssprache.de))

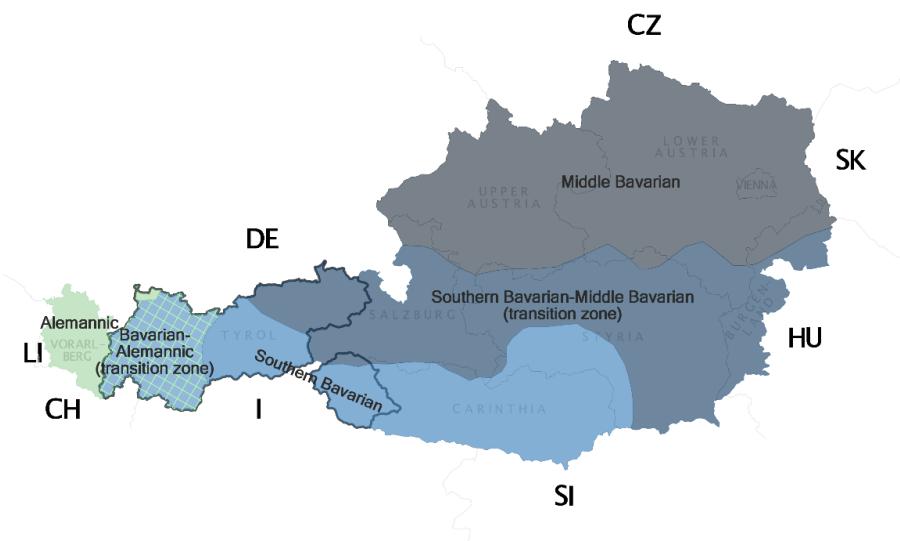


Figure 5: Dialect regions in Austria according to Wiesinger 1983. The borders of Tyrol are highlighted. (Map created by Y. Kathrein with SprachGIS – [regionalssprache.de](http://regionalssprache.de))

### 3.1.1 The Paznaun collection

On the western border of the province of Tyrol lies the Paznaun valley. The variety spoken there is assigned to Southern Bavarian, but is also quite strongly interspersed with Alemannic, since the Paznaun valley borders on the western province of Vorarlberg, whose dialects are Alemannic (see Figure 5) (cf. Wiesinger 1983: 836–842).

As shown in Table 1, the dialect differs from Alemannic mainly in that mhg. *a* is raised to /ɔ/ (mhg. *gäbele* > dial. /gɔble/ ‘fork’) in all positions, the umlauts are delabialized (mhg. *hächel* > dial. /haxlə/ ‘picking tool for blueberries’) and the mhg. diphthongs *î*, *û* and *iu* are diphthongized (New High German diphthongization) (cf. Wiesinger 1983: 830–831). Alemannic features are nevertheless present; for the phonetic domain, this applies, for example, to mhg. *ou*, which is preserved as /o:/ or /ou/ (cf. Wiesinger 1983: 831, 837). Thus, mhg. *boum* is realized as /po:m/. A peculiarity of the dialect in Paznaun (and in the neighboring Stanzertal to the north) is the treatment of mhg. *ei* as /a:/ (cf. Wiesinger 1983: 837) as in dial. /pa:/ for mhg. *bein* ‘bone’, which Kranzmayer regards as a “bavarianization of that older ā, as spoken for mhg. *ei* beyond the Arlberg pass in Montafon and in the Klosterthal [...].” (Kranzmayer 1956: 60, §20g4) [author’s translation]. A common feature with Alemannic is also found in the treatment of the final *-en* in words like *gelogen* (‘lied’ (past participle of *lie*)). It appears as /ə/ (/glo:ge/) (cf. Schatz 1903: 22, 92, Kranzmayer 1956: 115(§46h1))

Table 1: Dialect features of the Paznaun valley in contrast to the Alemannic dialect in the adjacent Montafon region

mhg.	mhg. example	Paznaun	example from collection	Alemannic (Montafon)
<i>a</i>	<i>gäbele</i>	/ɔ/	/gɔble/ ‘fork’	/a/
ä	<i>hächel</i>	/a/	/haxlə/ ‘picking tool for blueberries’	/ɛ/
î	<i>kîf</i>	/ai/	/kxaif/ ‘tight’	/i:/
<i>iu</i>	<i>schiuch</i>	/ui/	/ʃuix/ ‘shy’	/y:/
û	<i>strûche</i>	/au/	/ʃtrauxə/ ‘chill’	/u:/
<i>ou</i>	<i>boum</i>	/o:/, /ou/	/po:m/ ‘tree’	/o/
<i>ei</i>	<i>bein</i>	/a:/	/pa:/ ‘leg, bone’	/e:/
<i>-en</i>	<i>gelogen</i>	/ə/	/glo:ge/ ‘lied’	/ə/

Since 2011 there is an online collection for the Paznaun valley (<http://paznaunerisch.at>). With 1539 entries (as of August 2, 2021) it is the smallest collection to be analyzed here. It is based on several preliminary analogue works, mainly carried out by three persons: a retired elementary school principal, a retired elementary school teacher and the former valley doctor. This analog material was made available online to the general public and expanded by three private persons from this valley. People who speak the Paznaun dialect can contribute to the collection themselves by sending their words to those responsible. The contributions will then be checked and integrated into the collection.

The entries are presented on a single Wiki page (see Figure 6). Links for alphabetical search and two columns with the dialect word on the left and the translation on the right side make it quite a simple collection. In some cases, it is indicated in which part of the valley the word is used. In addition, there is sometimes information on whether the word is already disappearing (e.g. *äperlig*, *äperli* ‘poor (rather forgotten in the upper valley, still used in the lower valley)’ [author’s translation]), in which contexts it is used (e.g. *ausrucka* ‘to move out (concerning the music band, the rifle company)’ [author’s translation], or how it is used stylistically (e.g. *an Schråga* ‘a wooden frame; derogatory for an ugly person’) [author’s translation]. As I myself am a native speaker of the dialect spoken in the Paznaun valley, I was able to transcribe the items phonetically although the site does not provide audio material.

The Wiki is intended to a) serve as a reference work for the Paznaun dialect and its peculiarities<sup>14</sup> and b) document words that are about to disappear.<sup>15</sup> Whether this also means that the collectors want to preserve the dialect and thus carry it into the future, so that it strengthens again, is not entirely clear. In any case, however, this circumstance resonates when the three editors explain that the preservation of the dialect is close to their hearts.<sup>16</sup>

---

<sup>14</sup>“Diese Wiki soll als kleines Nachschlagswerk für diesen Dialekt und seine Eigenheiten dienen.” [‘This wiki should serve as a small reference work for this dialect and its peculiarities.’; author’s translation]

<sup>15</sup>“Viele Wörter [sic!] die mit der Zeit verloren/ vergessen wurden, können so einfach nachgeschlagen werden.” [‘Many words that have been lost/forgotten over time can be easily looked up this way.’; author’s translation]

<sup>16</sup>“Uns allen lag und liegt die Erhaltung des Dialekts am Herzen. Da sich die Welt im Tal aber gründlich geändert hat, ist heute schon vieles nicht mehr im Gebrauch. Somit ist einiges aus der alten Zeit hier zumindest konserviert.” [‘The preservation of the dialect was and is close to all our hearts. However, since the world in the valley has changed thoroughly, many things are no longer in use today. Thus, some of the old time is at least preserved here.’; author’s translation]

## Wörterbuch

Hier befindet sich das eigentliche Wörterbuch, um das sich dieses Wiki dreht!

### Inhaltsverzeichnis

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

#### A [Bearbeiten]

♦		♦
dr Å / di Å	=	der/die Eine (aber auch der/die Andere)
an' Å, Är	=	ein Ei (Galtür), Eier
ås	=	eins
åm	=	einem, zweien
Är, Äni	=	Einer, Eine
di Åna	=	die Anderen
dr / di Ándr, di Ándara	=	der / die Andere, die Anderen
an'etligi	=	einige
alå, är alå	=	allein; einer allein
atål	=	ein Teil (einer Gruppe), teilweise
åzacht	=	einzeln
dr/di Änzig	=	der/die Einzige
af Innschbrugg odr Wia	=	nach Innsbruck oder Wien
ågatlig, ågatli	=	eigentlich (die Endung -ig ist typisch fürs Obertal, -i fürs Untertal)
ågas	=	eigens
(da Rå) åb	=	(die stelle Wiese, den Rain) hinunter/hinab

Figure 6: The collection from the Paznaun valley is kept very simple: It only consists of links for alphabetical search and two columns.

### 3.1.2 The Ötztal collection

The second collection concerns the Ötztal that lies further east (see Figure 4). It is also located in the Southern Bavarian-Alemannic transition area (see Figure 5). Unlike in Paznaun, however, the final *-en* is realized as /n/ or /ŋ/ rather than as a vowel, so that it becomes /moxŋ/ ('to make') and /gə'lø:gŋ/ ('lied') (cf. Schatz 1903: 22, 92). Another phonetic characteristic concerns the palatal pronunciation of mhg. *o* and *uo*, for instance in *Loch* (dial. /løç/) ('hole') or *gut* (dial. /gøət/) ('good'), especially in the middle communities of the valley (cf. Klein et al. 1965: K. 30, 51).

At the morphological level, the following characteristics should be mentioned: the preservation of final short vowels, such as dial. /betə/ vs. std. /bɛt/ ('bed'), /siesə/ vs. std. /sy:s/ ('sweet') or /friesə/ ('feet'), whose *e* has been apocopied in other dialects (cf. Schatz 1903: 49, 93), and the preservation of the schwa sound in the prefix *ge-* before *h*, *w*, liquids, nasals, and plosives (/gəhɛvərn/ 'to belong',

/gəve:sŋ/ ‘have been’, /gə'le:gŋ/ ‘lied’, /gəmoxət/ ‘have made’, /gepləext/ ‘cried’) (cf. Kranzmayer 1956: 85, §29e).

Table 2: Some of the characteristics of the Ötztal dialect and its reflections in the collection

mhg.	mhg. example	Ötztal	example from collection
-en	<i>machen</i>	/n/	/moxn/ ‘to make’
	<i>gelogen</i>	/ŋ/	/gə'le:gŋ/ ‘lied’
<i>o</i>	<i>loch</i>	/ø/	/løç/ ‘hole’
<i>uo</i>	<i>guot</i>	/uə/	/gʊət/ ‘good’
-e	<i>bette</i>	/e/	/bete/ ‘bed’
	<i>süeze</i>		/siesə/ ‘sweet’
	<i>vüeze</i>		/fiesə/ ‘feet’
ge-	<i>gehören</i>	/gə/	/gəhəern/ ‘to belong’
	<i>gewest</i>		/gəvest/ ‘knew’
	<i>gelogen</i>		/gə'le:gŋ/ ‘lied’
	<i>gemeint</i>		/gəmɔət/ ‘meant’
	<i>geblèrt</i>		/gepləext/ ‘cried’

The collection, consisting of 6199 entries on August 2, 2021, is the most extensive of those to be analyzed. Again, there is a lot of preliminary analog material, which then resulted in the “Ötztal Dialect Dictionary” available online (<https://oetztalermuseen.at/dialektwoerterbuch>). It is run by the institution “Ötztaler Museen”, a cultural institution that unites three different museums of the valley and that is managed by a scientist. Ötztalers can also very easily provide word contributions themselves by inserting them into a prefabricated mask. After an internal check, they appear on the website.

In contrast to the collection from the Paznaun valley, there is also audio material included, which is gradually recorded by selected people from the valley and added to each contribution. It is arranged in blocks indicating the dialect word, the translation, the place to which the word applies, partially its syntactic embedding, an audio button to hear the pronunciation of the word and the person who contributed it (see Figure 7).

The website states that the aim of the joint collecting project is to raise awareness of language in general and of the peculiarities of the Ötztal dialect and to

**NEUESTE | A B C D E F G H I J K L M N O P Q R S T U V W X Z**

Suche nach... **SUCHE**

**tasigk**  
Schriftsprache: apathisch; lethargisch  
Erhoben in: Sölden

▶

Quelle: Markus Wilhelm

**geäärn, wen geäärchte?**  
Schriftsprache: welcher Familienabstammung  
gehörst du an?  
Erhoben in: Längenfeld

▶

Quelle: Bernhard Stecher

**dechtn**  
Schriftsprache: doch nicht  
Erhoben in: Sölden  
*Hall wearte dechtn scho geahn?*

▶

Quelle: Markus Wilhelm

**geän, güet geän**  
Schriftsprache: mit Schwierigkeiten bedient sein  
Erhoben in: Längenfeld  
*Bsp.: Do gangsche güet! = Diese Situation wäre nicht gerade optimal!*

▶

Quelle: Anna Praxmarer

Figure 7: The Ötztal collection is arranged in blocks and offers additional audio material as well as the source, sometimes also the word in its syntactic embedding.

appreciate both one's own identity and that of other regions. In this respect, the purpose of the collection can be seen as raising (regional) linguistic awareness and thereby achieving a comprehensive appreciation of all identities. The intention of the original collectors is not known to us. Indirectly, however, the purpose of documentation is also addressed when “personalities involved in the documentation of Ötztal *peculiarities* [emphasis by author]”<sup>17</sup> are mentioned on the website, whose collections have been integrated into the overall collection.

### 3.1.3 The Waidring collection

The dialects of the so called “Tiroler Unterland” (see Figure 4), i.e. especially those east of the area around Schwaz, are assigned to the Middle Bavarian transition area (see Figure 5). Among the characteristics are the vocalization of postvocalic *l* (dial. /hoits/ vs. Southern Bavarian /holts/ ‘wood’, dial. /vɔyt/ vs. Southern

<sup>17</sup>Original: “[...] auch zahlreiche weitere um die Dokumentation der Ötztaler Besonderheiten bemühte Persönlichkeiten [...]” (<http://oetztalermuseen.at/forschung/oetztaler-dialekt-woerterbuch/>) (Accessed December 31, 2022)

Bavarian /*velt*/ ‘world’, dial. /*muix*/ vs. Southern Bavarian /*milk*/ ‘milk’), the lenition of intervocalic *t* (dial. /*ve:də*/ vs. Southern Bavarian /*vēte*/ ‘weather’) and the realization of mhg. *ē* as a monophthong (dial. /*ſne:*/ vs. Southern Bavarian /*ſnev*/) ‘snow’). Another characteristic, but not limited to Middle Bavarian, is the fricative realization of *r* before dentals as in *Bart* (dial. /*bɔ:ft*/ ‘beard’) or *Herz* (dial. /*hēfts*/ ‘heart’) (cf. Kranzmayer 1956: 125, §50e3).

Table 3: Some of the characteristics of the Middle Bavarian dialect of Waidring as opposed to Southern Bavarian dialects

mhg.	mhg. example	Waidring	example from collection	Southern Bavarian
V- <i>l</i>	<i>holz</i>	/oi/	/hoits/ ‘wood’	/ol/
	<i>wēlt</i>	/ɔy/	/vɔyt/ ‘world’	/el/
	<i>milch</i>	/ui/	/muix/ ‘milk’	/il/
V- <i>t-V</i>	<i>wēter</i>	/d/	/ve:də/ ‘weather’	/t/
ē	<i>snē</i>	/e:/	/ſne:/ ‘snow’	/ɛə/
- <i>rt-</i>	<i>bart</i>	/ſt/	/bɔ:ft/ ‘beard’	/rt/
- <i>rz-</i>	<i>hērze</i>	/ſts/	/hēfts/ ‘heart’	/rts/

For the eastern part of the country, a collection supported by a dialect association called “Insa Tirola Mundart” was selected. Here, too, already existing collections were combined into an online collection (<https://www.tiroler mundart.at/woos-moast>) representing the northeasternmost part of the so-called Tiroler Unterland (see Figure 4). However, since the collections are based on dialects that differ more from each other in some respects than in the other two regions, a partial collection was selected from them that concerns the community of Waidring. It was created by a member of the association for his home village and contained 2949 entries at the time of the material extraction on April 12, 2021. It is arranged in columns and provides additional audio material that embeds the entry in a syntactic context. The audio file for the entry *Grieß* e.g. informs about the following: [ums goed is videramoe a vy:ds kri:s] ['Once again, there is a rather heated dispute over the money.]; author's translation]. Thus, the translation of *Grieß* as “starke Nachfrage” (‘strong demand’) is supplemented by the additional syntactic embedding in the audio file. Here, too, the search works both via an alphabetical bar and via text input, searching the columns “Mundart”

as well as “Hochdeutsch” (see Figure 8). Sometimes there are pictures linked to it that visually support the entry.

The screenshot shows a digital collection of dialect words. The interface includes a search bar with a magnifying glass icon and a 'x' for clearing the search. Below the search bar is a navigation bar with letters A through Z. The main table has columns for ID, Dialekt, Mundart, Hochdeutsch, Quelle, Audio, and Info. The 'Info' column contains small circular icons, some with arrows pointing to the right, indicating multimedia content like audio recordings or images. The 'Quelle' column shows 'Huawa' for most entries. The 'Audio' column has a speaker icon. The 'Info' column has a circular icon with a question mark. The 'Hochdeutsch' column contains definitions like 'geweint', 'brutal, grob', and 'zum Kranzgewobend Haarzopf'. The 'Mundart' column contains words like 'greascht, pläscht', 'grebbisch', and 'Gredlfrisur, Gredlfrisur'.

Dialekt		Mundart	Hochdeutsch	Quelle	Audio	Info
ID	Dialekt					
2185	Unterland	greascht, pläscht	geweint	✓ Huawa		
2186	Unterland	grebbisch	brutal, grob	✓ Huawa		
2187	Unterland	Gredlfrisur, Gredlfrisur	zum Kranzgewobend Haarzopf	✓ Huawa		
4244	Unterland	Gredlfrisur	die Haar werden am Kopf gezopft und zu einem Kreis fixiert	✓ Huawa		
4245	Unterland	greggat	klein verkümmt	✓ Huawa		
2189	Unterland	Grelei	Perlen am Rosenkranz, Kücklchn	✓ Huawa		
2190	Unterland	Griesch	Herz	✓ Huawa		
4246	Unterland	greisei	bisschen	✓ Huawa		
4247	Unterland	Grenggn	knorriges Stück Holz (auch für alten unguten Mann)	✓ Huawa		
2194	Unterland	griaß di	Gruß Gott (so wie's bei uns Tirolern der Brauch ist)	✓ Huawa		
2195	Unterland	griaß enk	grüß euch	✓ Huawa		
2196	Unterland	Griaßbeidl	Schmeichler	✓ Huawa		
2191	Unterland	griascht	gerürt	✓ Huawa		
2192	Unterland	Griaskoch	Griesbrei	✓ Huawa		
2193	Unterland	Griasia	Schmeichler	✓ Huawa		
4248	Unterland	Griedlstiera	Liebhaber	✓ Huawa		
4249	Unterland	Grieß	starke Nachfrage	✓ Huawa		
2197	Unterland	Griffischochtl	Federbinal	✓ Huawa		
4250	Unterland	grintig	grausig	✓ Huawa		
2198	Unterland	grippig	krank	✓ Huawa		
2199	Unterland	Grischbel	dünner, kleiner Mensch	✓ Huawa		
4251	Unterland	Grischpei	zarte Person	✓ Huawa		

Figure 8: The collection from Waidring was created by a collector who is called “Huawa”; it is integrated into a larger collection.

The purpose is similar to the one from the Paznaun valley: the provision of a reference work<sup>18</sup> and the preservation of local dialects.<sup>19</sup> Unlike the Paznaun collection, it is clear here that the association also wants to contribute to the continuity of the dialects.<sup>20</sup>

What unites the three collections, then, is their similar purpose. This puts them in line with very many other collections, as Baur explains:

<sup>18</sup>“Die Intention ist, möglichst viele Mundartliebhaber zu animieren, Mundartbegriffe aber auch fertige Wörterbücher zu integrieren, um eine umfassende Datenbank anbieten zu können.” [‘The intention is to encourage as many dialect lovers as possible to integrate dialect terms but also dictionaries in order to offer a comprehensive database.’; author’s translation] (<https://www.tiroler-mundart.at/woos-moast/>)

<sup>19</sup>“Wir sehen es als unsere Aufgabe, die Mundart durch Digitalisierung einerseits einfacher zu verstehen (sehen, hören statt lesen) und andererseits in die Zukunft mitzunehmen.” [‘We see it as our task, on the one hand, to make the dialect more understandable through digitalization (see, hear instead of read) and, on the other hand, to lead it into the future.’; author’s translation] (<https://www.tiroler-mundart.at/>)

<sup>20</sup>“Wir [...] wollen den Erhalt und die Pflege der Mundart des Tiroler Unterlands fördern.” [‘We [...] want to promote the preservation and cultivation of the dialect of the Tyrolean lowlands.’; author’s translation] (<https://www.tiroler-mundart.at/>) „Du wüst ins hoifen, das ma insa Mundart pflegen und fie die Zukunft dahoit’n?” [‘You want to help us maintain our dialect and preserve it for the future?’; author’s translation] (<https://www.tiroler-mundart.at/mitglied-insa-tiroler-mundart/>)

Die in ihrem Bestand als gefährdet angesehene Mundart soll für spätere Generationen [...] dokumentiert und vor dem Vergessen bewahrt werden. Darüber hinaus wollen nicht wenige durch ihre Sammelerarbeit die Mundart pflegen, sie konservieren und ihr neue Freunde, Sprecher und Leser, zuführen. Immer öfter wird in Vorworten der Dialekt als ein wichtiger Bestandteil örtlicher oder regionaler Kultur bezeichnet, den es zu erhalten gilt. Mundartwörterbücher sind also als ein Mittel, lokale Identität zu erhalten oder wiederzugewinnen.<sup>21</sup> (Baur 1987: 53)

### 3.2 Sampling

The samples for the Paznaun and Ötztal collection were taken on August 2, 2021, and those for Waidring on April 12, 2021. Due to the size of the collections, 70 randomly selected entries were taken for each collection, resulting in a total of 210 lemmas/syntagmas to be analyzed. However, for two collections, parts of it were already excluded in advance:

Due to a cooperation project with the Tyrolean Dialect Archive of the University of Innsbruck, the dialect documents of the scientist Eugen Gabriel had been integrated into the Ötztal collection. Of course, these are not amateur documents, which is why they were excluded from the sampling.

As already mentioned, only the partial collection of the collector “Huawa” from Waidring was selected as a sample for the collection of the “Tiroler Unterland”. I therefore call it the Waidring collection.

### 3.3 Documentation of deviations

All 210 entries were then cross-referenced with the Austrian dictionary (cf. *Österreichisches Wörterbuch: Vollständige Ausgabe mit dem amtlichen Regelwerk* 2018), the German variant dictionary (cf. Ammon et al. 2016), the DWDS (Digitales Wörterbuch der deutschen Sprache) and *Duden online* (Dudenredaktion n.d.). This means that, on the one hand, lexical and semantic deviations from both the Federal German (viz. German standard language associated with Germany) and

---

<sup>21</sup>‘The dialect, which is considered endangered in its existence, should be documented for later generations [...] and saved from oblivion. In addition, quite a few want to maintain the dialect through their collecting work, to preserve it and to bring it new friends, speakers and readers. More and more often the dialect is described in prefaces as an important component of local or regional culture, which must be preserved. Dialect dictionaries are therefore a means of preserving or regaining local identity.’ [author’s translation]

the Austrian standard language were documented. On the other hand, phonological and morphological deviations were documented and subsequently subjected to a phonological distance measurement according to Levenshtein (see Section 5.2). In addition, register, style, and age information, as well as the approximate diatopic distribution of each item, were recorded if the dictionaries provided information about that.

## **4 Questions and hypotheses**

As stated in Section 1, I consider lay dialect collections as a source for perceptual dialectology, since their contents indirectly shed light on what lay people consider their dialect. Presumably, the items prototypically represent those words or phrases, that are salient to them, that best characterize their dialect, that make it up.

This study therefore focuses on the following two questions:

- What do laypersons' dialect collections contain?
- How "dialectal" are the collections, i.e. where on the dialect standard axis are the items located?

"Dialekte sind die standardfernsten, lokal oder kleinregional verbreiteten Vollvarietäten." (Schmidt & Herrgen 2011: 59) Bearing in mind this definition of Schmidt & Herrgen (2011) according to which dialects are – among other things – furthest from the standard, I assume that

1. dialect collections created by laypersons contain mostly items that do not exist in the standard language.

Apart from the collectors' intention to record words that are in the process of being forgotten, this assumption is supported by Baur's observation that it is often "conspicuous, inexplicable or particularly old words" (Baur 1987: 66; author's translation) that are considered worth recording. Such words are hardly words of the standard language, but dialect words.

The peculiarities of a dialect which are also to be transported in such collections can find their expression not only on the lexematic, but also on the phonological and/or morphological level. Therefore, I further assume that

2. if the items do exist in the standard language, they
  - a) have a considerable phonological and/or morphological distance and/ or
  - b) are semantically distinguished from its equivalent in the standard language.

## 5 Results

### 5.1 Deviations on the linguistic levels

In keeping with the purpose of the collections described in Sections 3.1.1–3.1.3, I expect to find a relatively large number of words and/or phrases that demonstrate the distinctive features of the dialect in question; that is, I expect to find words that have no equivalent in the standard language either because of differences at the lexematic or semantic level.

It must be added that some words, of course, vary not only on one linguistic level, but on several. For example, the dialectal word *Gschlatter* is contrasted with its equivalent *Gschlader* [kʃla:də], which is marked as colloquial but can be found in the standard dictionary. The dialectal word refers to a semi-liquid mass (dough, mortar) that is (deliberately) sloppily mixed together<sup>22</sup> whereas the “standard” word refers to an inferior drink, a thin coffee<sup>23</sup>. Therefore, on the one hand, they differ in their semantics. In addition, the dialectal word is realized with a fortis plosive ([kʃlatr]), while the “standard” word is realized with a lenis plosive [kʃla:de]. Thus, they also differ at the phonological level. This leads to the classification as “phon-sem” (another example of a phonological, morphological and semantic deviation is *Plot* in Table 6).

#### 5.1.1 Lexical and semantic deviations

Table 4 shows the percentages of deviant elements on the different linguistic levels for all three samples. As can be seen, words that do not exist in the standard language or that exist but differ on a semantic level make up about half of the sample (53 % Paznaun, 57 % Ötztal, 50 % Waidring).

Which words make up this half? As far as the lexical level is concerned, there are words like *ånlag* (Paznaun) ‘comfortable, even (to walk)’, *vrkåltn* (Ötztal) ‘to put away’, *Schüehenooch* (Ötztal) ‘kick’ or *Kartoffelwiala* (Waidring) ‘dish from

<sup>22</sup>‘(oft absichtlich) schlampig zusammengerührte halbflüssige Masse (Teig, Mörtel)’

<sup>23</sup>‘minderwertiges Getränk, dünner Kaffee’

Table 4: Linguistic-levels-profile of the three samples. In the Paznaun sample, 53 percent have no lexical equivalent or different semantics, in the Ötztal sample it is 57 percent, in the Waidring sample 50 percent (all in bold letters).

	<i>Paznaun</i>	<i>Ötztal</i>	<i>Waidring</i>
<b>lex</b>	<b>46</b>	<b>49</b>	<b>37</b>
<b>only sem</b>	-	4	6
<b>phon-sem</b>	4	4	6
<b>phon-morph-sem</b>	-	-	1
<b>morph-sem</b>	3	-	-
<b><i>SUM</i></b>	<b>53</b>	<b>57</b>	<b>50</b>
only phon	33	23	29
phon-morph	10	14	11
phon-syn	2	2	-
only morph	1	1	3
no difference	1	3	7
<b><i>SUM</i></b>	<b>47</b>	<b>43</b>	<b>50</b>

steamed potatoes'. In this case, either the object/concept does exist in both systems but is realized in completely different ways (*ånlag*, *vrkåltn*, *Schüehenooch*) or the word refers to an object that is only locally common and has therefore no equivalent in Standard German (*Kartoffelwiala*; see Table 5).

Semantic differences, on the other hand, can be seen, for example, in the word *Muli*: There does exist the word *Muli* in Standard German, but it neither means 'donkey' nor 'drunkenness' as the dialect word suggests. It just means 'mule' in Standard German, which is a crossing between donkey and horse. Of course, other linguistic levels may be involved here, as outlined above; nevertheless, the semantic difference remains the most important for classification. This applies, for example, to the word *Plot*, which is pronounced [plot] in the local dialect<sup>24</sup>, while in Standard Austrian it is [plate]. (see Table 6).

This half (words non-existent in the standard with different semantics) forms, so to speak, what could be understood as a dialect collection in the narrowest sense: It consists of words which are completely excluded from standard dictionaries or which differ semantically from entries in standard dictionaries.

<sup>24</sup>Weak feminina are subject to the e-syncope in Middle Bavarian (Klein et al. 1965: K. 57)

Table 5: Items that differ on a lexical level

	standard equivalent	entry in dialects dictionaries
<i>ähag</i> (Paznaun) 'bequem, eben (zu gehen)' ('comfortable, flat (to walk)')	-	yes, but narrower range of meanings: 'vom Gelände, leicht ansteigend, wenig geneigt' ('slightly ascending terrain, little inclined'); Schatz 1993: 19)
<i>vrkältn</i> (Ötztal) 'verräumen' ('to put away')	-	no, merely <i>aukqltn</i> 'aufbewahren' ('to store sth.); Schatz 1993: 276) and <i>g(e)halte</i> " 'etw. in ein Behältniss legen, an seinem gehörigen Ort, Versteck aufheben, im Stand erhalten, eig. und bildl.' ('to put something in a container, to keep in its proper place/hiding place, to keep in state, proper and figurative'; Schweizerisches Idiotikon 1881ff: 1235)
<i>Schüehenooch</i> (Ötztal) 'Fußtritt' ('kick', literally 'foot kick')	-	no, only noun <i>fuessinärsch</i> (literally 'foot in the ass'; Schatz 1993: 193) and verb <i>fuessärsche</i> 'giving kicks in the bottom with your ass' (Schweizerisches Idiotikon 1881ff: 467)
<i>Kartoffelwiala</i> (Waidring) 'geriebene Erdäpfel gediinstet' (dish from steamed potatoes')	-	yes (WBÖ-Database <sup>a</sup> entry <i>Wirler</i> , nr. 1C.2a29)

<sup>a</sup><https://hie.diee.at/db>, accessed October 10, 2023.

Table 6: Items that differ on a semantic level

	standard equivalent	entry in dialects dictionaries
<i>Muli</i> (Waidring) 'Esel, Rausch' ('donkey', 'drunkenness')	<i>Muli</i> 'Maultier' ('mule')	yes, but only in parts of South Tyrol (Schatz 1993: 437)
<i>Plot</i> (f.) [plot] (Waidring) <sup>a</sup> 'Schulter' ('shoulder')	<i>Platte</i> [plate] among others 'verschiedenen Zwecken dienendes flaches Stück aus einem bestimmten Material, dessen Dicke im Verhältnis zu den anderen Abmessungen gering ist' ('flat piece of a certain material serving various purposes, the thickness of which is small in relation to the other dimensions')	no

<sup>a</sup>*Plot* additionally differs on phonological and morphological level.

A look at the dialect dictionaries compiled by scholars, on the other hand, reveals interesting aspects (see Table 5). First, some of the words collected by laymen (*vrkältn*, *Schüehenooch*, *Plot*) are not listed in them at all, which is probably because the question books on which the dictionaries are (coincidentally) based did not aim to do so (or else because the “question book makers” simply did not think to query them). Second, some words are listed, but not semantically differentiated in such a way that the local meanings emerging from the lay collections could have been taken into account (*ånlag*, *Muli*).

This small finding already shows how fruitful the inclusion of such lay collections would be in the compilation of dialect dictionaries, not only to have questionnaires created by scientists answered by lay people, but also to enrich the dictionary with introspectively generated lay data and thus come closer to the language reality.<sup>25</sup> Since these lay collections were often created by several authors in joint consultation, the material can certainly be considered reliable. Combining this with an online survey to verify entries would provide additional reliability (cf. Hofer & Meier 2015, Retti 1999). And it is a starting point to re-examine the question of subjective and objective data that has preoccupied perceptual dialectology since its inception: To what extent can subjective data on the dialect perception of laypersons (perception data) be reconciled with production data determined objectively by scientists? When is reconciliation being canceled? And why? (cf. Preston 1999) Applied to our collections, this would mean: What words have been introspectively included in a lay collection that do not appear in a dialect dictionary produced by scientists? And why? Is it because of the recording method? Or are individual entries too widespread and/or too close to the standard to be included in a dialect dictionary? Are the lay collectors not even aware of this and do such words belong to their understanding of dialect, which would then be very broad? Or are they aware of this and include such words in their collections anyway because they are used in everyday speech, no matter how base dialectal they are? To explore this is beyond the scope of this article. Qualitative interviews with lay collectors would be additionally necessary to shed light on these questions.<sup>26</sup>

<sup>25</sup>In a sense, the work of scientists is also introspection, since they, too, map only a part of reality through the creation of their questionnaires, albeit a very well-founded one. Hofer & Meier (2015) speak here of the fact that lexicographical competence and, in particular, the introspection of dictionary makers alone are actually not sufficient for scientific purposes to represent language in its natural environment: “Die Lexikographenkompetenz allein ist für einen Intersubjektivität anstrebenden wissenschaftlichen Standard jedoch ungenügend, denn durch allzu starke introspektive Ausrichtung [der WörterbuchmacherInnen, YK] wird der eigentliche Untersuchungsgegenstand – die Sprache in ihrem natürlichen Umfeld – verfehlt.” (Hofer & Meier 2015: 120)

<sup>26</sup>A first pilot interview in this context has already shown, for example, that laypersons are quite capable of observing language closely and that they are quite aware that there are gaps in the material. However, it also became clear that a lot of things happen unconsciously.

### 5.1.2 Phonological and morphological deviations from the standard language

As far as the other half is concerned, it consists of deviations that only occur on the phonological or morphological level or it is a combination of both (see Table 4). Of course they follow the sound laws of the respective dialects: It is words like *brat* [bra:t] vs. *breit* [brait] (= 'broad, wide') in the Paznaun sample, *Viich* [vi:ç] vs. *Vieh* [vi:] (= 'animal, livestock') in the Ötztal sample or *heazoagn* ['heetsøen] vs. *herzeigen* ['heøtsaign] (= 'to show') in the Waidring sample. Given its relative closeness to the standard variant, its occurrence in a dialect collection may be surprising. However, at least in the Paznaun sample, 21 out of 33 (= 63 percent) of this second half are marked either in terms of regional distribution (e.g. Austria and/or Southern Germany), style (colloquial, slang, elevated) or the age of the word (archaisms). Thus, words such as *flattiara* or *Dolba* make up about two thirds of those words only deviating on a phonological and/or morphological level (see Table 7).

For the Waidring and the Ötztal sample, this is slightly different: Stylistically marked words such as *potzn* 'to blot' in Waidring make up 49 percent (= 17 out of 35) whereas in Ötztal words such as *tochtlen* 'to slap someone (across the face)' only make up 19 percent (= 6 out of 31).

There are even a few items (7, 3 and 1 percent) that show no difference – at least no phonological difference – to the standard variant at all. However, such items are again mostly words that are marked either as colloquial or as slang or that only show a certain regional or functional distribution. The only visible difference is that they sometimes differ in terms of graphemes (see Table 8).

So, on the one hand, we can observe that about half of the items in our samples deviate lexically or semantically from standard language. Whether the first hypothesis ("dialect collections created by laypersons contain mostly items that do not exist in the standard language") as well as the second part of the second hypothesis ("if the items do exist in the standard language they are semantically distinguished from its equivalent in the standard language") can be confirmed by this might depend on what is meant by "mostly": If by this is meant "more than half", it can be said for at least two of the three samples that they consist mainly of words not found in the standard language. However, given the small number of items examined, this result can only be an approximation and we can state: It seems that words not found in the standard language are an important part of the dialect collections of laypeople.

The other half consists of words that differ on the phonological and/or morphological level, with a certain number of words marked (Paznaun about two thirds, Ötztal about half, Waidring one fifth). In this, the samples differ considerably.

Table 7: Examples of items that a) only differ on a phonological and/or morphological level and that b) are marked somehow.

dial.	std.	meaning	percentage of marked items
<b>Paznaun</b>			
<i>flattiera</i>	<i>flattieren</i> used in Switzerland, archaic in Austria and Germany	‘to flatter someone’	63
<i>Dolba</i>	<i>Dolm</i> Austrian, colloquial, offensive	‘stupid person’	
<b>Waidring</b>			
<i>potzn</i>	<i>patzen</i> Bavarian, Austrian	‘to blot’	49
<b>Ötztal</b>			
<i>tochtlen</i>	<i>dachteln</i> East Middle German, Southern German, Austrian, colloquial	‘to slap someone (across the face)’	19

Interestingly, there is a small percentage of words that do not differ from the standard pronunciation. But, again, almost all of these items are marked stylistically. To test the first part of the second hypothesis (“if the items exist in the standard language, they have a considerable phonological and/or morphological distance”), the deviations on these two levels have to be measured, which will be described in the next chapter.

## 5.2 Measuring phonological and morphological deviations

Here we now want to pursue the question of how much the items deviate phonologically and morphologically from Standard German. Thus, we are concerned

Table 8: All items of the samples that phonologically and/or morphologically do not differ from its standard equivalent.

dial.	std.	meaning	percentage of items not differing
<b>Waidring</b>			
<i>Zwiesl</i>	<i>Zwiesel</i> local	‘tree with forked trunk and two crowns’	7
<i>wist</i>	<i>wist</i>	‘Haw!’ [command to a horse to turn to the left]	
<i>Häfn, Hefn</i>	<i>Häfen</i> Austrian, colloquial	among other things: ‘prison’	
<i>sinnian</i>	<i>sinnieren</i> colloquial	‘to ponder’	
<i>Hanni</i>	<i>Hanni</i>	‘female surname’ (Johanna)	
<b>Ötztal</b>			
<i>Murgs</i>	<i>Murks</i> colloquial, pejorative, casually	‘sloppy, poor work’	3
<i>Scheniiror</i>	<i>Genierer</i> colloquial, esp. Austrian	‘shame’	
<b>Paznaun</b>			
<i>drinn</i>	<i>drin</i> colloquial	‘inside’	1

with the degree of dialectality of each item, i.e., the phonetic distance between an individual dialect word and its standard counterpart, which is indicated by the *d*-value (short for dialectality value).

For this purpose, each word was subjected to a normalized dialectality measurement according to Levenshtein (cf. Heeringa 2004).<sup>27</sup> That is, the individual items (in our case, the sounds) of one “chain” (in our case, the phonetically transcribed words) are compared with those of another chain. The goal is to calculate how many transformations must be made to the individual sounds in order to move from one chain to another. This is done by taking away, adding or substituting elements. However, each of these transformations “costs” something. The sum of the costs then gives the *d*-value. For example, it is calculated how much it costs to get from the dialect pronunciation [glandə] to the Standard German pronunciation [gelendə] (‘railing’).

Table 9: Measuring Levenshtein distance

1	2	3	4	5	6	7
g		l	a	n	d	ə
g	ɛ	l	ɛ	n	d	ə
					1	2

Each operation here costs<sup>28</sup>, which then gives a value of 2. To include the length of the word in the calculation, the value was normalized, which means that the result was divided by the number of alignment slots ( $2 \div 7 = 0.29$ ). Thus, the *d*-value of [glandə] compared to [gelendə] is 0.29. (cf. Beijering et al. 2008) This means that a little less than one third of the word shows deviations from the standard language.

For the calculations of the *d*-values, all words in question were broadly transcribed on the basis of the available audio files or on the basis of my own dialect

<sup>27</sup>These calculations were performed using Peter Kleiweg’s online tool: <https://urd2.let.rug.nl/~kleiweg/L04/webapp/bin/home> (accessed on December 26, 2022). The Levenshtein distance was preferred to the method developed by Herrgen/Schmidt (cf. Herrgen & Schmidt 1989, Lameli 2004), which is otherwise widely used in variational linguistics, for the simple reason that it provides an automation-based tool that allows for rapid measurements. Nevertheless, the Levenshtein distance measurement is also used in variational linguistics, especially in the Dutch area (cf. Gooskens & Heeringa 2004; a compilation of different methods can be found in Nerbonne & Heeringa 2009). Levenshtein also proves to be advantageous in evaluating other collections, since audio material is not always available, but one must infer pronunciation from the mere spelling of the items.

<sup>28</sup>A vowel-consonant substitution even costs 2, thus *sie* vs. *ʃø:n* (‘nice’) would be  $1 + 2 = 3$ .

competence (see Section 3.1.1). They have been transcribed broadly since I assume that allophones such as [x] ~ [ç], [r] ~ [r̥], [e] ~ [ə] are not conscious to laypersons or that these sounds are not the reason for inclusion in a collection. Standard Austrian pronunciation served as a basis for comparison, differing from Standard German pronunciation primarily in that /s/ is unvoiced in a vowel environment (e.g., Austrian /sa:gŋ/ vs. Federal German /za:gŋ/), short /i/ and /y/ are not centralized (e.g. Austrian /milç/, /glyk/ vs. Federal German /mɪlç/, /glyk/) and short final /e/ or those occurring in unstressed syllables are pronounced as /ɛ/ rather than /ə/ (cf. Wiesinger 2009). The words were generally transcribed in the allegro form (/maxn/ vs. /maxən/).

Figure 9 shows the *d*-values of phonologically and/or morphologically deviating items of the three samples. The medians (marked by a line) of the Paznaun and Ötztal samples are 0.5 (with average values of 0.45 and 0.46 resp., marked by an x), while the median of the Waidring sample is lower at 0.29 (average 0.34). This is consistent with the percentage of non-divergent elements (7 percent) and the lowest percentage of elements without a standard equivalent (50 percent, both see Table 4): The Waidring sample seems to be less dialectal than the other two collections.

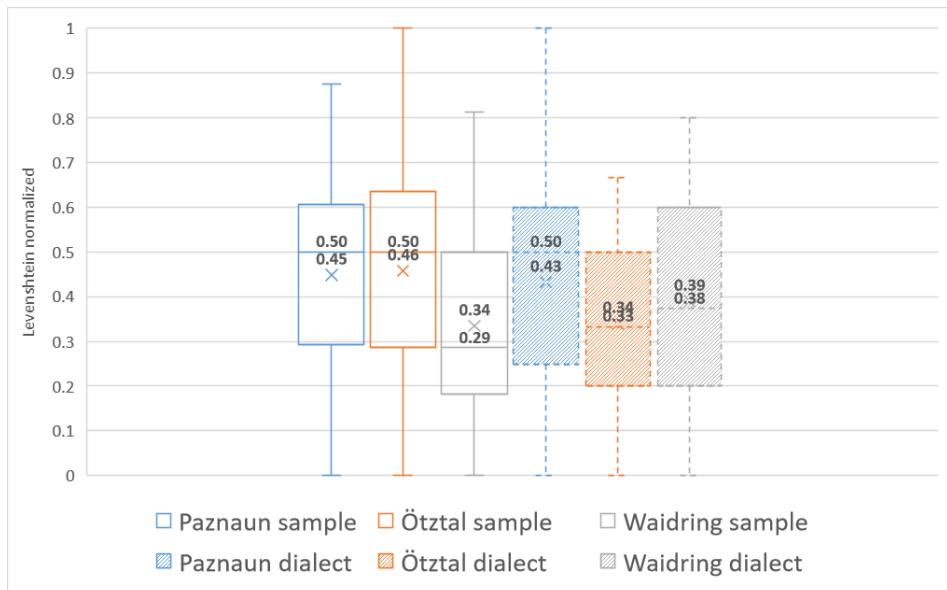


Figure 9: Normalized dialectality values of the samples (left half of the figure; Paznaun:  $n = 33$ , Ötztal:  $n = 30$ , Waidring:  $n = 35$ ) and of the underlying dialect (right half;  $n = 51$  each).

In order to rule out the possibility that it is not the dialect itself which has a lower degree of dialectality compared to the standard language, the *d*-value of the underlying dialect was also measured. For this purpose, the Wenker sheets available for the respective areas<sup>29</sup> were evaluated, since they allow a good comparison due to the uniform template. For the Paznaun valley, Ischgl was chosen as the reference corpus, for Ötztal it was Sölden. For Waidring, of course, it remained Waidring. The words used for comparison have not changed significantly since the 1930s (i.e. since the time of recording). If they changed or a different lexeme was used, an alternative word from the holdings of the Tyrolean Dialect Archive was used for comparison for all three places.

Since the vowel system of the underlying dialects deviates more from the norm than the consonant system, each Middle High German vowel was assigned a reference word from the Wenker sentences. Phonetic contexts with nasals and liquids were also considered (e.g. mhg. *-ân-* or mhg. *-ër-*). In the consonantal domain, Middle High German *t* (in intervocalic use) and *l* (in postvocalic use) were used. This is because South Central Bavarian (to which the dialect in Waidring belongs) is especially characterized by vocalization processes. (see Section 3.1.3)

This selection inevitably simplifies the linguistic reality. To mitigate this effect, the comparison was not made between individual vowels and consonants in isolation, but rather between entire words. These words were chosen specifically for their reference sounds in Middle High German. In addition, further phenomena (e.g. morphological) come into play (final *-en*, prefix *ge-* for participles, *n*-apocope, *e*-apocope for weak feminines, etc.). Examples of this comparison are shown in Table 10 (for the complete list see Section 6: Appendix).

The results of the calculations can be seen on the right part of Figure 9 (dashed bars). This shows that the *d*-value for the dialect of the Paznaun valley corresponds closely with that of the sample (blue boxes; 0.5 medians, 0.45 vs. 0.43 average). Thus, the collectors seem to be able to represent their dialect quite well in the collection. Words that differ phonetically/morphologically from the standard are not overrepresented in the sense that they would exceed the *d*-value of the Wenker material (which might be expected in a collection that aims to document idiosyncrasies that need not be in the lexical domain alone).

This is the case for the Ötztal sample (orange boxes): Here, a median *d*-value of 0.5 (or 0.46 on average) for the sample contrasts with a median *d*-value of 0.33 (or 0.34 on average) for the underlying dialect. In this collection, too, the focus is on the special features of the Ötztal dialect; however, in contrast to the

---

<sup>29</sup>The Wenker sheets are easily accessible via the REDE tool <https://apps.dsa.info/wenker/>. (Accessed August 08, 2024)

Table 10: Examples of words from the Wenker material whose local pronunciation was compared to the Standard Austrian pronunciation. The column “mhg.” indicates which sound, sound combination or morpheme was targeted in the respective word.

mhg.	word	dial.	std.	<i>d</i>
Ischgl		saf:fe		0.50
Sölden	ei/-e (fem.)	Seife	sæs fɛ	0.40
Waidring			fæs f	0.60
Ischgl		gloube		0.50
Sölden	ou/-en	glauben	gle:bm	glaubm
Waidring			gla:m	0.42
Ischgl		falt		0.25
Sölden	ël	Feld	falt	felt
Waidring			føit	0.75
Ischgl		to:		0.80
Sölden	ge-t/ân	getan	gɛtɔ:n	gɛta:n
Waidring			tū:	0.67
Ischgl		milx		0.00
Sölden	il	Milch	milx	milx
Waidring			muix	0.50
Ischgl		ſie		1.00
Sölden	-œn/-n <sup>a</sup>	schön	ſæn	ʃø:n
Waidring			ſe:	0.67
Ischgl		birſte		0.50
Sölden	-rst/-e (fem.)	Bürste	bi:xta	byrstɛ
Waidring			bi:ft	0.67

<sup>a</sup>after root vowel

Paznaun valley, the Ötztal idiosyncrasies on the phonological and morphological level are strongly condensed in the single collected items. Words such as dial. *Keer* [kxeer] versus std. *Gehör* [gehøe] ‘sense of hearing’ (*d*-value: 0.83) or dial. *Gewächsnr* [gevoksnr] versus std. *Erwachsener* [ɛvaksne] ‘adult’ (*d*-value: 0.75) will account for this.

The Waidring collection, on the other hand, seems to be less dialectal than it could have been in terms of phonological and morphological features. With a *d*-value of 0.29 (median; or 0.34 on average), the sample has the lowest degree of dialectality, as already noted. There, we find words such as *vazochn* ‘ill-bred’ (dial. [fɛtso:xn] vs. std. [fɛtso:gn] = *d*-value of 0.11) or *Ralweg* ‘bikeway’ (dial. [ra:lve:k] vs. std. [ra:dve:k] = *d*-value of 0.17). Relative to the *d*-value of the underlying dialect, which is 0.38 (median; or 0.39 on average), however, the result is somewhat relativized (gray boxes).

What does this mean with regard to the first part of the second hypothesis (“If the items do exist in the standard language, they have a considerable phonological and/or morphological distance.”)? We can state that the Ötztal sample has the highest degree of dialectality, compared to the degree of dialectality exhibited by the underlying dialect. The *d*-value of the sample is 17 (medians) or 12 (average) percentage points higher than the Wenker material, which is certainly not negligible. Possibly this has to do with the large number of collectors who were and are involved: The more people involved, the better they get to the heart of the matter (whereby “many cooks” could of course also “spoil the broth”). In addition, the institutional framework in which the collection is embedded could be effective here: The Ötztal museums do not only want to collect and convey knowledge, but also to conduct research.<sup>30</sup>

In this context, however, more detailed information would be needed about the possible intervention of the final editors, who are scientists. But perhaps this also has to do with a sharpened sense of one’s own variety; the linguistic awareness and self-image of the Ötztalers may have been spurred on by the inclusion of the Ötztal dialect as an intangible cultural heritage in 2010.<sup>31</sup> This may also be reflected in the fact that the phonologically and/or morphologically deviating words are the least frequently marked (19 percent; see Table 7). Thus, we can probably indeed speak of the degree of dialectality being considerable in the Ötztal sample, even if stylistic markings could be more frequent.

The other two samples do not have a higher *d*-value than the underlying dialect (we neglect the difference between 0.45 and 0.43 here), but they are still

<sup>30</sup><http://oetztalermuseen.at/uber-uns/> (Accessed December 29, 2022)

<sup>31</sup><https://www.unesco.at/kultur/immaterielles-kulturerbe/oesterreichisches-verzeichnis/detail/article/oetztaler-mundart> (Accessed December 29, 2022)

quite and relatively close to this value, respectively, which means that especially in the case of the Paznaun sample, with its high proportion of marked words (63 percent, see Table 7), one can also speak of a fairly substantial phonological/morphological difference to the standard language (always taking into account the *d*-value of the underlying dialect, of course).

The sample from Waidring, with its 49 percent of marked words (see Table 7) and its somewhat lower *d*-value than the underlying dialect indicates, can probably still be called dialectal, if not to a considerable, at least to a respectable extent.

Of course, one should not overinterpret these numbers in view of the small samples. But they can show a certain tendency, which is also confirmed with a similarly high number of lexical and semantic deviations, as we have already seen. And they show that lay dialect collections have sufficient linguistic evaluation potential, for example, with regard to the ideas lay people have about the vertical dimension of their dialect.

## **6 Conclusions and research perspectives**

The samples of lay dialect collections from Tyrol examined in the study, namely from Paznauntal, Ötztal, and Waidring, show a similar composition: About half of the respective material deviates from the standard language on a lexical and semantic level, i.e., this half forms the core of what can, among other things, be understood as dialect, namely the variety that has the greatest distance from the standard language. This half, therefore, will certainly do the most justice to those collectors who not only want to document the dialect, but also to save it from oblivion and perhaps even to make it fit for the future. The exact proportion of such individuals in the collections cannot be determined, as multiple people – and in the case of the Ötztal region, a particularly large number – were involved. But we know the intention of the editors of the collections, which in all three cases is to document the respective dialect with its peculiarities.

These peculiarities may be reflected not only at the lexical and semantic level, but also at the phonological and morphological level. This makes up the other half of the samples. In this, however, the three collections differ quite clearly: In the sample for Paznaun, almost two-thirds of phonologically/morphologically divergent words are marked in terms of distribution, style, or age. Moreover, these words have a mean *d*-value of 0.5, so the Paznaun sample is quite strongly dialectal after all. In Waidring, half of the phonologically/morphologically divergent words are marked. This is contrasted with the lowest *d*-value of 0.29 (with a *d*-value of 0.38 of the underlying dialect). Thus, this sample might be somewhat

more dialectal. In the Ötztal sample, only one fifth of the phonologically/morphologically deviant words are marked. However, the *d*-value must also be taken into account here, which is highest in this sample with 0.5 relative to the *d*-value of the underlying dialect (0.33). This might “allow” to include more unmarked words in a collection. Thus, the Ötztal sample is also very dialectal overall and can thus reflect the specifics of the dialect well.

The study of lay dialect collections naturally raises many questions. Most importantly, it stimulates further research, since such collections have not been perceived as a possible source for perceptual dialectology so far:

Apart from the fact that the samples would obviously have to be enlarged in order to provide representative results, it is obvious to conduct comparative studies, for instance with collections that have a different intention, e.g. those that have been created in a tourist context. There, clichéd, obscure, exotic, hard-to-pronounce, or otherwise strange words and phrases are to be expected with disproportionate frequency, just as a study of auto-stereotypes and meta-stereotypes (= supposed auto-stereotypes) would be worthwhile as a comparison, regardless of the intention with which a collection was created. Because what you believe others would not understand, find funny or know in a different context, you may be more likely to include in your collection than something that does not interest others because of its (supposed) information content. Presumably, attitudes towards one’s own or other people’s varieties also play a role in this context. In this respect, a comparison with collections from other (federal) states or dialect areas could also provide new insights into possible regional differences.

Theoretical penetration is one thing, practical processing is another. Ultimately, the linguistic considerations of “why?” are just as much a guess, for example: Why are words included whose dialectal value is very low to non-existent?

This leads to the question of whether measuring *d*-value is an appropriate way of addressing these collections at all, and to answer the question “why?”. After all, a word like [bra:t] ‘broad’ (compared to the standard word [brait]) may indeed have a normalized *d*-value of 0.3 according to Levenshtein. However, this says nothing about how “dialectal” this difference appears to laypeople, especially considering how this word is realized in surrounding dialects. It also says nothing about what laypeople are even aware of.

Certainly, saliency and prototypicality studies should therefore complement such analyses in which the subjects are asked, among other things, how they characterize their dialect or what is special/typical about it compared to other dialects in their region/dialects in other regions, and also what connects it to others. This is also conceivable in the opposite direction, if informants were asked

which given words best represent the subjects' dialect. This also suggests experimental arrangements in which the participants are asked to write down dialect words from their own dialect that occur to them spontaneously and to comment on them additionally or to discuss with each other afterwards why they have included the word (cf. e.g. the Pilesort method in the DFG project "Der deutsche Sprachraum aus der Sicht linguistischer Laien" ("The German language area from the point of view of linguistic laypersons"), in which the subjects were also asked about the motives for their sorting. There, the question of differentiation from other dialectal units also arose, but not on a vertical but on a horizontal level) (cf. Schröder 2017: 52–54). Therefore, qualitative interviews with amateur collectors could be another fruitful source to learn more about their motives for collecting and their conscious or unconscious approach, and thus to revisit the question of what makes their collections different from those of scientists, or what does not, and what are possible reasons for this.

It would also be interesting to obtain data through participant observation, for example, during an "editorial meeting" that takes place as part of the creation of a lay collection. Such data could provide insights into the process of creation and the resulting considerations, questions of awareness, or generally the evolution that a collection may take during this process.

In short, a whole range of methods that perceptual dialectology has developed and/or refined on its still rather young path could be applied. Or vice versa: lay dialect collections, if available, could complement the already existing perceptual dialectological research.

On the technical side, new possibilities for evaluation could also open up: The diploma thesis of Stefanie Kapferer, which was dedicated to the topic "Automation-supported measurement of the degree of dialectality in lay collections", shows that a combination of algorithm development and manual refinement could speed up the evaluation (Kapferer 2020). However, since the evaluation mode has changed several times since the completion of the work, further adjustments would be necessary for this.

What must not be lost sight of, however, in spite of all the methodology, is that the different degrees of knowledge of lay people, their awareness of problems, but also their ability to deal with linguistic problems, will also be reflected in the composition of the corpus. Thus, a collection aimed at preserving the language may nevertheless contain a relatively high proportion of words that differ from the standard language only on the phonological and/or morphological level. Whether the authors are always aware of this is questionable. And even if they are aware of it in individual cases, the stringency may suffer from the overall task to be accomplished.

The handling of linguistic issues, moreover, brings together laypeople and lexicographers of standard dictionaries. It is always a question of delimitation: How much and what of technical and group language(s), how much and what of foreign vocabulary, how much and what of regional forms, style levels and registers should be represented in the reference work (cf. Bergenholz 1989, Haß-Zumkehr 2001: 21, Brückner 2006, Lenz 2019)? If the vertical delimitation in a standard dictionary concerns the direction “down”, laymen must try to delimit in the direction “up”. Berthele summarizes this difficulty of conceptualizing varieties:

Sprachen und Varietäten existieren ja nicht in Sprachatlanten, Grammatiken und Wörterbüchern, sondern sie sind letztlich mehr oder weniger konvergierende, immer aber dynamische Mengen sozialer Praktiken. Eine Kummulation sozialer Praktiken zu konzeptualisieren ist weder für die Laien noch für die LinguistInnen einfach.<sup>32</sup> (Berthele 2010: 264)

Nevertheless, intuitive lay knowledge can sometimes be underestimated, as Berthele also notes:

Dabei dürfte sich nach meinem Dafürhalten langfristig zeigen, dass die laienlinguistischen Kategorisierungssysteme gerade im Bereich der Dialekte nicht grundsätzlich schlechter sind als die Systeme der ExpertInnen. Letztere versuchen in der Regel, hinreichende und notwendige Bedingungen zu formulieren, die Varietäten voneinander abgrenzen. Solche Bedingungen werden jedoch der Dynamik der Dialekte nicht gerecht [...].<sup>33</sup> (Berthele 2010: 263)

Whether consciously or intuitively, the knowledge that emerges from lay dialect collections deserves further attention, either because it is simply “on the doorstep” and has received little attention, even though it offers a very unmediated and immediate insight into what lay people consider “theirs”, or because it can provide another piece of the mosaic as a complement to other methods of perceptual dialectology.

---

<sup>32</sup>After all, languages and varieties do not exist in linguistic atlases, grammars, and dictionaries; they are ultimately more or less converging but always dynamic collections of social practices. Conceptualizing an assemblage of social practices is not easy, neither for the layperson nor for the linguist.’ [author’s translation]

<sup>33</sup>In my opinion, the lay language categorization systems are not fundamentally worse in the long run than the systems of the experts, especially in the field of dialects. Experts usually try to formulate sufficient and necessary conditions that distinguish varieties from each other. Such conditions, however, do not do justice to the dynamics of dialects [...].’ [author’s translation]

## Appendix

mhg. sound/ combination of sounds/ phonolog. process	example word	Paznaun	Ötztal	Waidring	Standard
i	<i>bist</i>	bif	bif	bist	bist
i/monosyllabic lengthening	<i>Tisch</i>	tif	tif	ti:f	ti:f
e	<i>Bett</i>	bet	betse	bet	bet
ë/monosyll. length.	<i>schlecht</i>	flëxt	flext	fle:xt	flëxt
ää/Dim.	<i>Bänklein</i>	banjkxli	banjkxle	banjkxé	benjklain
a/-e (masc.)	<i>Affe</i>	øf	øfe	øf	afé
o/-e (fem.)	<i>Woche</i>	voxe	vøxa	vox	vøxé
ö/Dim.	<i>Vögelchen</i>	feigeli	fe:gele	fe:ge	fœ:glçen
u/monosyll. length.	<i>Luft</i>	luft	luft	luft	luft
ü/-rst-	<i>Bürste</i>	pirste	bi:cta	bi:st	byrsté
â	<i>Abend</i>	ɔ:bet	ɔ:bnt	o:bnt	a:bënt
ê	<i>weh, Schnee</i>	væ ſnæ	væ ſnæ	ve: ſne:	ve: ſne:
î	<i>bleib</i>	blaip	blaip	blaip	blaip
ô	<i>groß</i>	grøas	grøas	grous	gro:s
œ	<i>größer</i>	grøesər	grøeser	gresə	grø:se
û	<i>laut</i>	laut	laut	laut	laut
iu (U)/-er	<i>Häuser</i>	haisər	haiser	haise	hɔise
iu (< eu)	<i>neu</i>	nui	nui	noi	nai
ie/-er	<i>lieber</i>	liebər	lieber	liebe	li:be
uo	<i>gut</i>	guet	gøet	guet	gu:t
üe/-e (masc. Pl.)	<i>Füße</i>	fies	fiesé	fies	fy:șe
ei/-e (fem.)	<i>Seife</i>	sa:fe	søfe	søef	saifé
ou/-en	<i>glauben</i>	gloube	glo:bm	gla:m	glaubm
öu/-e (masc. Pl.)	<i>Bäume</i>	bem	ba:mən	ba:m	bɔimə
-en	<i>fliegen</i>	fliege	fliegj	fliəj	fli:gj
-ên/-en	<i>stehen</i>	ſtie	ſteen	ſtē:	ſte:n
-œn/-n (after root vowel)	<i>schön</i>	ſie	ſœn	ſē:	ſœ:n
il	<i>Milch</i>	milx	milx	muix	milx
al	<i>Salz</i>	solts	solts	sɔits	salts
ol	<i>Holz</i>	holts	holts	hoits	holts
el/-en	<i>stellen</i>	ſtelle	ſteln	ſtoin	ſteln
ël	<i>Feld</i>	falt	falt	føjt	felt
or	<i>Korb, Dorf</i>	kxɔerp	kxarp	darf	kxorp
ôr	<i>Ohr</i>	ɔ:er	ɔ:er	ɔu	ɔe
œr/-en	<i>hören</i>	heere	heern	heen	hœen
ër	<i>Berg</i>	bërk	bark	bërk	bërk
-t/-er	<i>Wetter</i>	vetr	ve:ter	ve:də	vëte
-st/-er	<i>Schwester</i>	ſvestər	ſvester	ſvestə	ſvestə
-rt-	<i>fertig</i>	fërtik	fextik	festik	fertik
-rst/-e (fem.)	<i>Bürste</i>	birſte	bi:xta	bi:st	byrsté

mhg. sound/ combination of sounds/ phonolog. process	example	Paznaun word	Ötztal	Waidring	Standard
-rts-	<i>schwarz</i>	<i>svorts</i>	<i>svoxts</i>	<i>svofts</i>	<i>svarts</i>
ge-s	<i>gestorben</i>	<i>kſtørbe</i>	<i>kſtarbm</i>	<i>gſtorm</i>	<i>geſtørbm</i>
ge-l	<i>gelernt</i>	<i>gleernt</i>	<i>ggleernt</i>	<i>gleent</i>	<i>gelernt</i>
ge-w/strong, weak inflection	<i>gewesen</i>	<i>gvest</i>	<i>geve:sn</i>	<i>gve:sn</i>	<i>geve:sn</i>
ge-n/uo	<i>genug</i>	<i>gnuek</i>	<i>genuæk</i>	<i>gnuek</i>	<i>genuk</i>
ge-k	<i>gekannt</i>	<i>kxent</i>	<i>kxent</i>	<i>kxent</i>	<i>gékant</i>
ge-b/-en	<i>gebrochen</i>	<i>broxe</i>	<i>gębroxn</i>	<i>broxn</i>	<i>gębroxn</i>
ge-t/ān	<i>getan</i>	<i>to:</i>	<i>gętɔ:n</i>	<i>tū:</i>	<i>gęta:n</i>
ge-f/-en	<i>gefallen</i>	<i>gfole</i>	<i>gfɔln</i>	<i>gfoin</i>	<i>gęfahn</i>
-n (after root vowel)	<i>Wein</i>	<i>vai</i>	<i>vain</i>	<i>vai</i>	<i>vain</i>

## References

- Ammon, Ulrich, Hans Bickel & Alexandra N. Lenz. 2016. *Variantenwörterbuch des Deutschen: Die Standardsprache in Österreich, der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen*. 2nd edn. Berlin, Boston: de Gruyter.
- Anders, Christina Ada. 2008. Mental Maps linguistischer Laien zum Obersächsischen. In Helen Christen & Eveline Ziegler (eds.), *Sprechen, Schreiben, Hören: Zur Produktion und Perzeption von Dialekt und Standardsprache zu Beginn des 21. Jahrhunderts*, 201–227. Wien: Edition Praesens.
- Anders, Christina Ada, Nicole Palliwoda & Saskia Schröder. 2014. „in dem moment wo ich es dann erkenne dann ist es auch gleich wieder weg“: Salienzefekte in der Sprachperzeption. *Linguistik Online* 66(4). 51–69.
- Auer, Peter. 2004. Sprache, Grenze, Raum. *Zeitschrift für Sprachwissenschaft* 23(2). 149–179.
- Auer, Peter. 2014. Anmerkungen zum Salienzbegriff in der Soziolinguistik. *Linguistik Online* 66(4). 7–20.
- Baur, Gerhard W. 1987. Mundartwörterbücher für alle: Zu Möglichkeiten des Sammelns, Ordnens, Erklärens und Publizierens von Dialektwortschatz. In Landesstellen für Volkskunde in Freiburg und Stuttgart (ed.), *Beiträge zur Volkskunde in Baden-Württemberg*, vol. 2, 53–84.
- Beijering, Karin, Charlotte Gooskens & Wilbert Heeringa. 2008. Predicting intelligibility and perceived linguistic distance by means of the Levenshtein algorithm. *Linguistics in the Netherlands* 25(1). 13–24.

- Bergenholtz, Henning. 1989. Probleme der Selektion im allgemeinen einsprachigen Wörterbuch. In Franz J. Hausmann, Oskar Reichmann, Herbert E. Wiegand & Ladislav Zgusta (eds.), *Wörterbücher: Ein internationales Handbuch zur Lexikographie*, vol. 1 (Handbooks of Linguistics and Communication Science (HSK) 5/1), 772–779. Berlin: de Gruyter.
- Berthele, Raphael. 2010. Der Laienblick auf sprachliche Varietäten: Metalinguistische Vorstellungswelten in den Köpfen der Deutschschweizerinnen und Deutschschweizer. In Christina Ada Anders, Markus Hundt & Alexander Lasch (eds.), „*Perceptual Dialectology*“: *Neue Wege der Dialektologie* (Linguistik: Impulse & Tendenzen 38), 245–267. Berlin, New York: de Gruyter.
- Brückner, Dominik. 2006. Zur Lemmaauswahl im Klassikerwörterbuch. *Lexicographica* 22. 173–186.
- Denz, Josef, Bernd D. Insam, Anthony Rowley & Hans Ulrich Schmid. 2002. *Bayrisches Wörterbuch (BWB)*, vol. 1 (Bayerisch-Österreichisches Wörterbuch. II. Bayern). München: Oldenbourg.
- Diercks, Willy. 1988. Mental Maps: Linguistisch-geographische Konzepte. *Zeitschrift für Dialektologie und Linguistik* 55(3). 280–305.
- Dollmayr, Viktor & Eberhard Kranzmayer. 1963. *Wörterbuch der bairischen Mundarten in Österreich (WBÖ) / Band 1, 1. Lieferung: Vorwort, Einleitung, A - Achtung*. Wien: Böhlau.
- Dudenredaktion. N.d. *Duden online*. <https://www.duden.de/>.
- Eickmans, Heinz. 1980. Zur Gestaltung lokaler Mundartwörterbücher: Überlegungen anhand niederrheinischer Beispiele. *Niederdeutsches Wort* 20. 33–55.
- Elmentaler, Michael, Joachim Gessinger & Jan Wirrer. 2010. Qualitative und quantitative Verfahren in der Ethnodialektologie am Beispiel von Salienz. In Christina Ada Anders, Markus Hundt & Alexander Lasch (eds.), „*Perceptual Dialectology*“: *Neue Wege der Dialektologie* (Linguistik: Impulse & Tendenzen), 111–149. Berlin, New York: de Gruyter.
- Gooskens, Charlotte & Wilbert Heeringa. 2004. Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language Variation and Change* 16. 189–207.
- Haß-Zumkehr, Ulrike. 2001. *Deutsche Wörterbücher: Brennpunkt von Sprach- und Kulturgeschichte*. Berlin: de Gruyter.
- Heeringa, Wilbert. 2004. *Measuring dialect pronunciation differences using Levenshtein distance*. Groningen: University of Groningen. (Doctoral dissertation).
- Herrgen, Joachim & Jürgen E. Schmidt. 1989. Dialektalitätsareale und Dialektabbau. In Wolfgang Putschke, Werner Veith & Peter Wiesinger (eds.), *Dialektographie und Dialektologie: Günter Bellman zum 60. Geburtstag von seinen*

- Schülern und Freunden* (Deutsche Dialektgeographie 90), 304–346. Marburg: Elwert.
- Hettler, Ivonne. 2018. *Salienz, Bewertung und Realisierung regionaler Sprachmerkmale in Bremen und Hamburg* (Deutsche Dialektgeographie 124). Hildesheim: Olms.
- Hofer, Lorenz & Stefanie Meier. 2015. Mitwirkung der Sprachgemeinschaft im lexikographischen Prozess eines Dialektwörterbuchs. In Regula Schmidlin, Heike Behrens & Hans Bickel (eds.), *Sprachgebrauch und Sprachbewusstsein: Implikationen für die Sprachtheorie*, 117–132. Berlin, Boston: de Gruyter.
- Hundt, Markus. 2010. Bericht über die Pilotstudie „Laienlinguistische Konzeptionen deutscher Dialekte“. In Christina Ada Anders, Markus Hundt & Alexander Lasch (eds.), „*Perceptual Dialectology*“: *Neue Wege der Dialektologie* (Linguistik: Impulse & Tendenzen 38), 179–219. Berlin, New York: de Gruyter.
- Kapferer, Stefanie. 2020. *Entwurf eines computativen Messverfahrens zur Berechnung des Dialektalitätswertes von Laiendialektsammlungen: Samt Konzeption und Implementierung des Verfahrens in Form des “dialect measurement tools”*. Universität Innsbruck. (MA thesis).
- Klein, Karl Kurt, Ludwig Erich Schmitt & Egon Kühebacher. 1965. *Tirolischer Sprachatlas*, vol. 1: Vokalismus (Deutscher Sprachatlas: Regionale Sprachatlanten 3). Marburg: N. G. Elwert.
- Kranzmayer, Eberhard. 1956. *Historische Lautgeographie des gesamtbairischen Dialektraumes: Mit 27 Laut- und 4 Hilfskarten in besonderer Mappe*. Wien: Böhlau.
- Lameli, Alfred. 2004. *Standard und Substandard: Regionalismen im diachronen Längsschnitt* (ZDL-Beihefte 128). Wiesbaden: Steiner.
- Lameli, Alfred, Christoph Purschke & Roland Kehrein. 2008. Stimulus und Kognition: Zur Aktivierung mentaler Raumbilder. *Linguistik Online* 35(3). 55–86.
- Landolt, Christoph & Tobias Roth. 2020. Schweizerisches Idiotikon: Wörterbuch der schweizerdeutschen Sprache. In Alexandra N. Lenz & Philipp Stöckle (eds.), *Germanistische Dialektlexikographie zu Beginn des 21. Jahrhunderts* (ZDL-Beihefte 181), 143–173. Stuttgart: Steiner.
- Lenz, Alexandra N. 2010. Zum Salienzbegriff und zum Nachweis salienter Merkmale. In Christina Ada Anders, Markus Hundt & Alexander Lasch (eds.), „*Perceptual Dialectology*“: *Neue Wege der Dialektologie*, 2nd edn. (Linguistik: Impulse & Tendenzen 38), 89–110. Berlin, New York: de Gruyter. DOI: 10.1515/9783110227529.1.89.
- Lenz, Alexandra N. 2019. Variatio delectat? Vom Verhältnis von Lexikographie und sprachlicher Vielfalt. In Die Präsidenten der Berlin-Brandenburgischen Akademie der Wissenschaften und der Akademie der Wissenschaften zu Göttingen (ed.), *Paradigmenwechsel in der Lexikographie: Herausforderung*

- und Chance*, 64–88. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften.
- Nerbonne, John & Wilbert Heeringa. 2009. Measuring dialect differences. In Peter Auer & Jürgen E. Schmidt (eds.), *Language and space: An international handbook of linguistic variation*, vol. 1: Theories and Methods (Handbooks of Linguistics and Communication Science (HSK) 30/1), 550–567. Berlin, New York: de Gruyter Mouton.
- Österreichisches Wörterbuch: Vollständige Ausgabe mit dem amtlichen Regelwerk*, 43rd edn. 2018. Bearbeitet von Magdalena Eybl et al.; Redaktion: Christiane M. Pabst, Herbert Fussy, Ulrike Steiner. Wien: Österreichischer Bundesverlag.
- Palliwoda, Nicole & Saskia Schröder. 2016. Perceptual Dialectology, speech samples, and the concept of salience: Initial findings from the DFG-project “Lay linguists’ perspectives on German regional varieties: Reconstructing lay linguistic conceptualizations of German in a perceptual dialectology approach”. In Jennifer Cramer & Chris Montgomery (eds.), *Cityscapes and perceptual dialectology: Global perspectives on non-linguists’ knowledge of the dialect landscape* (Language and Social Life 5), 257–274. Berlin, Boston: de Gruyter Mouton.
- Preston, Dennis R. 1999. *Handbook of perceptual dialectology*, vol. 1. Amsterdam: John Benjamins.
- Preston, Dennis R. 2010. Mapping the geolinguistic spaces of the brain. In Alfred Lameli, Roland Kehrein & Stefan Rabanus (eds.), *Language and space: An international handbook of linguistic variation*, vol. 2: Language mapping (Handbooks of Linguistics and Communication Science (HSK) 30/2), 121–153. Berlin, New York: de Gruyter Mouton.
- Purschke, Christoph. 2011. *Regionalsprache und Hörerurteil: Grundzüge einer perzeptiven Variationslinguistik* (Zeitschrift für Dialektologie und Linguistik – Beihefte 149). Stuttgart: Franz Steiner Verlag.
- Purschke, Christoph. 2014. „I remember it like it was interesting“: Zur Theorie von Salienz und Pertinenz. *Linguistik Online* 66(4). 31–50.
- Retti, Gregor. 1999. Ein Internetfragebogen zur Verifizierung von Lexikoneinträgen. *Linguistik Online* 3(2). DOI: 10.13092/lo.3.1039.
- Schatz, Josef. 1903. Die tirolische Mundart: Mit einer Karte. *Zeitschrift des Ferdinandeums für Tirol und Vorarlberg* 47(3). 1–94.
- Schatz, Josef. 1993. *Wörterbuch der Tiroler Mundarten* (Schlern-Schriften 119/120). Neudruck der Auflage 1955/56. Innsbruck: Universitätsverlag Wagner.
- Schmidt, Jürgen E. & Joachim Herrgen. 2011. *Sprachdynamik: Eine Einführung in die moderne Regionalsprachenforschung* (Grundlagen der Germanistik 49). Berlin: Schmidt.

- Schnabel, Michael, Manuel Raaf & Daniel Schwarz. 2020. Bayerisches Wörterbuch. In Alexandra N. Lenz & Philipp Stöckle (eds.), *Germanistische Dialektlexikographie zu Beginn des 21. Jahrhunderts* (ZDL-Beihefte 181), 47–76. Stuttgart: Steiner.
- Schröder, Saskia. 2017. Die Verortung der eigenen Sprechweise im Makrobereich durch linguistische Laien. In Markus Hundt, Nicole Palliwoda & Saskia Schröder (eds.), *Der deutsche Sprachraum aus der Sicht linguistischer Laien: Ergebnisse des Kieler DFG-Projektes*, 47–82. Berlin, Boston: de Gruyter.
- Schröder, Saskia. 2019. *Sprachräumliche Praxis: Sprachraumkartierung in der Wahrnehmungsdialektologie* (Kieler Forschungen zur Sprachwissenschaft 10). Berlin, Bern, Wien: Peter Lang.
- Schweizerisches Idiotikon. 1881ff. *Wörterbuch der schweizerdeutschen sprache. Begun by Friedrich Staub and Ludwig Tobler [...] Vol. I ff.* Frauenfeld: Huber & Basel: Schwabe.
- Stöckle, Philipp. 2020. Wörterbuch der bairischen Mundarten in Österreich (WBÖ). In Alexandra N. Lenz & Philipp Stöckle (eds.), *Germanistische Dialekt-Lexikographie zu Beginn des 21. Jahrhunderts* (ZDL-Beihefte 181), 11–46. Stuttgart: Steiner.
- Wiesinger, Peter. 1983. Ergebnisse dialektologischer Beschreibungen: Areale Bereiche deutscher Dialekte im Überblick. In Werner Besch, Ulrich Knoop, Wolfgang Putschke & Herbert E. Wiegand (eds.), *Dialektologie: Ein Handbuch zur deutschen und allgemeinen Dialektforschung* (Handbooks of Linguistics and Communication Science 1/2), 807–900. Berlin, New York: de Gruyter Mouton.
- Wiesinger, Peter. 2009. Die Standardaussprache in Österreich. In Eva-Maria Krech, Walter Haas & Mariana Alvarez (eds.), *Deutsches Aussprachewörterbuch*, 229–258. Berlin, New York: de Gruyter.



# Chapter 15

## Applying the state-of-the-art tonal distance metrics to a large dialectal dataset

Ho Wang Matthew Sung<sup>a</sup>, Jelena Prokić<sup>a</sup> & Yiya Chen<sup>a</sup>

<sup>a</sup>Leiden University

From Séguy's (1971, 1973) early dialectometric studies to the application of Levenshtein distance in dialectometry (e.g. Heeringa 2004) nowadays, the calculation of phonetic distances between dialects has largely been focused on segments. Despite the fact that tonal languages make up to 70% of the languages in the world (Yip 2002: 1), tones are still largely neglected or simplified in comparative dialectological studies. In the current literature, there are very limited attempts to quantify tone distances between dialects. Furthermore, these methods were mostly designed for perceptual studies and most importantly, they were tested with a rather small dialect dataset (20 or fewer dialects, e.g. Yang & Castro 2008, Tang 2009). When the method is not tested on a larger dataset, it is unclear how many tones a particular representation used in a distance calculation method can differentiate, and if it can be applied across different languages. An ideal method to calculate tone distances should be able to differentiate all possible tones in a given set of dialects, and not be bounded to specific datasets only.

In this chapter, we have compared four ways in calculating tone distances using data from Yue and Pinghua dialects, two Sinitic branches from the Sino-Tibetan language family which are spoken in Guangdong and Guangxi provinces in Southern China. They are chosen for this study because Yue and Pinghua provide a huge amount of data, both segmental and tonal. Furthermore, their tonal inventories consist of a wide range of tone types, differing in pitch, contour, shape, and duration. Our results show that some methods are more suitable for dialectometry, though they still require further improvements before we can apply them to investigate tonal variation.



## 1 Introduction

Traditional dialectology (also known as dialect geography, Chambers & Trudgill 1998: 14) originated in Europe, pioneered in Germany and France, has primarily been centering around phonetics and lexicon, and syntax has entered the scene more recently. The study of phonetic variation has been heavily focused on segments, and not tones, despite the fact that 60-70% of the world's languages also include lexical tones. This focus is not surprising, however, since European languages mostly do not use pitch to differentiate word meaning. Although the traditional methodology of dialect geography spread to different corners of the world, the application of computational mapping, later computational processing of dialectal data, has mostly been developed and applied to European languages, such as Dutch (Heeringa 2004) and French (Goebl 1984). For the same reason, these computational methods do not concern tonal variation, and whether these methods are suitable for the tonal data is a current issue yet to be solved.

In Chinese dialectology, for example, there are numerous studies on dialects spoken in China (e.g. Chao's (1928) survey on Wu dialects, Grootaers's work on Shanxi (Grootaers 2003), Zhan's (2002) *Introduction to Yue dialects in Guangdong*), and it has a century-long tradition, but most studies on tonal variation are descriptive. This means that studies usually report the tonal inventory of a dialect after a fieldwork investigation, and/or tones are analyzed in terms of how they correspond to historical tone categories (from the Middle Chinese period, based on the rhyme dictionary descriptions). An example of such studies is Zhan (2002). Although there is a huge amount of dialect data available for Chinese (in the form of IPA transcriptions, including tones), currently we are still at the exploration stage in finding a methodology which allows dialectologists to measure tone distances for the purpose of dialect classification. The lack of analytic tools is not only a problem for dialectologists working with Chinese, but for dialectologists, as well as linguists, around the world. While most of the world's languages are tonal, we do not have proper methods which would allow us to investigate dialectal variation of tones, the most important aspect of tonal languages by which they differ from the non-tonal languages.

The structure of this paper is as follows: Section 1 briefly sketches what tonal languages are and their relative isolation in dialectometry. Section 2 gives an overview of the current approaches to quantifying phonetic tone distances, with more focus on the replicable methods. Section 3 introduces the Yue and Pinghua dialects, which serve as the case study for this paper, and their sources of data, as well as a sketch of the methodology used in comparing different methods in calculating tone distances introduced in Section 2. Section 4 presents the results

of the comparisons as well as their interpretations and lastly, Section 5 and Section 6 offer the discussion and conclusions respectively.

## 2 Tonal languages and dialectometry

According to Hyman (2006: 229), tones can be defined as the use of “pitch... [in] the lexical realization of at least some morphemes”. It has been estimated that around 60–70% of the languages in the world are tonal (Yip 2002: 1). Some relatively well-studied tonal languages include Cantonese, Mandarin and Yoruba.

The main types of tone contour are Level, Falling, Rising, as well as complex tones Concave (Falling-Rising) and Convex (Rising-Falling) tones. A level tone is a tone with a contour that is kept flat throughout the whole tone production.<sup>1</sup> A rising tone is when the pitch of the offset is higher than the onset, whereas a falling tone is the opposite. Concave and convex tones are the combinations of Rising and Falling tones. Languages and dialects may have more than one tone of the same contour. For example, Cantonese has three level tones, two rising tones and one falling tone. It has been found that across 737 Sinitic varieties (including over 500 Mandarin dialects), Falling tone is the most common tone (1125 tokens), closely followed by Level tones (1086 tokens). Rising (790 tokens) and Concave (352 tokens) tones are less common and Convex (80 tokens) tones are the least common (Cheng 1973).

There are very limited attempts to quantify tone distances between dialects, like Yang & Castro (2008) and Tang (2009). Furthermore, these methods were mostly designed for perceptual studies and most importantly, they were tested on a rather small dialect dataset (20 or fewer dialects). When the tone representation used in a particular method is not tested on a larger dataset, it is unclear how many tones it can differentiate, and if it can be applied across different languages. An ideal method to calculate tone distances should be able to differentiate all possible tones in a given set of dialects, and not be bounded to specific datasets only.

In this study we rely on the Yue and Pinghua dialects to explore the possibility of measuring tone distances between dialects using computational methods. We made this choice because Yue and Pinghua provide a huge amount of data, especially on tones. The tonal inventories of Yue and Pinghua dialects consist of a wide range of tone types, differing in pitch, contour, shape, and duration. A

---

<sup>1</sup>In reality, some of the Level tones also have a slight falling or rising F0 trajectory, but this paper focuses on impressionistic transcription data, which does not always capture such fine trajectories.

wide range of tone types is necessary for the development of a way to measure tone distances, since this method is aimed to be applicable to all tonal languages in the world, provided the suitable transcriptions are given. It should be noted that the transcriptions of tones for Yue and Pinghua are impressionistic, meaning that the distances cannot capture fine trajectories and it may be subjected to transcribers' differences. However, this is understandable, since a lot of the Yue and Pinghua data, as well as other Chinese varieties and possibly other tonal language dialect surveys, were collected before portable recorders were available for data collection.

### 3 Previous approaches to measuring tone distances

#### 3.1 Background

Most studies in dialectology (descended from the dialect geographical tradition) have been conducted on Indo-European languages and other non-tonal languages (including pitch-accent languages like Japanese). Therefore, there is a lack of dialect geographical studies that involve tones. For tonal languages such as Sinitic languages, most research is rather descriptive, i.e., listing the tonal inventory of a dialect or showing diachronic correspondences with historical tone categories, as in Zhan's (2002) *Introduction to the Yue dialects in Guangdong*, for example.

The same problem can be found in dialectometry, the subdiscipline of dialectology which applies quantitative and computational methods to study dialects. From Séguy's (1971, 1973) enterprise of aggregating linguistic features to the application of Levenshtein distance (e.g. Heeringa 2004), segments have always been the focus of the studies in dialectometry. Up to date, there are no large-scale dialectometric studies on tones; the highest number of dialects involved are no more than 20 (see Yang & Castro 2008, Tang 2009). In some dialectometric studies, tones were neglected (e.g. Wichmann & Ran 2019), others used a rather simplified method (e.g. Stanford 2012). In addition, there are studies on the correlation between phonetic distance and the perception of tones (e.g. Yang & Castro 2008), which do not focus on the application of these measures on dialect classification.

At least for Sinitic languages, there is an abundance of data available for tones (transcriptions), but not enough is known on the incorporation of these data with segments. Tones in Chinese dialectology are most commonly written in Chao's (1930) notation, known as the *tone letters*. The tone letters is a transcription system which consists of 5 digits, 1, 2, 3, 4, 5, and they represent different (possible)

contour levels in a tone. 1 represents the lowest contour level and 5 represents the highest, while the rests sit in between. When combined (as two digits or three digits) together, they can indicate a change in the contour, which represent the shape of the tone. For example, <sup>53</sup> is a falling tone, whereas <sup>213</sup> is a dipping tone (a falling contour followed by a rising contour).

In the following subsections we will discuss four previously proposed methods for measuring tone distances and point some drawbacks (see Section 3 for the data).

### 3.2 Onset-Contour-Offset

Onset-Contour-Offset (OCO hereafter) is a representation of tones proposed by Yang & Castro (2008). This representation gives a more phonetic representation of tones, instead of an abstract one, as its purpose is to approximate multiple cues of tones in the distance measure in order to generate a more accurate prediction for intelligibility between dialects.

OCO involves a transformation of the tone letters (5-level transcription, Chao 1930) into a representation which consists of three components: *Onset*, *Contour* and *Offset*, each represented with one character, except for Contour, which can have up to two characters. Onset and Offset are the starting and ending contour levels of the tone, and the Contour is the shape of the tone. For the contour levels, the original 5-level transcription is converted into three categories, which are *H(igh)*, *M(id)* and *L(ow)*. H represents levels 4 and 5, M represents 3 and L represents 1 and 2. For contours, the basic shapes include *R(ising)*, *F(alling)*, *L(evel)*, and the complex tones are represented by the combination of the basic shapes, hence it has up to two characters.<sup>2</sup> Examples of the Contour representations can be found in Table 1.

As an example, the OCO representation of 221 would be LLFL, and for 24, it would be LRH. To calculate tone distances, Yang & Castro (2008) applied the Levenshtein distance algorithm on the OCO representation. This is illustrated in Table 2.

When two tones with different lengths are compared (length of three and four, like in Table 2), the Onset (Slot 1) and Offset (Slot 4) are always aligned together. In this example, we can find two substitutions and one deletion out of four alignment slots. This yields a (3 / 4 =) 0.75 difference between the tone pair.

---

<sup>2</sup>The complex tones are additions to Yang and Castro's original proposal, as the original article did not account for these types of tone, but they are present in our dataset.

Table 1: Contours in OCO representation with examples

Representation	Contour	Examples
L	Level	11, 33
R	Rising	12, 35
F	Falling	31, 52
RF	Convex	131, 253
FR	Concave	213, 424

Table 2: Calculation of Levenshtein Distance between 221 and 24 in OCO representation

Slot 1	Slot 2	Slot 3	Slot 4	Operations	Distance
L	L	F	L	-	-
L	R	F	L	Substitution of L > R	1
L	R	-	L	Deletion of F	1
L	R		H	Substitution of L > H	1
			Sum		3

Last but not least, the aggregated distance is calculated by summing the tone distances of the tone pairs and dividing the sum by the number of words compared in pairwise comparison between two dialects. This is done for any dialect pair in the dataset.

OCO has shown high correlation with mutual intelligibility for both Zhuang and Bai in Yang & Castro’s (2008) study. Do & Lai’s (2021) results basically agree with Yang and Castro. However, Tang (2009: 122–125) found that OCO fails to yield a classification which can differentiate Mandarin from other Sinitic languages. Based on this result, Tang (2009: 125) argued that this representation does not “provide a handle on dialect affinity”.

It should be mentioned that in Yang & Castro (2008), a related representation, Onset-Contour, correlates significantly higher than OCO with intelligibility as a predictor for Zhuang. However, because this representation can differentiate even fewer tones than OCO (see Section 4), in this paper we focus on the OCO representation.

### 3.3 Tone-to-string

The tone-to-string method applies the Levenshtein distance algorithm directly on Chao's (1930) tone letters. The differences in the digits are not accounted in this method, i.e. a substitution from 2 to 1 costs the same distance as from 4 to 1. In addition, when a two-digit tone is compared with a three-digit tone, the first digit of the two-digit tone aligns with the second digit of the three-digit tone<sup>3</sup>, as shown in the example below. Note that in this approach short tones are not distinguished from the other tones in the dataset.

In the example in Table 3, we illustrate how tones 325 and 15 are aligned and the distance between them is calculated using Levenshtein algorithm. In this example, one substitution and deletion are required to convert 325 to 15, which yields distance of (2 / 3 =) 0.67 between two tones.

Table 3: Calculation of Levenshtein Distance between 325 and 15 with tone-to-string method

Slot 1	Slot 2	Slot 3	Operations	Distance
3	2	5	-	-
-	2	5	Deletion of 3	1
	1	5	Substitution of 2 > 1	1
Sum				2

The aggregated distance is again calculated as the sum of the tone distances of the word pairs, divided by the number of words compared in pairwise comparison between two varieties. Thus, the aggregated distance is calculated for any pair of varieties in the dataset.

Yang & Castro (2008) has found that applying Levenshtein distance directly on Chao's (1930) tone letters has a lower correlation to mutual intelligibility than OCO. Tang's (2009) evaluation also shows that this method yields quite a high number of misclassifications when compared to the traditional classification (separating Mandarin from other Sinitic languages).

### 3.4 Binary comparison

Cheng (1997: 53) proposed to calculate the similarities between dialects by measuring the ratio of shared items to all the items (segments and tones) that are

<sup>3</sup>This alignment is based on Tang (2009).

occurring in both varieties in a pairwise comparison. The shared and non-shared items are indicated in binary form, i.e. with 1 (present) and 0 (absent) in the data matrix. This is similar to Goebl's (1984) *Relative Identity Value*, but the data matrix is comprised using the bag-of-words approach (based on the sound inventory of all dialects in the data). Tang (2009: 105–106) has found that this method yields a number of misclassifications when compared to the traditional taxonomy (Mandarin vs. other Sinitic languages). In addition, Tang (2009: 114–115) weighted the tone inventories by lexical frequency (out of 764 items in the database). This method yields one misclassification more than the binary inventory comparison. Another related approach by Cheng (1991: 88–89) includes Middle Chinese sound categories (as indicated by the ancient rhyme dictionaries *Guangyun*<sup>4</sup>) in the data matrix, in addition to adding lexical weighting and comparing tone inventories, using more than 2700 words. The compared elements are not simply the synchronic sound segments in the inventories, but reflexes of certain Middle Chinese sound categories. For example, instead of stating Dialect A has 79 tokens of [p<sup>f</sup>], Dialect A now has 40 tokens of [p<sup>h</sup>] as a reflex of \*p<sup>h</sup> and 39 tokens of [p<sup>h</sup>] as a reflex of \*b. Tang's (2009: 131–132) evaluation shows that the result from Cheng's method highly resembles the traditional classification. Due to lack of proto-Yue reconstructions that we could use to annotate our data, the above method cannot be applied to our dataset. Therefore, we use the binary comparison approach.

Comparing the tonal inventory between dialects might be too simplistic, since dialects could share the same tonal inventory (with the same phonetic tone values), but the lexical distribution (Wells 1982: 78) of the tones could be different. For example, the Taishan and Kaiping dialects are members of the Siyi dialects which have an identical tonal inventory, with 5 tonemes which are phonetically the same (Zhan & Cheung 1990: 85–87). However, not all lexical items share the same tones. In Table 4, we have listed three examples where each item has a different tone in each dialect.

In order to account for the lexical distribution differences between dialects, the binary comparison has been modified from the inventory level (comparing differences in the tonemic inventory only) to the lexical level (comparing tone differences in word pairs). Instead of comparing the proportion of elements (tones) shared, the tone distance for each lexical item is calculated by identifying whether the tone is identical (distance = 0) or not (distance = 1). An overall aggregate distance for each dialect pair is then calculated.

---

<sup>4</sup>In Chinese dialectology, *Guangyun*, a rhyme dictionary which represents the phonology of Middle Chinese, is often used as a proto-system for modern Chinese dialects (You 2016: 85–86).

Table 4: Examples of lexical distribution differences in tones of two Siyi dialects

Item	Taishan	Kaiping
First syllable of ‘lychee’ 荔	lai <sup>31</sup>	lai <sup>21</sup>
‘inner’ 内	<sup>n</sup> dui <sup>31</sup>	lui <sup>22</sup>
‘caldron’ 锅	fu <sup>21</sup>	fu <sup>55</sup>

The binary method is a harsh measure which implies that there is no difference in the distance between similar and very different tones, i.e. 11 vs. 22 and 11 vs. 523 are equally distant.

### 3.5 Gandour-Harshman-Tang tone distance

Unlike the previous methods in Section 3.2–Section 3.4, the Gandour-Harshman-Tang tone distance is based on the perceptual experiment of tone distances published in Gandour & Harshman (1978). Their stimuli included a range of tones, which differ in contour, pitch and duration, and they asked listeners to rate their distances. Based on the perceived distances of the listeners (Thai, Yoruba and English), five dimensions have been extracted and interpreted as *Average Height*, *Direction* (Rising, Level, Falling), *Length*, *Extreme End Point* and *Slope*. These dimensions are underlying ‘cues’ which can explain the variance found in the perceptual tone distance matrix and have been implemented as multivalued features by Tang (2009: 125–126) as a way to measure objective tone distances. This is the tone distance measure which accounts for perception, unlike other methods, which have been criticized as “very crude and unrealistic”, since any substitution or changes in the pitch and contour have equal importance (Tang 2009: 125). However, a method as sophisticated as this still gave “highly unsatisfactory results” according to Tang’s (2009: 127) evaluation method, since it also cannot differentiate Mandarin from other Sinitic languages, just like OCO.

Based on the contour of the tone, a feature value is assigned for each cue under the criteria listed in Table 5 to represent a tone, following Tang’s (2009: 125) implementation. It should be noted that in Tang’s dissertation, only the last two digits of the tone are taken into account for Direction and Slope, which means that complex tones and their counterparts without their onset element will share the same values for these two cues.

It should also be noted that we made a slight modification to Tang’s (2009) implementation. For Direction, instead of assigning the values 0, 1, 2 to Level,

Table 5: Cues and their relative feature values in GH-T

Cue	Features	Feature value	Maximal difference
Average Height	Lower than 2.5	1	4
	Between 2.5 and 3.5	3	
	Higher than 3.5	5	
Direction <sup>a</sup>	Falling	0	2
	No change/ one digit tone	1	
	Rising	2	
Duration	1-digit (or short 2-digit) tone	1	2
	2-digit (or short 3-digit) tone	2	
	3-digit tone	3	
Slope <sup>b</sup>	Difference smaller than 3	0	1
	Difference bigger than 3	1	
Extreme endpoint <sup>c</sup>	Ends with 2, 3 or 4	0	1
	Ends with 1 or 5	1	

<sup>a</sup>last two digits

<sup>b</sup>last two digits

<sup>c</sup>final digit

Falling, and Rising, respectively, we have made Level as the middle value (i.e. 1). This change can make the distances between the directions more continuous (Rising-Level-Falling forming a continuum), rather than being arbitrary categorical values. Furthermore, since this method is Tang's (2009) implementation of the perceptual dimensions found by Gandour & Harshman (1978), this method will be addressed as the GH-T hereafter, which stands for Gandour-Harshman-Tang.

Here is an example how tone distance is calculated with GH-T, which is also illustrated in Table 6. The maximum difference two tones can have is 10, based on sum of the maximal difference each cue can have ( $4 + 2 + 2 + 1 + 1$ ). Take 551 and 12# as an example (# indicates a shorter duration of the tone, often found in checked syllables), their feature specifications of these tones are listed in Table 6. After deducting the absolute difference of each feature value and sum the differences, this value is the distance between the tone pair. 551 and 12# is one of the

possible pairs of tones with the greatest distance in Chao's (1930) transcription system, since it yields the maximum difference score this tone representation can produce.

Table 6: Example of tone distance calculation using GH-T

Tone string	Height	Direction	Duration	Slope	Extreme		Total
					end point		
551	5	0	3	1		1	
12#	1	2	1	0		0	
Difference	4	2	2	1	1		10

For the computation of GH-T, the Manhattan distance has been used. Manhattan distance in this context is the sum of absolute differences of the values for all the cues shown in Table 5 (maximum distance is 10). In addition, all the differences are divided by 10 so that the range of differences are kept between 0 and 1, just like the other metrics in the previous subsections. Again, the overall aggregated difference is calculated between all dialect pairs in the data.

### 3.6 Interim summary

The state-of-the-art methods in measuring tone distances range from a strict binary approach to a metric which deals with of multi-valued features (GH-T). It is important to compare these methods systematically in order to explore the applicability of these methods to a larger dialect dataset in the classification context. As we have previously mentioned, these methods have only been applied to datasets with 20 or fewer dialects. We would like to know which method(s) are more linguistically coherent and suitable for dialect classification and identify possible improvements, so that they can be applied to various tonal languages in the world.

## 4 Data and methodology

### 4.1 Data

The data of this study focuses on the dialects within the Yue-Pinghua-speaking area. Yue and Pinghua are two sub-branches of the Sinitic branch of the Sino-Tibetan language family, which are spoken in Southern China and the diaspora communities in North America, the UK, Malaysia etc. (Wu 2012).

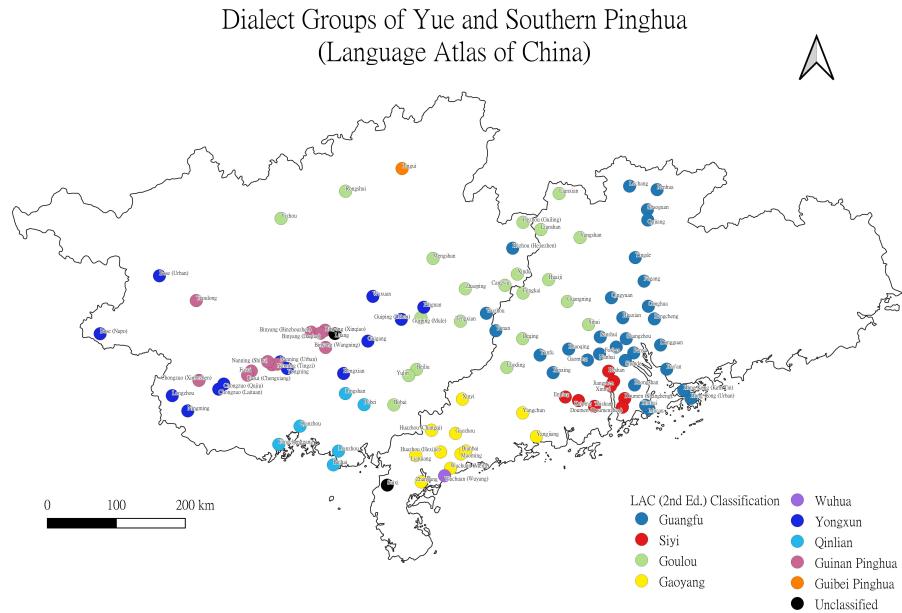


Figure 1: Dialect areas of Yue and Pinghua in Guangdong and Guangxi provinces (based on the *Language Atlas of China 1<sup>st</sup> Edition*, Li and Wurm1987)<sup>5</sup>

#### 4.1.1 Yue and Pinghua

Yue and Pinghua are mainly spoken in Southern China, namely in the Guangdong and Guangxi provinces (Chinese Academy of Social Sciences 2012). Their geographical distribution is shown in Figure 1<sup>5</sup> (based on the *Language Atlas of China 1<sup>st</sup> Edition*, Li and Wurm1987)<sup>6</sup>.

Traditionally, Yue has seven dialect groups, whereas Pinghua has two major dialect groups (Northern vs. Southern). There have been several attempts in classifying the dialects within Yue. Yue-Hashimoto (1988) performed the most extensive comparison over a number of Yue dialects, with over 100 features. However, her study suffered from the lack of data in certain regions, leaving her classification incomplete. Other classifications were either focused on one area (e.g. Zhan 2002 for the Guangdong province only) or unknown criteria were used

<sup>5</sup> All the maps in this paper are produced with QGIS (QGIS development team 2022).

<sup>6</sup> The first edition is used because the maps provide a more precise delimitation of the Yue and Pinghua areas than the second edition, since it is based on political boundaries instead of linguistic boundaries.

(e.g. Zhan 1981, Yuan 2001). The classification from the *Language Atlas of China* (LAC hereafter, 2nd Edition, Chinese Academy of Social Sciences 2012) covers the whole Yue-speaking region, and it stated clearly which criteria were used in the classification. A potential problem with the LAC classification is that very few features were used, and the motivation behind the choice of features has not been explained. However, since it is the only existing complete classification of Yue, it remains the representative of the traditional classification of Yue.

The status of Pinghua as a major branch of Sinitic languages has been controversial ever since its first proposal in the 1980s. Scholars have been arguing whether Pinghua belongs to Yue or not, and there have been proponents for both sides of the arguments. A group of scholars represented by Liang & Zhang (1999), Wei (1996), Li (2000) cited in Tan (2012) focused more on the differences (of features) between Pinghua and Yue, and argued that Pinghua should be independent from Yue. The opponents represented by Wu (2001), Tan (2000) and Liang (1997), cited in Tan (2012) examined a bunch of features and they mostly argued that a sub-branch of Pinghua, Guinan (Southern) Pinghua, is actually very similar to Yue, therefore it should be grouped under Yue. Liang (1997) made an even more radical claim that the whole Pinghua branch should be merged with Yue after examining seven features.

One of the reasons why the status of Pinghua is controversial is that each scholar focused on a different set of features, some on the differences (e.g. the reflex of Middle Chinese \*k<sup>h</sup>- is often a [k<sup>h</sup>-] in Pinghua, but [h-] in Yue, Li (2000): 36), while others on the similarities (e.g. the retention of codas -p, -t, -k, Wu (2001): 137).

#### 4.1.2 Segmental dialectometric classification of Yue and Pinghua

Our preliminary dialectometric result on the segmental variation (using classic Levenshtein distance) of Yue and Pinghua somewhat resembles the opponents of the Yue-Pinghua dichotomy, that Northern Pinghua is very different from the rest of the Yue and Southern Pinghua dialects (see Figure 2).

Figure 2 is a multidimensional scaling plot which reduces dimensions and approximates the distances calculated between all dialects in a 2D plot. The axes, labelled as dimension 1 and 2 represent underlying patterns which explains the greatest amount of variance found in the aggregated dialect distances. We can see that most of the orange dots (traditionally classified as Northern Pinghua) are located on the left (with high negative values in the first dimension). There is one exception, i.e. the Lingui dialect, which is located much closer to 0 in dimension 1, and also towards the rest of the dialects. Furthermore, there seems

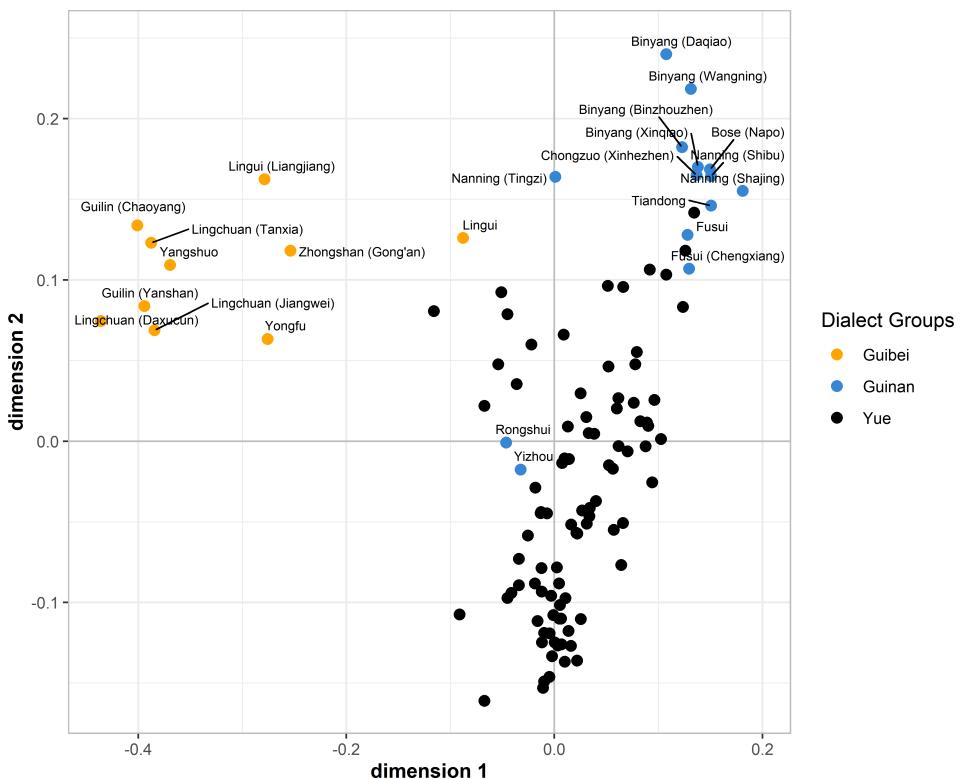


Figure 2: MDS plots of 113 Yue and Pinghua dialects according to LAC,  
 $r^2 = 0.59$

to be no continuum between the Northern Pinghua dialects and the rest of the dialects, which suggests abrupt boundaries between Northern Pinghua and the Yue continuum (in black). This result resembles the traditional analysis. To gain an understanding of how well the current methods in measuring tonal distances are, we have decided to focus on the Yue continuum, and remove the Northern Pinghua varieties. Like removing outliers in other statistical analyses, the Yue continuum (including Southern Pinghua) can then become less skewed due to the Northern Pinghua as outliers (in other words, less squished in dimension one), so that tonal variation can be interpreted more easily. Note that Lingui is now treated as part of Southern Pinghua based on its similarity with the rest of the dialects.

### 4.1.3 Data

The data used in this chapter is digitised by Sung et al. (2024).

A dialect survey is the published result of the data collected during fieldwork, like a linguistic atlas, but in tabular form (Francis 1983: 105–106). This is a typical method of publication in Chinese dialectology. The data used in this study only consist of IPA transcriptions of monosyllabic words, since not all dialect surveys have polysyllabic words recorded.<sup>7</sup> Tones, using Chao's (1930) tone letters, are also included in the transcriptions.

In total, data from six dialect surveys (Zhan & Cheung 1987, 1994, 1998, Shao 2016, Xie 2007, Chen & Lin 2009, Chen & Liu 2009) were extracted and digitized. These dialect surveys were conducted between the 1980s and 2010s and each of them contains transcriptions of more than 3000 items. Additional data from localities which the surveys did not cover came from mostly homonymic syllabaries in published sources (Liu 2015, Zhong 2015, Huang 2006, Chen 2009, Yang 2013, Tan 2017, Shi 2009, Chen & Weng 2010).<sup>8</sup> The localities and the data sources can be found in the map in Figure 3.

Around 130 (monosyllabic) words<sup>9</sup> are extracted and digitized from the original sources. The exact number of words per variety in the database varies as not every dialect contains the same number of transcribed words. The database contains 104 Yue and Southern Pinghua varieties (after removing the Northern Pinghua localities).

## 4.2 Methodology

We evaluate each tone distance measures based on i) number of tones distinguished, ii) comparison with perceptual dimensions (Gandour & Harshman 1978), iii) local incoherence, iv) comparison with traditional classification, and v) comparison with segmental classification (dialectometry).

An ideal tone distance measure is able to distinguish all the tones found in the data. By computing the distances between all the tones in the data, we can

<sup>7</sup>A modified version of the IPA is used in Chinese dialectology, with a number of non-standard characters such as the apical vowel [ɿ] which roughly represents [ɿ], ['] for aspiration, [A] and [E] for roughly [a] and [ɛ]/[ɛ] (between [e] and [ɛ]) respectively (Handel 2015).

<sup>8</sup>A list of (monosyllabic) words organized by their pronunciation; words are grouped by their rhymes, then their onset, and finally by their tone

<sup>9</sup>The word list is based on around 80 words from the Swadesh list, and the rest consists common words include an expansion in the domain of numbers, animals, directions, colours, as well as items which contain some notable Yue phonological features. The additional items in complementary to the Swadesh list is used to balance out some missing features that are not found in the Swadesh list.

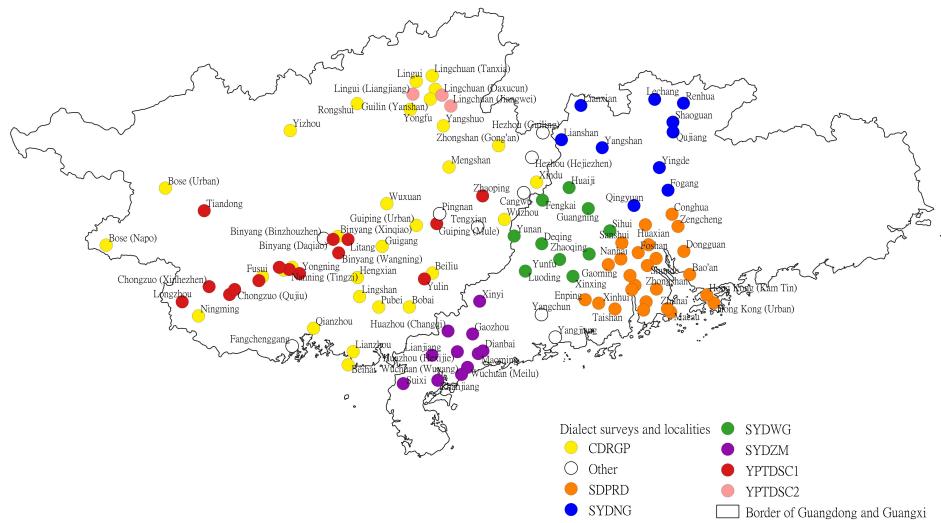


Figure 3: Localities and their respective sources

visualize these distances on an MDS plot, and inspect how much overlap there is between the converted tone representations, i.e. how much do tones share the same representation under each method. We also interpret the first two dimensions and check whether they match Gandour & Harshman's (1978) perceptual dimensions of tones. Previous studies in dialectometry have validated their methods through comparing the linguistic distances with perceptual distances (e.g. Gooskens & Heeringa 2004) and we apply the same approach in this study. For tones, Gandour & Harshman (1978) and Gandour (1983) have repeatedly found average pitch and direction as the most important dimensions in tone distance perception of native speakers from several tone languages. An ideal tone distance measure should also yield similar dimensions in the tone distances that each method produces.

Local incoherence is another external validation method used in dialectometry. Previous studies have shown that dialects closer to each other tend to be more similar than distant ones (the *Fundamental Dialectological Postulate*, Nerbonne & Kleiweg 2007). This pattern has been found not only within dialectal variation, but also in other domains (in Geography, it is known as *Tobler's First Law of Geography*, Tobler 1970). The idea behind local incoherence is to measure how much dialectal variation (in terms of the distances between the nearest dialects in any locality) matches the tendency stated above, when using a particular

method. Since the optimal score is 0, a tone distance calculation method that gets a small local incoherence score is considered as a more suitable method. Lastly, we want to know how similar or different the classifications generated by each method compare to the traditional and segmental classifications (from segmental dialectometry). The comparison can inform us whether tones behave similarly to segments as well as whether tones were used in the classification of dialects in the LAC (Chinese Academy of Social Sciences 2012), which may not have been explicitly stated.

The comparison between the tonal classifications and the traditional classification or the segmental classification requires additional steps. Firstly, cluster analysis (Ward's method) was performed on the segmental classification as well as the classification for each tone distance measure. The traditional classification in the *Language Atlas of China* originally gives ten different groups. However, in our dataset one of them consists of only two dialects (Suixi and Litang under 'Unclassified'<sup>10</sup>; the other two groups consist of one dialect (Huazhou (Hexijie) under 'Wuhua') and Lingui under 'Guibei Pinghua'). We have merged these two dialects with the Gaoyang dialect group, whereas Litang and Lingui now group with Guinan Pinghua, based on their geographical proximity in order to avoid clusters consisting of only one element.

To compare the similarity of classification (dialect groups) between each cluster solution, we have chosen to use the Adjusted Rand Index as an indicator to measure how much each cluster overlaps between a Reference Classification and an Observed Solution.<sup>11</sup> The Adjusted Rand Index (ARI hereafter) is a method used for comparing two different clustering solutions (with chance correction, Hubert & Arabie 1985), which derived from the Rand Index (Rand 1971). We are assuming that the traditional classification and the segmental classification are Reference Classifications, and the classification generated with each tone distance measure are the Observed Solutions, which are compared against the Reference Classification. If two classifications completely overlap, the score is 1, if they only overlap on the chance level, the score is 0. We are using the ARI scores for the comparison instead of regression for the following reasons. Firstly, the classification in the LAC is not empirically supported, which means we do not have a solid classification to compare to yet (see Section 3.1). This study also does not include a regression analysis to see how much each method can add to the explained variance to the segmental distances. Tones as a separate linguistic level has not been studied in dialectometry, and this current study aims to fill this gap.

<sup>10</sup>The LAC did not include this locality in their classification.

<sup>11</sup>In the literature, the reference classification is known as the Gold Standard. However, since the classifications we are using are not the absolute correct solution in our context (since there is no one 'correct' solution in dialectology), we use an alternate name instead.

The classification maps for the Reference classification and each of the tone distance measure can be found in the appendix.

## 5 Results

As described in Section 2, there are 4 tone distance measures that are found in the literature which are also easy to replicate. Firstly, we calculate the pairwise distances between all the dialects by calculating the tone distances for each lexical item (see Section 2 for the details of each method). We then calculate the aggregate pairwise distance between each dialect in the data by summing all the lexical tone distances divided by the total number of comparisons in the pair. This procedure is repeated for all the dialects in the data. For each method, the aggregate pairwise distances are stored in a distance matrix. These matrices are then used for the further analyses in the following sub-sections.

### 5.1 Tone overlap and perceptual dimensions

In this section, we will present Multidimensional Scaling (MDS) plots to visualize the distances between all the tones in the Yue data under different tone distance calculation methods introduced in Section 2. MDS is a dimension reduction method which represents “measurements of similarity (or dissimilarity) among pairs of objects as distances between points in a low-dimension multidimensional space” (Borg & Groenen 2005: 3). In our case, an MDS plot would represent the distances between tones (represented by labelled points in the plot), and the further the points are from each other, the more different they are. The MDS plots for the tone distances calculated using each method are shown from Figures 4–8.<sup>12</sup> Note that unless specified, the classical MDS has been used in the analysis.

The first evaluation on the tone distance calculation methods is based on the number of overlapping representations of tones in the dataset. An MDS plot is useful in visualizing tone overlaps, which are indicated by the total number of dots in the plot and the additional lines pointing at each overlapping representation in each plot. Gandour & Harshman (1978) and Gandour (1983) used the same technique to extract perceptual dimensions.

Table 7 summarizes the number of tones each method distinguishes from the original dataset, in both raw token and percentage. The total number of distinctive tones in the data is 73.

---

<sup>12</sup>Plots and explained variances were produced and calculated with *LED-A* (Heeringa et al. 2022), except Figure 8.

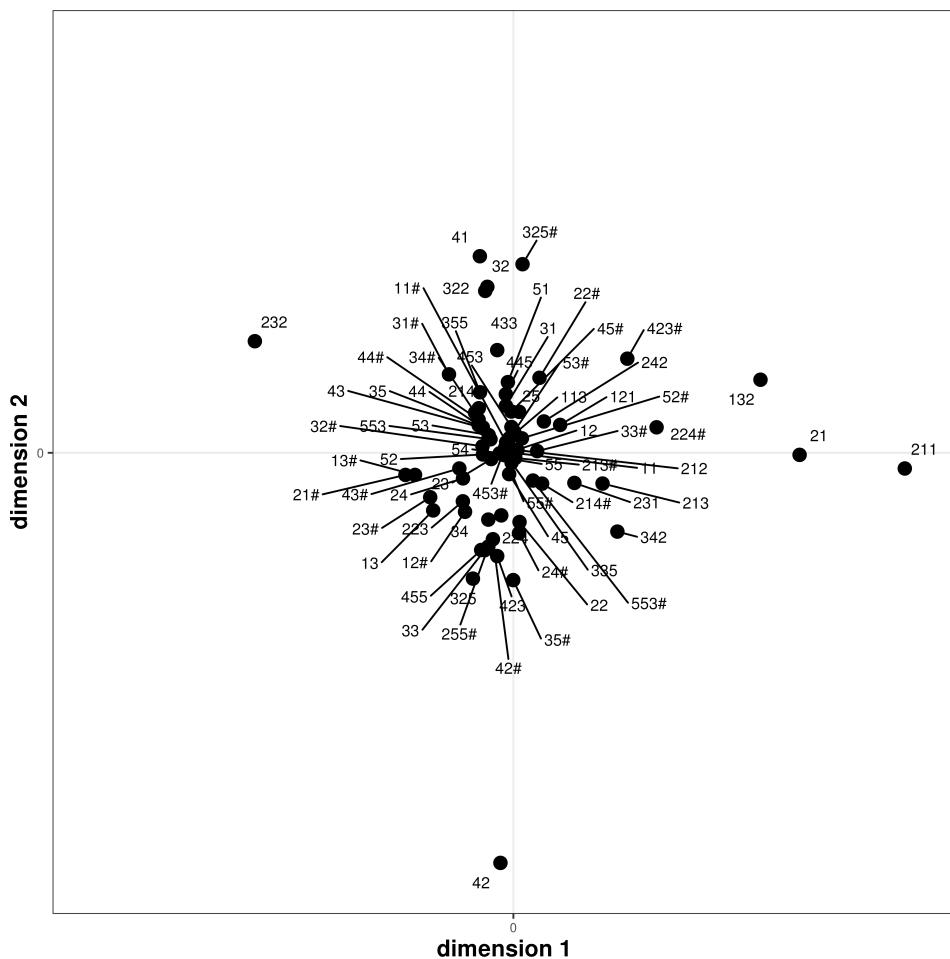


Figure 4: MDS plot for Binary comparisons (Classic MDS),  $r^2 = \text{n.a.}$

Table 7: Differentiation of tones under each tone distance measure

Method	Tones differentiated	Distinction rate
Binary	73	100%
Tone-to-string	52	71.2%
OCO	32	43.8%
GH-T	40	54.8%

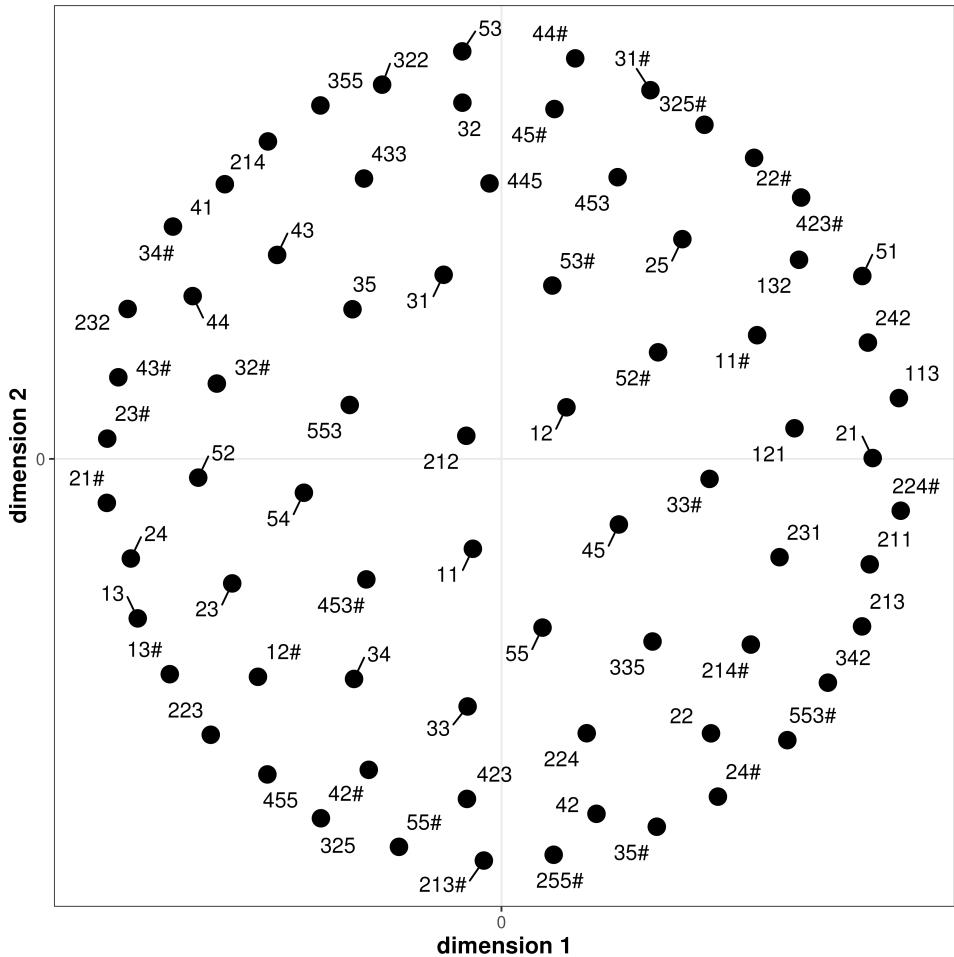


Figure 5: MDS plot for Binary comparisons (Sammon's MDS),  $r^2 = \text{n.a.}$

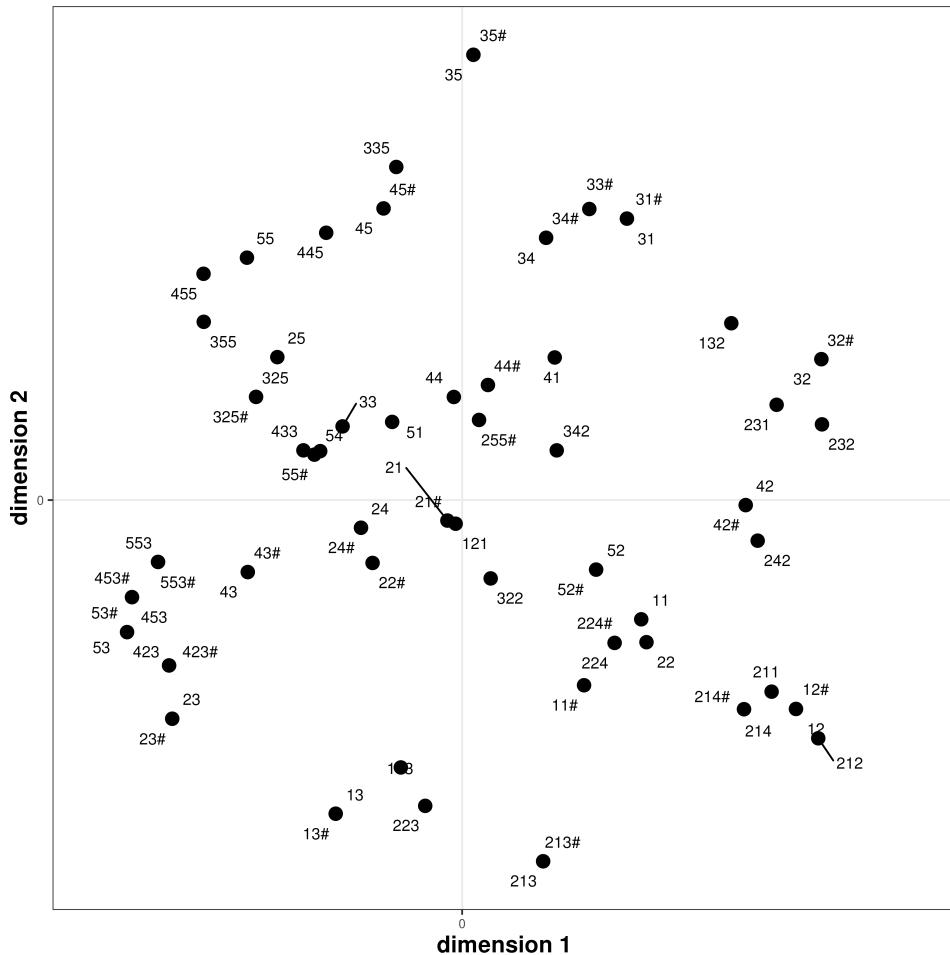


Figure 6: MDS plot for Tone-to-string,  $r^2 = 0.26$

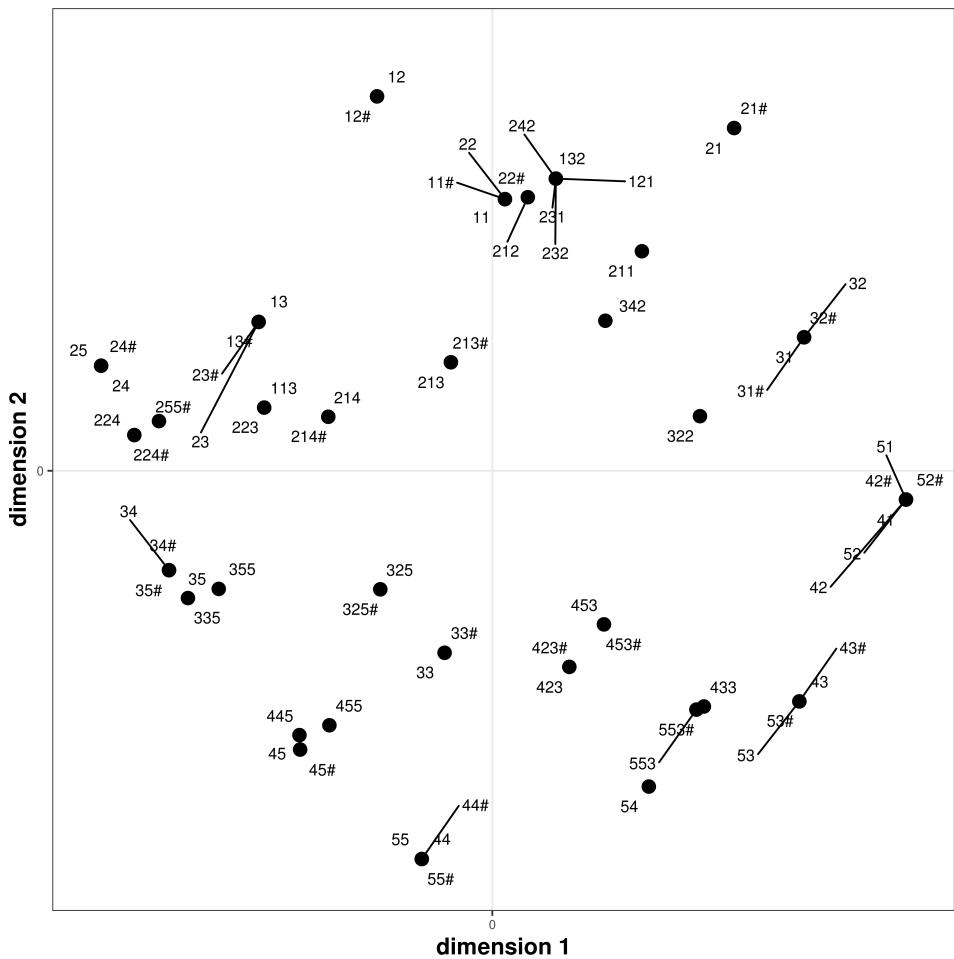


Figure 7: MDS plot for OCO,  $r^2 = 0.59$

In Table 7, we can see that the Binary method yields full distinction between all the tones in the data, and Tone-to-string can differentiate slightly over 70% of the tones in the data. For GH-T, only 54.8% of the tones can be differentiated, and lastly, OCO performs the poorest out of all methods, with under half of the tones in the data only.

The overlap of tones for the Tone-to-string method mainly occurs in cases where there is long and short distinction (indicated by the ‘#’ in Figure 6). Both Yang & Castro (2008) and Tang (2009) did not specify how they dealt with short phonetic tones, therefore, the current analysis has removed the short tone marker in the data. For the OCO method (Figure 7), in addition to the lack of length distinction, the contrasts between several tones with similar contour have been collapsed. For example, we can see a cluster of six tones on the right consisting of 41, 42, 42#, 51, 52, 52# sharing the same OCO representation. This overlap is the result of collapsing a five-level contour transcription system into a three-level representation. Lastly, in the GH-T representation (Figure 8), there are also overlapping tones with similar contours (e.g. the cluster at the top, consisting 42, 32, 43) or tones which share the same latter two digits (e.g. the cluster at the bottom consisting 23, 24, 224#, 423#). This is due to how tones are encoded with their features (see Table 5) as well as how tones with two and three digits align (see Section 3.5). A potential method in maximizing the distinctions of each method will be discussed in Section 5.

Gandour & Harshman (1978) and Gandour (1983) have consistently found that cross-linguistically, the two major perceptual dimensions in tones are average pitch and direction. The first two dimensions extracted from the MDS plots are summarized in Table 8.

Table 8: Interpretation of the first two dimensions in the MDS plots of each tone distance measure

Method	Dimension 1	Dimension 2
Gandour & Harshman (1978), Gandour (1983)	Average Pitch	Direction
Binary	Uninterpretable	Uninterpretable
Tone-to-string	Uninterpretable*	Uninterpretable*
OCO	Direction	Average Pitch
GH-T	Average Pitch	Direction

Firstly, the MDS plots for the Binary method requires some explanations. Figure 4 is plotted using the Classical MDS and Figure 5 is plotted using Sammon’s

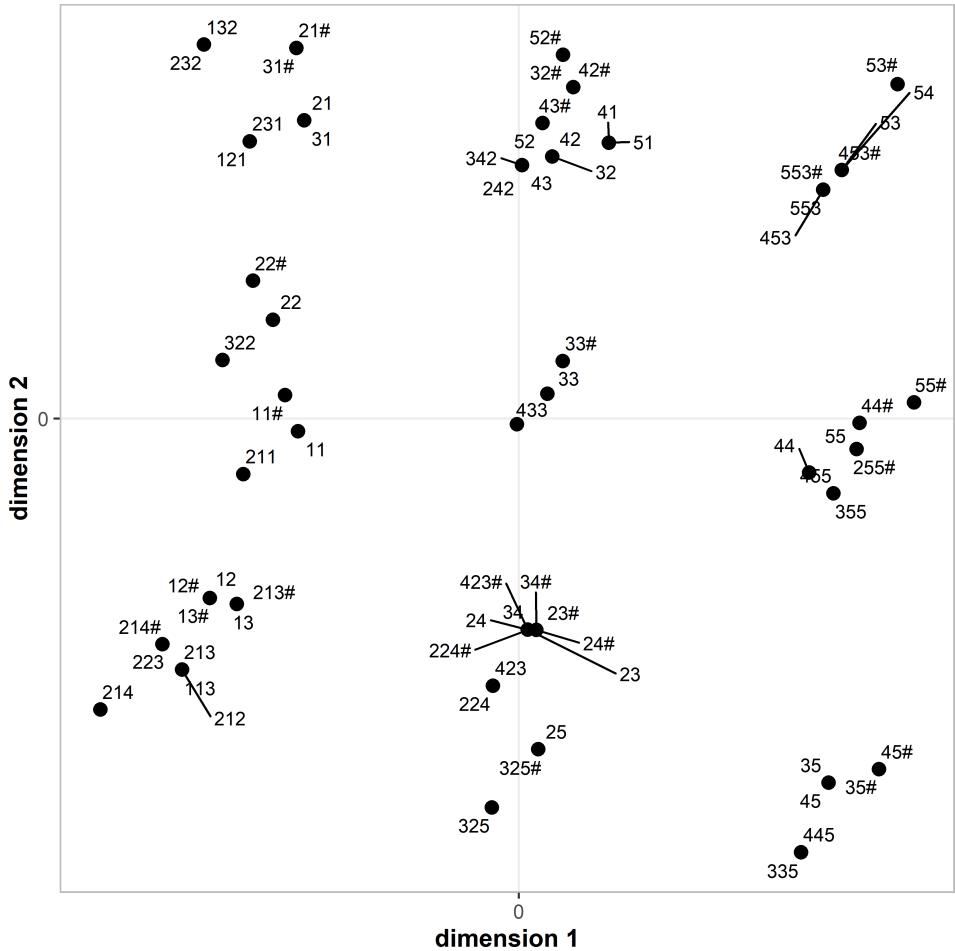


Figure 8: MDS plot for GH-T,  $r^2 = 0.74$

Mapping. Classical MDS was first used because the distance data is ratio data (Borg & Groenen 2005: 23). However, we realized that the plot in Figure 4 does not make much sense. There is a central cluster, with some tones in isolation, e.g. 42, 211, but the nature of the binary method should not yield any clusters (as shown in the middle of Figure 4), since all the tones are equally distant to each other. Since the relationship between the tones may not be 2-dimensional<sup>13</sup>, this gave us the motivation to go for Sammon's Mapping (Figure 5), as it is designed for non-linear data structure (Sammon 1969), and the plot resembles the distance matrix that we produced, where each tone is only similar (distance = 0) to itself, and all the tones are dispersed.

Summarized in Table 8, the first two dimensions of the Binary method (Figure 5) are uninterpretable. One of the reasons could be because the distance matrix was constructed based on how all the tones in the data are equally different (except itself), so that the MDS plot is just displaying all the tones without taking into account the room of gradual similarity between the tones in the data (which the Binary method did not measure at all). The MDS plot for Tone-to-string (Figure 6) is also uninterpretable if we only focus on Dimensions 1 and 2.<sup>14</sup> However, tones in Figure 6 are clustered together loosely by the pitch level that the tone ends in. This suggests the plot is a product of the operations and calculations on the string (of the tone letters), instead of the actual linguistic component of the tones. The OCO representation (Figure 7) shows a clearer picture than the previous two methods. For Dimension 1, we can see that on the right of the y-axis, the tone contours tend to be falling (e.g. 32, 52 and 53), and on the left, the contours tend to be rising (e.g. 12, 25, 45). For the level tones, they somewhat sit near the middle (y-axis). For Dimension 2, we can see that the tones above the x-axis tend to have a low onset, and their average pitches are mostly below 2.5. On the other hand, the tones under the x-axis are the opposite. Lastly, the GH-T plot (Figure 8) shows nine clear clusters across the plot. The first dimension also shows average pitch differences; if we look at the dots around the x-axis, we can easily identify that the low-level tones are clustered on the left, mid-level tones are clustered in the middle, and high-level tones on the right. The second dimension also quite clearly suggests direction. If we focus on the dots around the y-axis, we will dis-

<sup>13</sup>However, LED-A did not provide the explained variance value, nor did Gabmap work with this type of distance matrix (everything has a distance of 1 except itself).

<sup>14</sup>In terms of the procedures, it is possible to extract dimensions beyond dimension 2. With three and four dimensions, the cumulative explained variances are 0.40 and 0.55 respectively. However, if the first dimension, which has the highest variance explained for the distance matrix, is not even interpretable, the higher dimensions, which have lower explained variances, probably will not yield more insights and validity to the method.

cover that the clusters of tones go from falling to level to rising, starting from the top.

## 5.2 Local incoherence

Local Incoherence shows how different/ similar a dialect is to its nearest dialects (see Section 4.2). The local incoherence scores of each of the methods are calculated using *Gabmap* (Nerbonne et al. 2011, Leinonen et al. 2016). The results are shown in Table 9.

Table 9: Local Incoherence scores for each tone distance measure and segments. Lower values represent better fit to the Fundamental Dialectological Postulate.

Distance measure	Local incoherence
OCO	4.51
Tone-to-string	3.63
Binary comparison	3.74
GH-T	4.36
Segments (classic Levenshtein)	2.10

The optimal score of Local Incoherence is 0, and from Table 9, we see that none of the tone distance measures gets closer to 0 than the segmental level. Within the tone distance measures, Tone-to-string gets the lowest score, although it is not much lower than Binary comparison. OCO appears to be the method that produces the highest local incoherence, compared to the more linguistically naïve methods, while GH-T sits in between Binary and OCO.

In general, all tone distance measures have higher local incoherence than segments. This suggests two possibilities. The first possibility is that none of these measures are currently good enough to capture geographical variation of tones. The second possibility is that tones behave differently from segments, i.e. tones do not follow the *Fundamental Dialectological Postulate*. An additional contribution to a high local incoherence might come from similar varieties (Guangfu and Yongxun dialects, see Appendix Map 1) distributed in such a long distance apart, with the Yongxun dialects surrounded by dissimilar dialects in the west (Sung 2023). In order to explore these possibilities further, the tone distance measure should be refined first, so that inherent problems of each measure can be eliminated from the investigation, and a smaller region (e.g. Guangdong) can be inves-

tigated, in order to avoid the irregularity of distance introduced by the Yongxun dialects in Guangxi.

### 5.3 Adjusted Rand Index (ARI)

The cluster solutions of each of the tone distance measures (calculated with *Gabmap* (Nerbonne et al. 2011, Leinonen et al. 2016)) are compared to the traditional taxonomy from the *Language Atlas of China* as well as the segmental distances (also calculated with classical Levenshtein distance with *Gabmap*<sup>15</sup>) by using the ARI. ARIs indicate how much cluster solutions overlap with the Reference Classification (see Section 3.2). Results are shown in Table 10.

Table 10: ARI Similarity matrix between tone distance measures, traditional classification and segmental classification. Higher scores represent more overlap.

	LAC	Segments	Binary	Tone-to-string	OCO	GH-T
LAC	1.00	0.47	0.09	0.21	0.25	0.18
Segments	0.47	1.00	0.13	0.17	0.19	0.14
Binary	0.09	0.13	1.00	0.42	0.32	0.34
Tone-to-string	0.21	0.17	0.42	1.00	0.44	0.42
OCO	0.25	0.19	0.32	0.44	1.00	0.50
GH-T	0.18	0.14	0.34	0.42	0.50	1.00

The ARI score of 1 represents a complete overlap of cluster solutions between two classifications. In Table 10, the highest ARI score belongs to the pair of OCO and GH-T. In fact, OCO, GH-T and the Tone-to-string method all share high resemblance in terms of the classifications they yield. In fact, they are even higher than the ARI between the traditional classification and the segmental classification.

When we set the Reference Classification to LAC, the closest classification to the traditional taxonomy is the classical Levenshtein segmental classification, with 0.47. In terms of tone classifications, OCO had a score of 0.25 followed by Tone-to-string with 0.21. Between the segmental classification and the tone

<sup>15</sup>Another segmental distance measure, PMI Levenshtein (Wieling et al. 2011), which automatically infers segmental distances from 0 to 1 (instead of purely 0 or 1 in classical Levenshtein) based on the co-occurrence of segments, is not used. This is because the segmental distances inferred do not resemble to their acoustic distances (e.g. vowels do not show height and backness), unlike other languages tested using the same method in Wieling et al. (2011).

distance calculation methods, OCO resembles segments the most, with an ARI score of 0.19, while Tone-to-string follows with an ARI of 0.17. Overall, the Binary method has the lowest ARI score with almost all classifications. Although OCO and Tone-to-string resemble the traditional and segmental classifications the most respectively, their ARI scores are rather low, suggesting rather dissimilar classifications.

## 6 Discussion

### 6.1 Summary of results

The overall results are summarized in Table 11.

Table 11: Overall results of evaluation across 4 tone distance calculation methods

Method	Overlaps	Perceptual dimensions	Local incoherence	ARI	
				Traditional classification	Segmental classification
Binary	None	Uninterpretable	Intermediate	Lowest	Intermediate
Tone-to-string	Intermediate	Uninterpretable	Lowest	Intermediate	Intermediate
OCO	Highest	Resembles GH	Highest	Highest	Highest
GH-T	Intermediate	Identical to GH	Intermediate	Intermediate	Lowest

### 6.2 Which method to choose?

Each tone distance calculation method seems to do better than the others in one specific task. OCO has shown the best result with respect to two evaluation measures, while others performed best in one task each. This does not automatically make OCO the go-to method for the dialect classification of tones, though. OCO suffers the most from the overlapping representations, whilst the Binary method could distinguish all tones. The Binary method, however, cannot differentiate similar vs. very distinct tones, which makes the whole distance calculation rather arbitrary. For this reason, the Binary method is not suggested to be used in dialectometry.

In terms of perceptual dimensions, OCO shows some resemblance to Gandour & Harshman's (1978)'s findings, but the relative importance of the dimensions

is not in the right order, as reflected in the MDS plot (see Table 6). This aspect, however, could be fixed, so OCO has some potentials in this sense.

Looking at local incoherence, the best performing method is Tone-to-string, but like the Binary method, this method is also linguistically naïve. In addition, it's first two dimensions on the MDS are uninterpretable, hence this method is also not recommended in dialectometry. However, readers should be reminded that local incoherence should be taken as a grain of salt here, as mentioned in Section 5.2: The possibility that tones do not behave like segments and the dialect islands (Yongxun dialects) in the Guangxi province surrounded by dissimilar dialects both might contribute to the higher local incoherence.

As suggested in Section 5.3 and based on the observation that there is low resemblance between the traditional, segmental and tonal (OCO, GH-T) classifications, there is a possibility that the nature of tonal variation is different from segments. Therefore, we cannot simply use the ARI scores as a determining factor for choosing one tone distance calculation over another. OCO shares the highest ARI score with GH-T (and to some extent, the perceptual dimensions), which makes both methods valid to be used in dialectometry, given that some improvements are made (e.g. increase the distinction rate). However, there is one consideration which makes us favour one method over another – the ability to combine tonal and segmental analysis. The segmental analysis is done using Levenshtein Distance, as does the OCO representation. Ideally, the method for calculating tone distances is not limited to a tone-only analysis, but also a combined analysis with segments. This will not be possible with GH-T, since it requires a separate specification of tone features. For this reason, we argue that the OCO method should be chosen and further improved.

### 6.3 Possible improvements

The most fundamental problem OCO has is being unable to distinguish tones in the dataset (not even 50%). The first thing we should do is to find a way to take duration into consideration, and secondly, modifying the three-level distinction into five, since having five levels seems to be the maximum used by any language (Yip 2002: 20). Another issue which has been pointed out is that some tones, although they have different contours, are still yielding the same difference under OCO.<sup>16</sup>

We illustrate this with two tone pairs, 53 vs. 31 and 12 vs. 31. The former pairs are Falling tones (with the same slope), only differing in pitch levels. The latter

---

<sup>16</sup>Thanks to J. Fruehwald for his comment at Methods in Dialectology XVII.

pair are Rising and Falling tones respectively, and they have different degrees of slope. Many linguists would agree that one tone should be more different from 31 than the other. If we use the OCO representation to calculate tone difference, we get the same distance for these two tones. This is illustrated in Figure 9.

	53 vs. 31				12 vs. 31		
	Onset	Contour	Offset		Onset	Contour	Offset
53	H	F	M	12	L	R	L
1 subs.	<b>M</b>	F	M	1 subs.	<b>M</b>	R	L
1 subs.	M	F	<b>L</b>	1 subs.	M	<b>F</b>	L
31	M	F	L	31	M	F	L
Total: 2/3				Total: 2/3			

Figure 9: Comparison of tones 53 and 12 with 31

As illustrated in Figure 9, 53 vs. 31 only consisted of pitch differences, while 12 vs. 31 involves contour as well. However, using OCO representation this cannot be detected. In order to distinguish differences in these cases, perhaps weighting (onset, offset and/or contour) should be tested on each parameter<sup>17</sup>. Alternatively, other features (based on perceptual cues, see Gandour & Harshman 1978) shall be explored.

#### 6.4 Tonal variation in Yue and Pinghua

Lastly, we would like to make a remark on the patterns of tonal variation in the Yue and Pinghua-speaking so far. Although in general the ARI scores of the tone distances generated from each method and the segmental distances are relatively low, by visual inspection of the cluster maps (see Appendix), Tone-to-string, OCO and GH-T yield some clusters which resemble the segmental clusters, as

<sup>17</sup>Reviewer 2 has suggested introducing *steps* when dealing with substitution in the OCO representation. For example, if Onset or Offset has a High, then a substitution to Mid would be 1, but to Low would be 2. Similarly, for Contour, a substitution from Rising to Level would be 1, but to Falling would be 2.

well as the LAC a lot. For instance, Siyi dialects (consisting of 5 localities, Taishan, Kaiping, Enping, Xinhui and Doumen) can be identified, which align with both segmental and traditional classification. Another notable observation for the same methods is that we can see the East-most cluster (traditional Guangfu dialects) spread westward towards Guangxi. This pattern matches the segmental analysis, which can be explained by historical migration around 150 years ago (de Sousa 2022: 268). The other patterns are yet to be explained<sup>18</sup>.

To further understand the tonal variation of Yue and Pinghua (and any other tonal languages), therefore, we first need an improved way to calculate tone distances. That will allow us to find ways tones differ between dialects (e.g. is it gradual or abrupt?), to explore whether tones may behave differently from segments (as we have suggested in Section 4.2 with local incoherence) as well as the possibility of combining segments and tones in the same analysis.

## 7 Conclusion

Traditionally dialectology has focused on segments, and not much attention has been paid to tones, despite the wealth of data we have on tone languages, such as Sinitic languages. The limited research in exploring tone distance measures has only been tested on a small dataset. Hence, far from enough systematic comparison has been made on the existing methods for the purpose of dialect classification. This paper offers such systematic comparison, with a dataset which comprises of 104 dialects. Our results show that the OCO and GH-T representations make meaningful contributions to dialectometry, but they are not yet ready for the use in measuring tonal distances. We have suggested possible improvements and discussed the coherence problem between the tone distance calculation methods and tone changes.

Throughout the comparison, we repeatedly see that tones do not follow the same variation patterns as segments. Perhaps tonal variation really has a different nature in how it varies geographically. This is a rather unexplored area which awaits more research and improved methodology, and we hope that this paper is the first step in that direction.

---

<sup>18</sup>Since the submission of this chapter, modifications of the OCO representation has been made and further analyses on tonal variation have been done. Please refer to Sung & Prokić (2024) for the follow up study.

## Appendix

Dialect Groups of Yue and Southern Pinghua  
(Language Atlas of China 1st Ed.)

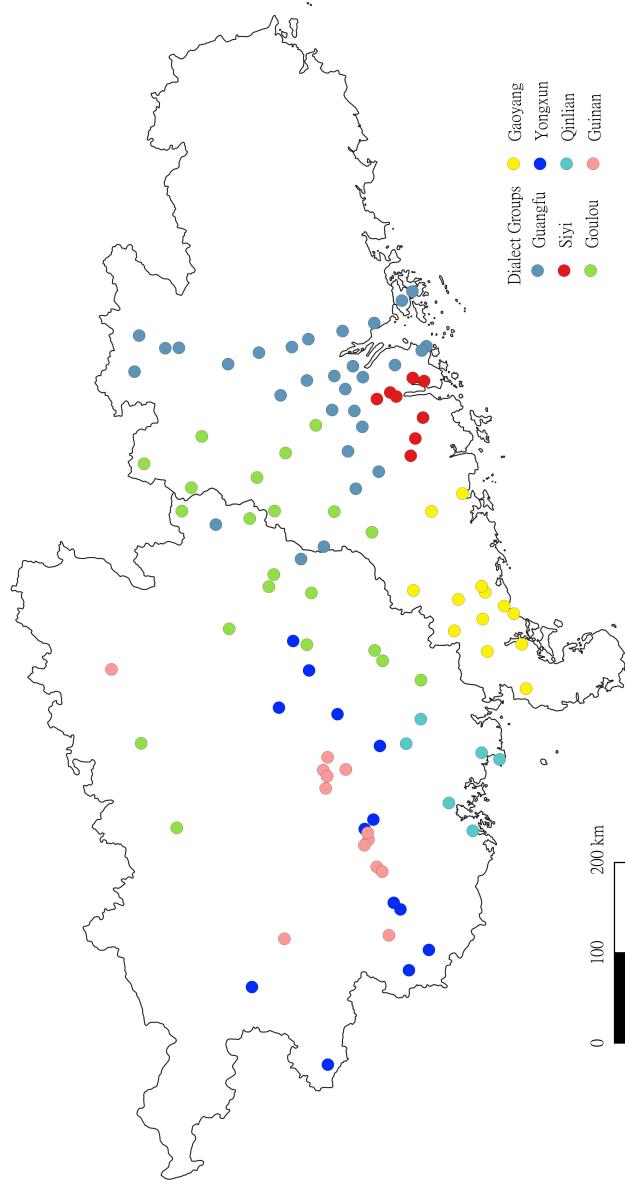


Figure 10: Traditional classification of the dialects in the Yue-Pinghua data (based on Language Atlas of China)

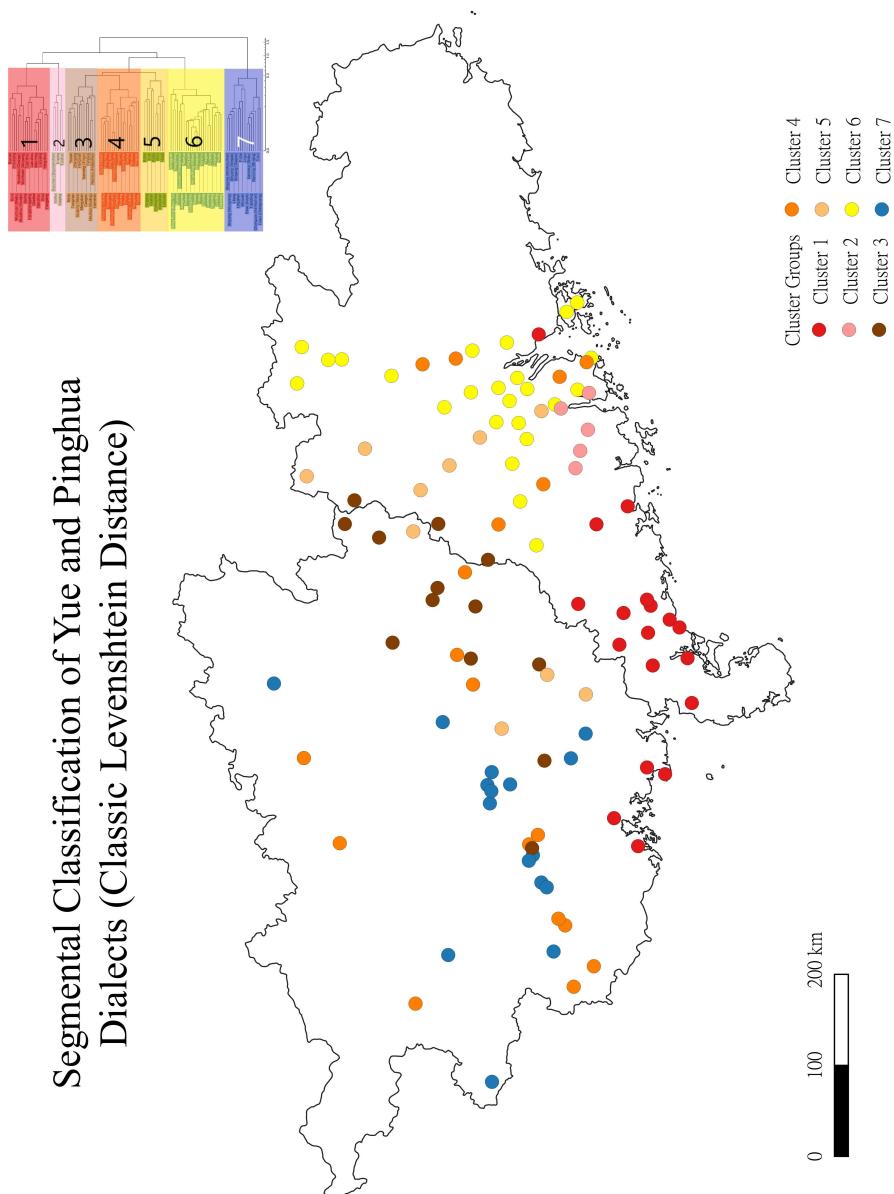


Figure 11: Cluster map (Ward's method) of the segmental classification of the Yue-Pinghua data

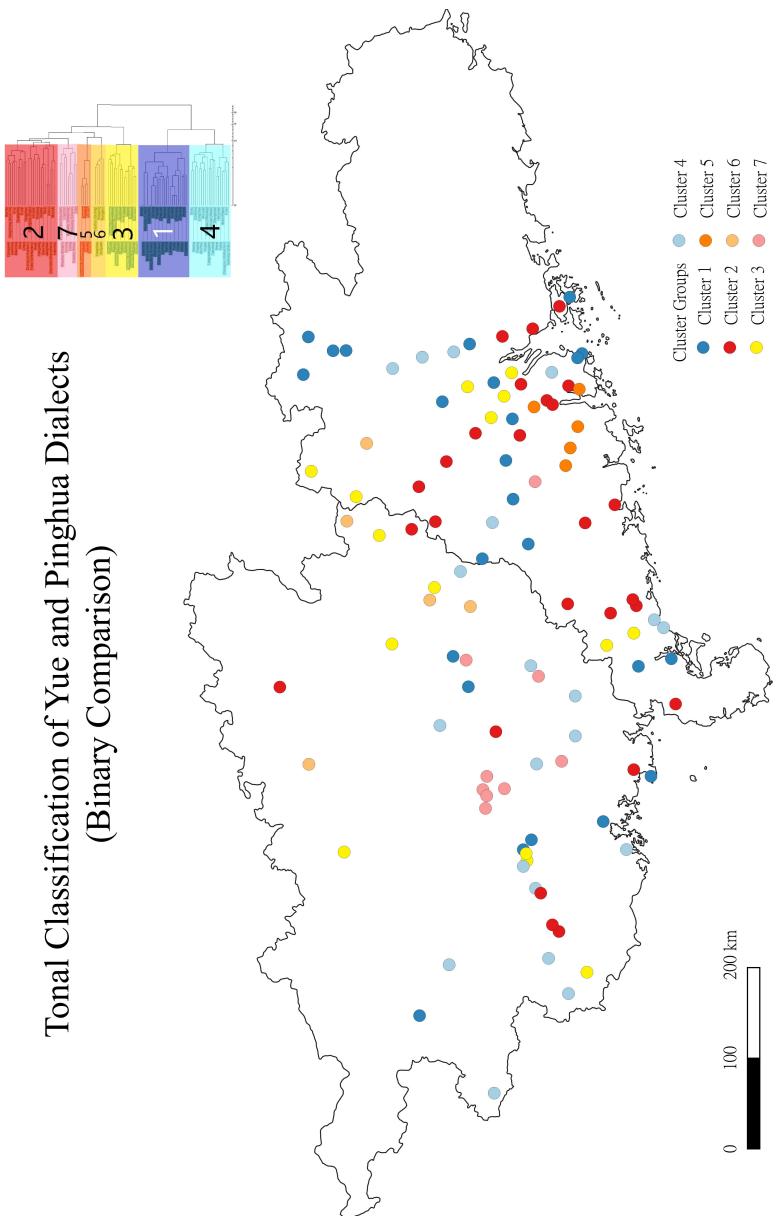


Figure 12: Cluster map (Ward's method) of the tonal classification of the Yue-Pinghua data (Binary comparison)

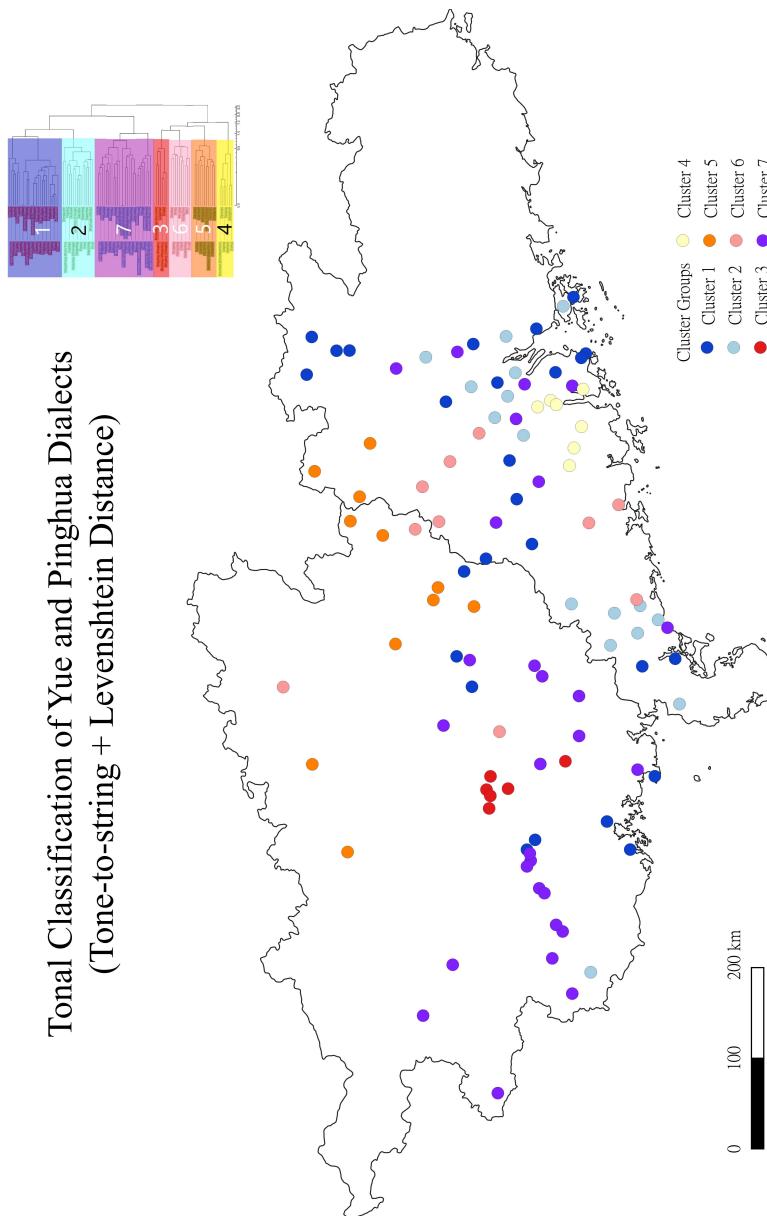


Figure 13: Cluster map (Ward's method) of the tonal classification of the Yue-Pinghua data (Tone-to-string + Levenshtein distance)

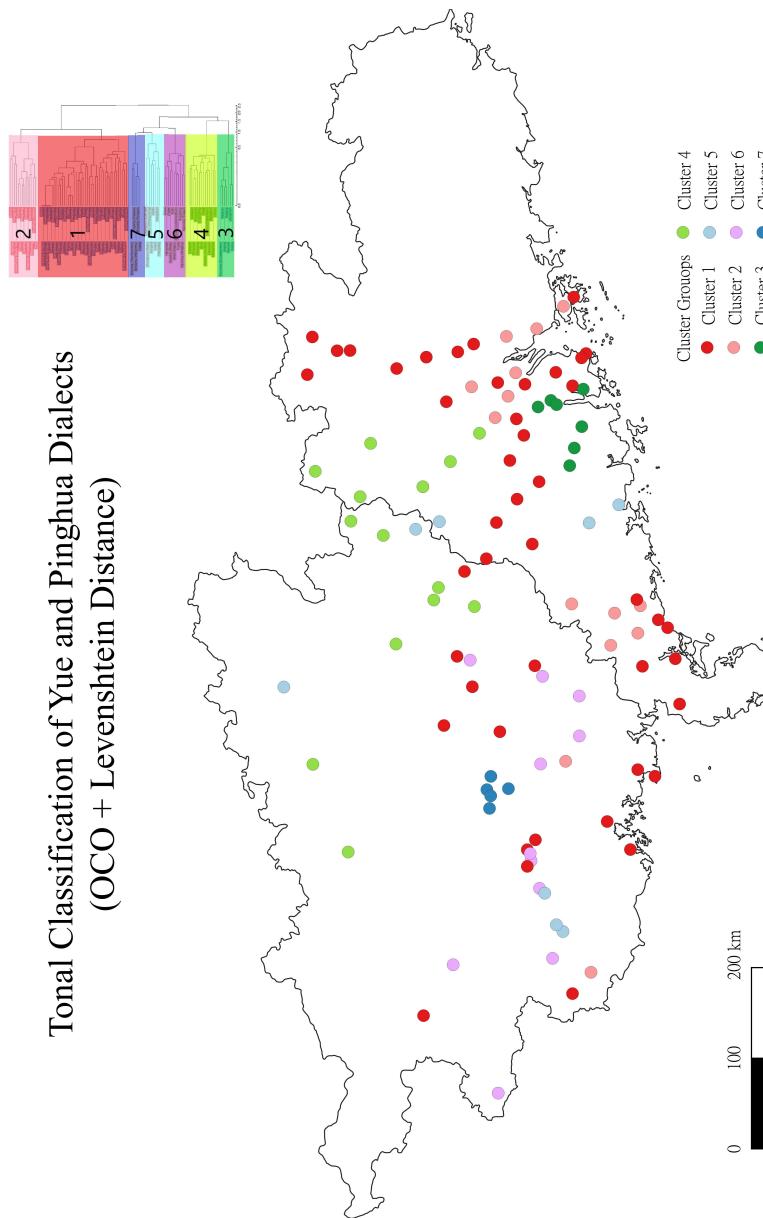


Figure 14: Cluster map (Ward's method) of the tonal classification of the Yue-Pinghua data (OCO + Levenshtein distance)

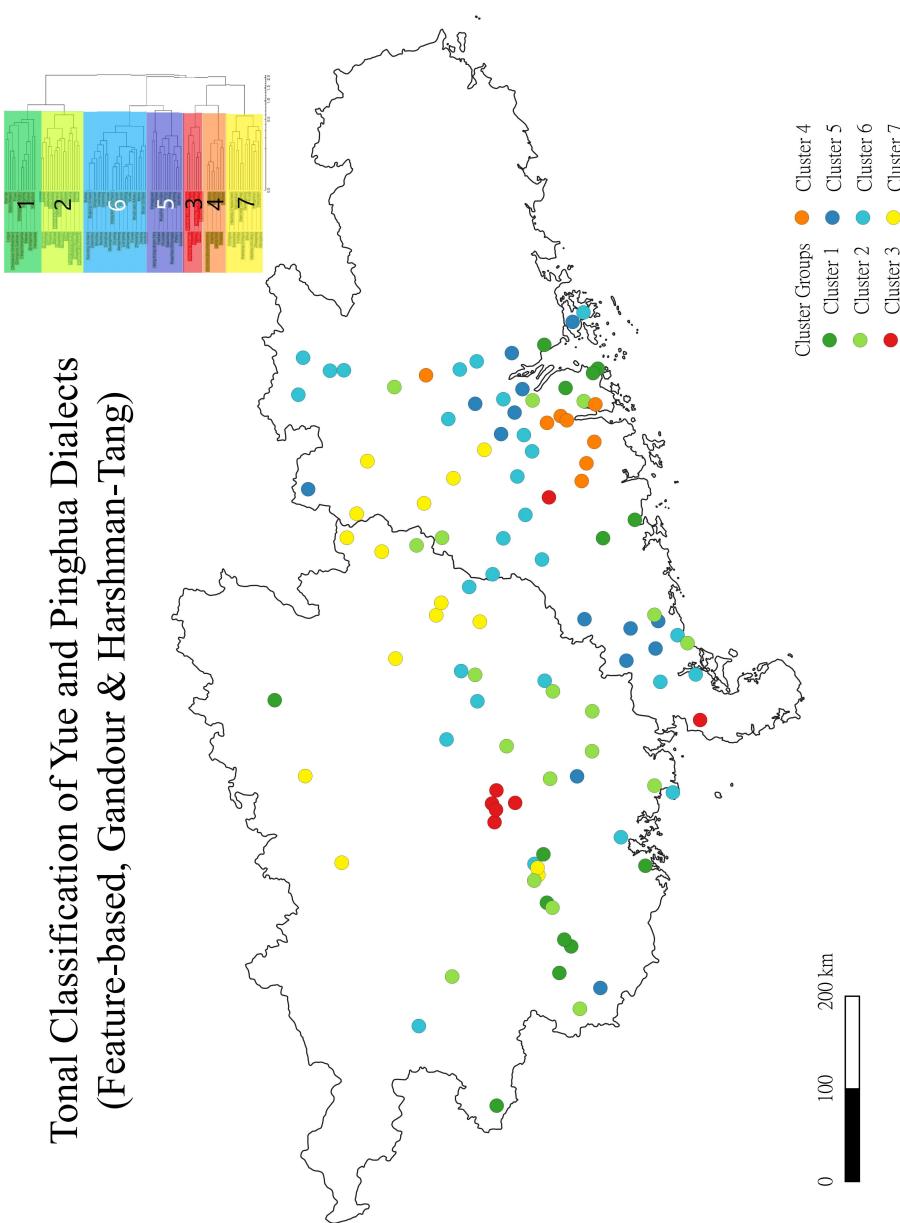


Figure 15: Cluster map (Ward's method) of the tonal classification of the Yue-Pinghua data (GH-T)

## References

- Borg, Ingwer & Patrick J. F. Groenen. 2005. *Modern multidimensional scaling: Theory and applications* (Springer series in statistics). New York: Springer. DOI: 10.1007/0-387-28981-X.
- Chambers, Jack K. & Peter Trudgill. 1998. *Dialectology*. 2nd edn. Cambridge: Cambridge University Press.
- Chao, Yuen-Ren. 1928. 現代吳語的研究 [*Studies in the modern Wu-dialects*]. Beijing: 清華學校研究院 [Tsinghua School Research Institute].
- Chao, Yuen-Ren. 1930. A system of tone letters. *Le maître phonétique* 8(45). 24–27. <http://www.jstor.org/stable/44704341>.
- Chen, Hailun & Yi Lin. 2009. 粵語平話土話方音字彙第1編: 廣西粵語、桂南平話部分 [*The lexicon of Yue, Pinghua and Tuhua : Guangxi Yue and Guinan Pinghua*], vol. 1. Shanghai: 上海教育出版社 [Shanghai Education Publishing].
- Chen, Hailun & Cunhan Liu. 2009. 粵語平話土話方音字彙第2編: 桂北、桂東及周邊平話、土話部分 [*The lexicon of Yue, Pinghua and Tuhua : Pinghua and Tuhua in Guibei, Guidong and the surrounding areas*], vol. 2. Shanghai: 上海教育出版社 [Shanghai Education Publishing].
- Chen, Xiaojan. 2009. 廣西賀州八步(桂嶺)本地話音系 [The phonology of Hezhou Babu (guiling) local vernacular in Guangxi]. 方言 [*Dialect*] 1. 53–71.
- Chen, Xiaojin & Zewen Weng. 2010. 粵語西翼考察 [*An investigation in the west of the Yue-speaking area*]: 广西贵港粤语之个案研究 [*A case-study on Guigang Yue in Guangxi*]. Guangzhou: 暨南大學出版社 [Jinan University Press].
- Cheng, Chin-Chuan. 1973. 漢語聲調計量研究 [A quantitative study of Chinese tones]. 中國語言學報 [*Journal of Chinese Linguistics*] 1(1). 93–110.
- Cheng, Chin-Chuan. 1991. 漢語方言親疏計量關係 [Quantifying affinity among Chinese dialects]. 中國語言學報專題系列 [*Journal of Chinese Linguistics Monograph Series*] 3. 76–110.
- Cheng, Chin-Chuan. 1997. Measuring relationship among dialects: DOC and related resources. *International Journal of Computational Linguistics and Chinese Language Processing* 2(1). 41–72.
- Chinese Academy of Social Sciences. 2012. 中國語言地圖集 [*Language atlas of China*]. 2nd edn. Beijing: Commercial Press.
- de Sousa, Hilário. 2022. On Pinghua and Yue: Some historical and linguistic perspectives. *Crossroads* 19(2). 257–295. DOI: 10.1163/26662523-bja10004.
- Do, Youngah & Ryan Ka Yau Lai. 2021. Accounting for lexical tones when modeling phonological distance. *Language* 97(1). 39–67. DOI: 10.1353/lan.2021.0012.
- Francis, Winthrop Nelson. 1983. *Dialectology: An introduction*. London: Longman.

- Gandour, Jackson. 1983. Tone perception in far Eastern languages. *Journal of Phonetics* 11(2). 149–175. DOI: 10.1016/S0095-4470(19)30813-7.
- Gandour, Jackson & Richard A Harshman. 1978. Crosslanguage differences in tone perception: A multidimensional scaling investigation. *Language and speech* 21(1). 1–33. DOI: 10.1177/002383097802100101.
- Goebel, Hans. 1984. *Dialektometrische Studien: Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF* (Beihefte zur Zeitschrift für romanische Philologie 191–193). Berlin, Boston: Max Niemeyer Verlag.
- Gooskens, Charlotte & Wilbert Heeringa. 2004. Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language Variation and Change* 16. 189–207.
- Grootaers, Willem A. 2003. 漢語方言地理學 [*Contributions to Chinese dialect geography*]: Translated by R-J. Shi [石汝杰] and R. Iwata [岩田礼]. Shanghai: Shanghai Jiaoyu Press.
- Handel, Zev. 2015. Non-IPA Symbols in IPA transcriptions in China. In Rint Sybesma (ed.), *Encyclopedia of Chinese language and linguistics online*. Bedfordshire, UK: Brill. DOI: 10.1163/2210-7363\_ecll\_COM\_00000292.
- Heeringa, Wilbert. 2004. *Measuring dialect pronunciation differences using Levenshtein distance*. Groningen: University of Groningen. (Doctoral dissertation).
- Heeringa, Wilbert, Vincent van Heuven & Han Van de Velde. 2022. *Levenshtein edit distance app*. [www.led-a.org](http://www.led-a.org).
- Huang, Qun. 2006. 賀州市賀街本地話同音字匯 [The homonymic syllabary of the Hezhou local vernacular in Hezhou]. 桂林師範高等專科學校學報 [*Journal of Guilin Normal College*] 20(3). 5–13.
- Hubert, Lawrence & Phipps Arabie. 1985. Comparing partitions. *Journal of Classification* 2(1). 193–218. DOI: 10.1007/BF01908075.
- Hyman, Larry M. 2006. Word-prosodic typology. *Phonology* 23(2). 225–257. DOI: 10.1017/S0952675706000893.
- Leinonen, Therese, Çağrı Çöltekin & John Nerbonne. 2016. Using gabmap. *Lingua* 178. 71–83. DOI: 10.1016/j.lingua.2015.02.004.
- Li, Lianjin. 2000. 平話音韻研究 (*Research on Pinghua phonology*). Nanning: Guangxi Renmin Press.
- Liang, Jinrong. 1997. 桂北平話語音研究 [*Research on Guibei Pinghua phonology*]. Guangzhou: Jinan University. (Doctoral dissertation).
- Liang, Min & Junru Zhang. 1999. 廣西平話概論 [An introduction to Pinghua in Guangxi]. 方言 [*Dialect*] 1. 24–32.

- Liu, Caifeng. 2015. 廣東兩陽粵語語音研究 [*A phonological study on Cantonese in Liangyang area of Guangdong*]. Guangzhou: Jinan University. (Doctoral dissertation).
- Nerbonne, John, Rinke Colen, Charlotte Gooskens, Peter Kleiweg & Therese Leinonen. 2011. Gabmap: A web application for dialectology. *Dialectologia Special Issue II*. 65–89.
- Nerbonne, John & Peter Kleiweg. 2007. Toward a dialectological yardstick. *Journal of Quantitative Linguistics* 14(2–3). 148–166. DOI: 10.1080/09296170701379260.
- Rand, William M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association* 66(336). 846–850. DOI: 10.2307/2284239.
- Sammon, John W. 1969. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers* C-18(5). 401–409. DOI: 10.1109/T-C.1969.222678.
- Séguy, Jean. 1971. La relation entre la distance spatiale et la distance lexicale. *Revue de linguistique romane* 35(139–140). 335–357. DOI: 10.5169/seals-399508.
- Séguy, Jean. 1973. La dialectométrie dans l'atlas linguistique de la Gascogne. *Revue de linguistique romane* 37(145–146). 1–24. DOI: 10.5169/seals-658403.
- Shao, Huijun. 2016. 粵西湛茂地區粵語語音研究 [*The phonological study of the Yue dialects spoken in the Zhan-Mao area in Western Guangdong*]. Guangzhou: 中山大学出版社 [Sun Yat-Sen University Press].
- Shi, Rihai. 2009. 廣西防城區粵語音系 [The Phonology of the Fangcheng Yue dialect in Guangxi]. 百色學院學報 [*Journal of Baise University*] 22(2). 106–116.
- Stanford, James N. 2012. One size fits all? Dialectometry in a small clan-based indigenous society. *Language Variation and Change* 24(2). 247–278. DOI: 10.1017/S0954394512000087.
- Sung, Ho Wang Matthew. 2023. *Is a typologically, genetically different language similar to European languages? A dialectometrical analysis on Yue and Pinghua*. Conference Presentation at 73. Studentische Tagung Sprachwissenschaft (StuTS), Frankfurt, Germany, 25–29 May 2023. <https://73.stuts.de/>.
- Sung, Ho Wang Matthew, Jelena Prokic & Yiya Chen. 2024. A new dataset for tonal and segmental dialectometry from the Yue- and pinghua-speaking area. In Michael Hahn, Alexey Sorokin, Ritesh Kumar, Andreas Shcherbakov, Yulia Otmakhova, Jinrui Yang, Oleg Serikov, Priya Rani, Edoardo M. Ponti, Saliha Muradoglu, Rena Gao, Ryan Cotterell & Ekaterina Vylomova (eds.), *Proceedings of the 6th workshop on research in computational linguistic typology and multilingual NLP*, 25–36. St. Julian's, Malta: Association for Computational Linguistics. <https://aclanthology.org/2024.sigtyp-1.3>.

- Sung, Ho Wang Matthew & Jelena Prokić. 2024. Exploring tonal variation using dialect tonometry. *Languages* 9(12). 378.
- Tan, Yuanxiong. 2000. 桂南平話研究 [Research on Guinan Pinghua]. Guangzhou: Jinan University. (Dissertation).
- Tan, Yuanxiong. 2012. 平話和土話 [Pinghua and Tuhua]. In Rong Li, Stephen Adolphe Wurm, Theo Baumann & Mei W. Lee (eds.), 中國語言地圖集 [Chinese language atlas], 152–158. Beijing: 商務印書館 [The Commercial Press].
- Tan, Yuanxiong. 2017. 廣西賓陽縣(賓州鎮)本地話音系 [The phonology of Binzhouzhen in the Binyang county in Guangxi]. 梧州學院學報 [Journal of Wuzhou University] 27(5). 58–71.
- Tang, Chaoju. 2009. *Mutual intelligibility of Chinese dialects: An experimental approach*. Leiden: Leiden University. (Dissertation).
- Tobler, Waldo. 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography* 46 (Supplement). 234–240. DOI: 10.2307/143141.
- Wei, S. 1996. 試論平話在漢語方言中的地位 [The status of Pinghua within Chinese dialects]. *語言研究 [Language research]* 2. 95–101.
- Wells, John C. 1982. *Accents of English: An introduction*, vol. 1. Cambridge: Cambridge University Press.
- Wichmann, Søren & Qibin Ran. 2019. ASJP 模式的漢語方言計算分析——以65個漢語方言語檔為例 [A phylogenetic study on 65 Chinese doculects: With ASJP tools]. *现代语文 [Modern Chinese]* 5. 4–13.
- Wieling, Martijn, Eliza Margaretha & John Nerbonne. 2011. Inducing phonetic distances from dialect variation. *Computational Linguistics in the Netherlands Journal* 1. 109–118. <https://clijournal.org/clinj/article/view/10>.
- Wu, Wei. 2001. 論桂南平話的粵語系屬 [On the Yue affiliation of Guinan Pinghua]. 方言 [Dialect] 2. 133–141.
- Wu, Wei. 2012. 粵語 [Yue]. In *Language Atlas of China*, 2nd edn.
- Xie, Jianyou. 2007. 廣西漢語方言研究 [Studies on the Chinese dialects in Guangxi]. Nanning: 廣西人民出版社 [Guangxi People's Publishing House].
- Yang, Cathryn & Andy Castro. 2008. Representing tone in Levenshtein distance. *International Journal of Humanities and Arts Computing* 2(1–2). 205–219. DOI: 10.3366/E1753854809000391.
- Yang, Shiwen. 2013. 廣西藤縣濤江方言音系 [The phonology of the Tengxian Mengjiang dialect in Guangxi]. 方言 [Dialect] 1. 71–85.
- Yip, Moira. 2002. *Tone* (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press. DOI: 10.1017/CBO9781139164559.
- You, R. 2016. 漢語方言學教程 [A course in Chinese dialectology]. Shanghai Educational Publishing House.

- Yuan, Jiahua. 2001. 漢語方言概要 [*An introduction to Chinese dialects*]. 2nd edn. Beijing: Wenzi Gaige Press.
- Yue-Hashimoto, Anne. 1988. A preliminary investigation into the subclassification problem of the Yue dialects. *Computational Analysis of Asian and African Languages* 30. 7–38.
- Zhan, Bohui. 1981. 現代漢語方言 [*Modern Chinese dialects*]. Wuhan: 湖北人民出版社 [Hubei People's Press].
- Zhan, Bohui (ed.). 2002. 廣東粵方言概要 [*Introduction to the Yue dialects in Guangdong*]. Guangzhou: 暨南大學出版社 [Jinan University Press].
- Zhan, Bohui & Yat-Shing Cheung. 1987. *A survey of dialects in the Pearl river delta*, vol. 1: Comparative morpheme-syllabary. Guangzhou: 廣東人民出版社 [Guangdong People's Publishing House].
- Zhan, Bohui & Yat-Shing Cheung. 1990. *A survey of dialects in the Pearl River Delta*, vol. 3: A synthetic review. Guangzhou: People's Publishing House of Guangdong.
- Zhan, Bohui & Yat-Shing Cheung. 1994. 粵北十縣粵方言調查報告 [*A survey of Yue dialects in North Guangdong*]. Guangzhou: 暨南大學出版社 [Jinan University Press].
- Zhan, Bohui & Yat-Shing Cheung. 1998. *A survey of Yue dialects in West Guangdong*. Guangzhou: Jinan University Press.
- Zhong, Ziqiang. 2015. 廣西蒼梧本地話音系 [The phonology of Cangwu Local Vernacular in Guangxi]. 方言 [*Dialect*] 2. 177–192.

# Chapter 16

## Convergence and divergence of tone paradigms across Tai dialects in the 21st century

Chingduang Yurayong<sup>a,b</sup>, Saknarin Pimvunkum<sup>b</sup> & Yuttaporn Naksuk<sup>b</sup>

<sup>a</sup>University of Helsinki <sup>b</sup>Mahidol University

The present study examines tone paradigms of Tai dialects spoken in Laos, Malaysia, Myanmar, Thailand, and Vietnam. The focus is on the homophones between the paradigmatic tone D in checked syllables and tones A, B, and C in smooth syllables. The method comprises a conventional historical-comparative approach to dialectal classification and language contact among the Tai languages, complemented by a computer-aided quantitative approach which makes treatment of data collected from 315 different dialect speakers of 17 Tai languages possible. The data illustration generated through a Neighbor-Net algorithm identifies shift of the tone paradigm in some diaspora dialects which have converged towards models from sociolinguistically dominant dialects in their speaking locations.

### 1 Introduction

The Tai languages form a second level subbranch under the Tai-Kadai language family, one of the indigenous language groups in Mainland Southeast Asia. According to the mainstream view, the Proto-Tai-Kadai language is believed to have been spoken ca. 5000 years ago in South China (Ostapirat 2005: 126) before its dispersal into several first level subbranches: 1) Kra, 2) Hlai, and 3) Kam-Tai. The taxonomical structure, diversification of the Tai-Kadai language family and estimated dates of dispersals at different stages of intermediate protolanguages are given in Figure 1.

Chingduang Yurayong, Saknarin Pimvunkum & Yuttaporn Naksuk. 2025. Convergence and divergence of tone paradigms across Tai dialects in the 21st century. In Susanne Wagner & Ulrike Stange-Hundsdörfer (eds.), *(Dia)lects in the 21st century: Selected papers from Methods in Dialectology XVII*, 403–434. Berlin: Language Science Press. DOI: 10.5281/zenodo.15006623 



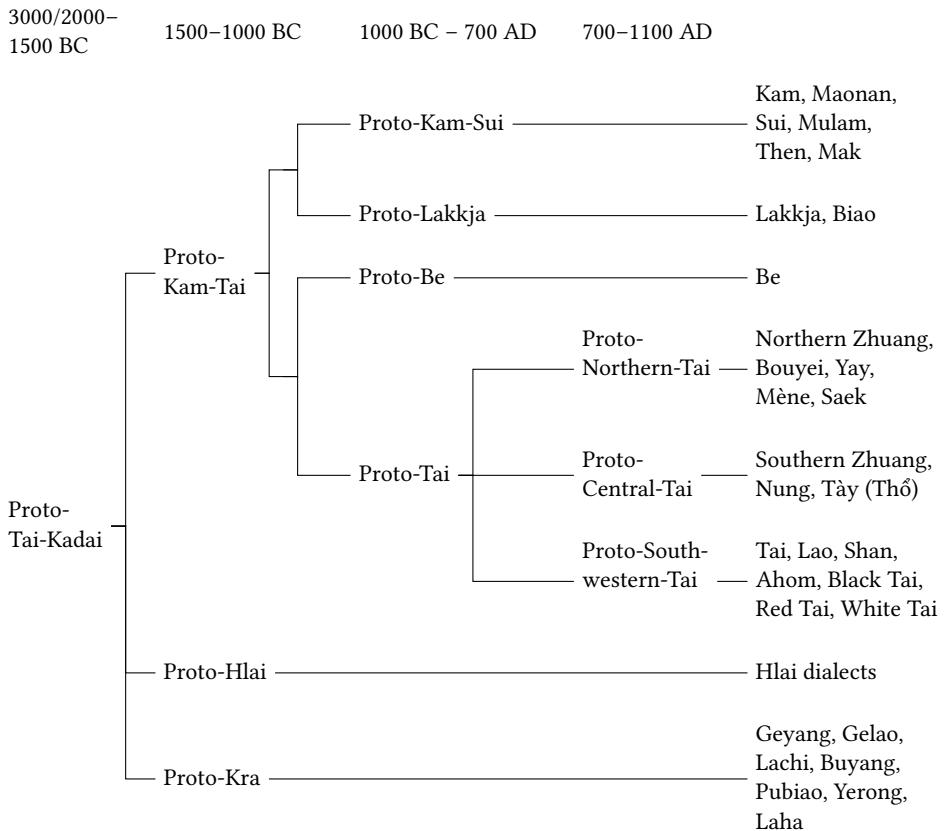


Figure 1: Diversification of Tai-Kadai languages (based on Mitani 1977, Edmondson & Solnit 1997, Pittayaporn 2014).

As is indicated in Figure 1, the Tai branch stems off from the Kam-Tai intermediate protolanguage, which makes the Kam-Sui languages linguistically closer to the Tai languages, compared to the more distantly related Kra and Hlai languages. This close genealogical tie is evident from a higher proportion of shared lexical items between Kam-Sui and Tai (see Luo 1988, Edmondson & Solnit 1997: 4).

In terms of geographical distribution, the Tai branch and its linguistic populations have spread out from their hypothetical Urheimat in South China towards South Asia and Mainland Southeast Asia. Figure 2 illustrates the speaking areas of modern Tai-Kadai languages. Today, there is a total of ca. 102 million Tai-Kadai speaking people (Liao 2023a: 199), of which ca. 21 million are living in China. As can be seen from Figure 2, the Tai branch has extended its speaking area to the

south and west, significantly farther than its sister languages, Kam-Sui, Kra, and Hlai.

Focusing on the Tai branch, previous studies have proposed a taxonomical structure with the following three major groups distributed across Mainland Southeast Asia (Chamberlain 1975, Li 1977; see also Luo 1997 for an alternative classification with four groups):

1. Northern Tai – the majority spoken in China, and the Saek language spoken in Laos, Thailand, and Vietnam
2. Central Tai – spoken in China and Vietnam
3. Southwestern Tai – spoken in China, India, Laos, Malaysia, Myanmar, Thailand, and Vietnam.

In any case, such a subgrouping remains disputable when certain isoglosses concerning sound changes occur across subbranches. By only applying a conventional comparative method of historical linguistics, common patterns and clustering of phonological innovations namely do not give a straightforward dialectal classification. This is likely due to later language contact among Tai dialects, resulting from migration and relocation of the linguistic populations and causing convergence among Tai dialects spoken in adjacent areas (Pittayaporn 2009, Liao 2023b). A method which carefully deals with potential contact-induced changes and incorporates them into the classification yields a result illustrated in Figure 3.

For instance, innovation Q concerns a sound change involving the merger  $*kr-$  =  $*h_r-$ , which is the case in Shan, Thai (Siamese), Black Tai and Lue, but not in Sapa (Pittayaporn 2009: 301).

In terms of population history, it has been the case especially in Laos and Thailand where people were compelled to migrate to new locations in the region, often as captives during the multiple war time periods between 1750 and 1850 (Piyabhan 1998). The series of such historical events has given rise to new diaspora communities dispersing around Thailand in particular. The resettlement of Tai dialect speakers primarily concerned community members from the first generation, whereas the younger second and third generations of dialect speakers already exhibit signs of shifting towards a national language or the dominant regional dialect of their current location (Akharawatthanakun 2003, Bunyasathit et al. 2016).

For many decades, the evolution and convergence among Tai dialects have been popular research topics in Thai and Tai-Kadai linguistics among researchers and students, particularly in Thai universities. Among hundreds of individual

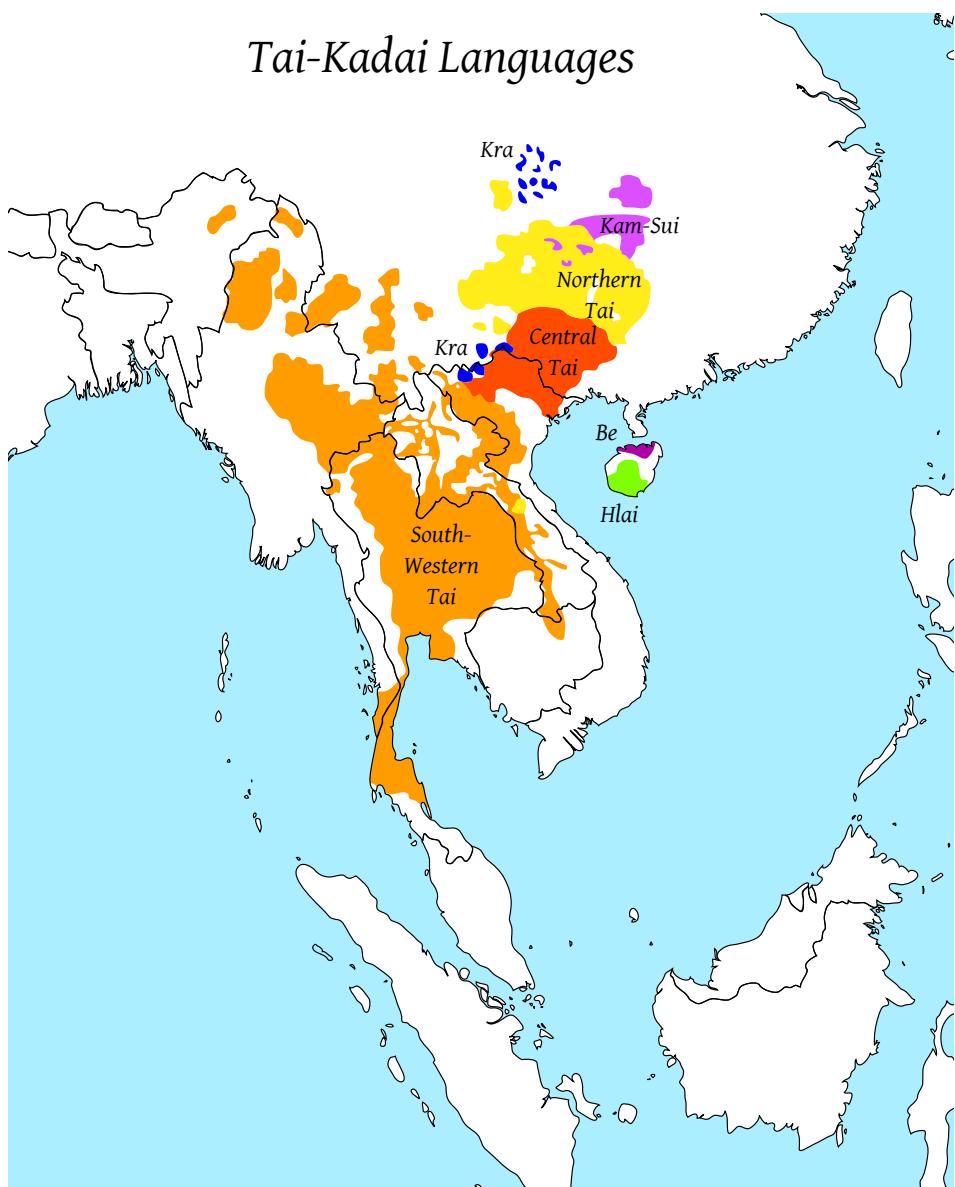


Figure 2: Geographical distribution of modern Tai-Kadai languages (public domain, <https://upload.wikimedia.org/wikipedia/commons/7/71/Taikadai-en.svg>).

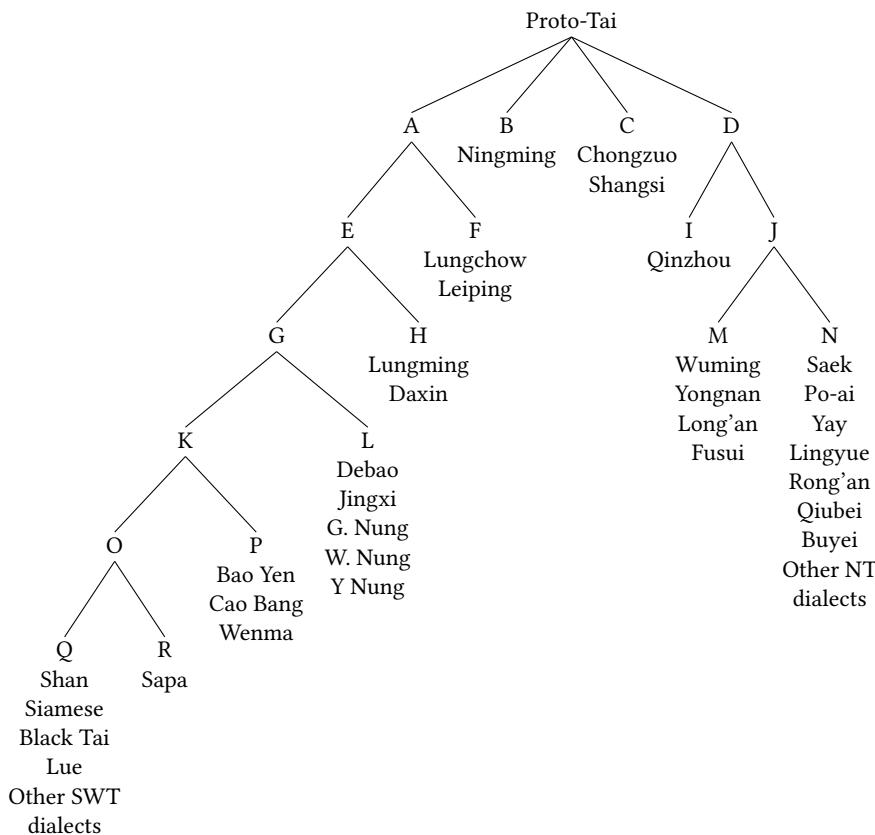


Figure 3: A classification of Tai dialects with reference to contact-induced changes (Pittayaporn 2009: 298).

studies and theses published after the comprehensive handbook of Tai dialects by Li (1977), extensive works on Tai dialects by Gedney (edited by Hudak 2008) and a comparative work on Tai tone paradigms by Brown (1985), there have been several large-scale comparative studies of tone paradigms. To name just few, the comparative approach has been applied for the investigation of tonal variation in Yo dialects (Koowatthanasiri 1981), Phu Thai dialects (Srithonrat 1983), Lao dialects (Akharawatthanakun 2003), Thai dialects in Malaysia (Damanhuri 2004), Central Thai dialects (Canilao 2010), Black Tai dialects (Buruspahat 2012), and Tai Lue dialects (Akharawatthanakun 2020). Consequently, the consensus is continuously evolving due to newly collected data from individual dialects which can contradict and deviate from earlier proposals in the genealogical classification of the Tai languages.

Participating in the ongoing discussion in dialectological studies of Tai-Kadai languages, our first goal in the present study is to identify change in progress. We choose homophones in tone paradigms of different Tai dialects as our object of study. The patterns of these homophones may have either resisted changes or diverged from their protosystems as we have arrived in the 21st century. Concretely, our attention is directed towards (i) paradigmatic tones DL and DS, which represent checked syllables with final stops [-p/-t/-k/-ʔ], and (ii) their alignment with tones A, B and C, which represent smooth syllables with final vowels or sonorants (see Section 2 for the detailed definition of tone paradigm).

The nature of the current study involves historical-comparative linguistics in conjunction with a quantitative approach, utilising computer-aided tools to investigate data from more than 300 different dialect speakers under investigation and aiming to trace significant signals of convergence among their tone paradigms. With this method, our second goal is to demonstrate how to address language changes occurring in the 21st century and their potential impact on how we read and interpret the taxonomical structure of the Tai languages.

We divide our study into five sections. After introducing the research questions, hypotheses, and historical background in Section 1, in Section 2 we present a diachronic framework for the studies of tones, which has been previously employed for Tai as well as other Southeast Asian languages. Building upon the state of the art, in Section 3 we provide a description of how to apply a computer-aided quantitative method to investigate and analyse tone paradigm data from different Tai dialect speakers. Based on the data analysis, in Section 4 we discuss the scenarios of convergence and divergence of tone paradigms in general and in individual cases, which can be explained by sociolinguistic factors. Finally, in Section 5 we give some conclusional thoughts on the functionality of the applied quantitative method and highlight some areas which deserve attention in future studies and data collection, in order to conduct diachronic studies in an even more efficient manner.

## **2 Tone system in Tai and other Southeast Asian languages**

Tones in Tai as well as in other languages in Southeast Asia, such as the cognate Kam-Sui, Hlai and Kra languages, along with neighbouring Chinese, Karenic, Hmong-Mien, and Vietic languages, are organised within a paradigm. A general principle dictates that the distribution of tones in paradigm is controlled by two factors: (i) the type of initial consonant, and (ii) the type of syllable. Both factors offer clues which can be traced back to phonological shapes at the protolanguage

stage. Therefore, the tone paradigm serves as an important piece of evidence for reconstructing the lexicon of the protolanguage because sound changes, such as merger or loss of specific consonant types, may have left their traces on tones. This mechanism is called tonogenesis, which gave rise to tones as compensation for loss of certain phonetic features, most notably a voice distinction in the domain of consonants (see Wulff 1934, Haudricourt 1954, Ferlus 2004, Kingston 2011, Michaud & Sands 2020). Naturally, this approach implies that the protolanguages did not originally possess tones, as has been discussed, most notably, for Chinese (Handel 2014: 593, Hill 2019: 182).

In the case of Tai languages, a classification of four types of initial consonants has been associated with different pitch patterns as given below (see e.g. Maddieson 1984 and Ratliff 2015 for the effect of voicing in pitch). One of the common practices to annotate pitches and contours in the studies of East Asian linguistics uses numbers to indicate pitch height: 1 = lowest pitch vs. 5 = highest pitch.

1. Aspirated – high pitch
2. Voiceless – mid pitch
3. Implosive – mid pitch
4. Voiced – low pitch

Note that the relation between manner of articulation and pitch presented here should not be taken as an absolute linguistic universal. Some Tai-Kadai languages, namely, exhibit no significant correlation between aspiration and high pitch (see Zhang 1980: 38, Edmondson 1990: 188, Liao 2016: 170–171, Zhu et al. 2016: 18). Furthermore, the status of the aspirated initials remains a subject of dispute, with questions surrounding whether they can be reconstructed as far as to Proto-Tai or only to the Proto-Southwestern-Tai intermediate stage (see arguments in Liang & Zhang 1996, Pittayaporn 2009, with a summary provided in Liao 2023a). Another dimension to consider is the syllable type, which leads to a division of four tone classes, a.k.a. tonal splits: A, B, C (smooth syllables ending with vowels or sonorants), and DL, DS (checked syllables ending with stops [-p/-t/-k/-?]).

The two factors discussed above serve as the foundation for the tool known as “Gedney’s tone box” (see Table 1), which was initially proposed by William J. Gedney (1972) and is recognised as one of the main comparative tools in the Tai-Kadai linguistics scholarship. Based on the original tone box by Gedney, we have systematised and relabelled the tone slots for the comparative purpose of

the present study (see Table 2). According to Pittayaporn (2009: 271), contour characteristics of each tone are reconstructed for Proto-Tai as follows: A = mid-level, B = low-rising, C = high-falling, and D = low-rising.

Table 1: The original Gedney's (1972) Tai tone box for Proto-Tai

		Proto-Tai tones					
		A	B	C	D-short	D-long	
Initials at time of tonal splits	Voiceless	Friction	5th tone	2nd tone	3rd tone	2nd tone	2nd tone
			1st tone				
	Voiced			3rd tone	4th tone	4th tone	3rd tone
Smooth syllables				Checked syllables			

Table 2: The modified Tai tone box (based on Gedney 1972).

Initial consonant class at time of tonal splits	Initial consonant class				
	A	B	C	DL(ong)	DS(hort)
Aspirated	A1	B1	C1	DL1	DS1
Plain	A2	B2	C2	DL2	DS2
Implosive	A3	B3	C3	DL3	DS3
Voiced	A4	B4	C4	DL4	DS4

In any case, it is important to note that Gedney's tone box has been primarily developed within the context of the Southwestern branch of Tai languages. Consequently, it often has issues when applied to explain the tone paradigms in other subbranches, i.e. Central and Northern Tai, not to mention the more distantly related Tai-Kadai languages like Kam-Sui, Kra and Hlai (see the criticisms on cross-family validity of Gedney's tone box tool in Liao 2023a).

A similar type of tone box is also employed in the studies of tones in Chinese, Karenic, Hmongic and Vietic languages. However, some languages may distinguish only a contrast between voiceless and voiced initial stops, while the aspirated and implosive classes are grouped within the same category as voiceless

stops, making only the voice distinction between the initial consonant classes. Similarly, the vowel length distinction is not considered a meaningful category in some languages. See, for instance, a tone box with only two initial consonant classes, based on the reconstructed Middle Chinese tone paradigm and used to describe tone paradigms across modern Chinese dialects in Table 3. Interestingly, certain Middle Chinese tone names appear to record the contour characteristics pronounced during the Middle Chinese period (cf. the reconstruction of Proto-Tai tone contours in Pittayaporn 2009: 271).

Table 3: The Chinese tone box

Middle Chinese name of tone				
	平 <i>píng</i> 'level'	上 <i>shǎng</i> 'rising'	去 <i>qù</i> 'falling'	入 <i>rù</i> 'entering'
Tone label	A	B	C	D
Voiceless initials	A1	B1	C1	D1
Voiced initials	A2	B2	C2	D2

From the paradigms in modern Tai languages, we know that tone D can be traced back to a checked syllable structure ending with stops [-p/-t/-k] in Proto-Tai. However, there is unfortunately no internal evidence available for syllable structures of tones A, B and C prior to the emergence of tones in Proto-Tai-Kadai. This is in contrast to some other neighbouring languages, for example, Chinese where sufficient evidence for syllable structures during the pre-tonal Old Chinese stage is attested in earlier written sources and phonological adaptation of Sanskrit loanwords in Old Chinese, as well as Old Chinese loanwords in Korean (Handel 2014: 594, Hill 2019: 184–185). A reconstruction of the Old Chinese syllable structures is given as follows:

1. A – a syllable ending with a vowel or sonorant
2. B – a syllable ending with a glottal stop [-?]
3. C – a syllable ending with a sibilant [-s]
4. D – a syllable ending with a stop [-p/-t/-k]

Similar patterns and mechanisms have also been applied together with evidence from Chinese loanwords to explain tonogenesis in Hmong-Mien (Ratliff

2010: 183–184) and Vietic (Thurgood 2002), both of which have undergone early contact with Chinese. Interestingly, in the field of Tai-Kadai linguistics, there has been no serious attempt to reconstruct distinct original syllable types for tones A, B and C in Proto-Tai-Kadai and their corresponding reflexes in subsequent intermediate protolanguages, although there are ample early Chinese loanwords available for similar analysis (cf. Pittayaporn 2014), as has been done for Hmong-Mien and Vietic.

Relevant to the current study which focuses on Tai tone paradigms is the observation that tones DL and DS always align paradigmatically with some of the twelve phonemic tone slots in columns A, B, and C (cf. the similar low-rising contour reconstructed for tones B and D as given in Pittayaporn 2009: 271). However, in practice, the maximum number of phonemic tone distinctions observed in Tai-Kadai languages is nine, often due to the merger of tones between the plain (tones A2/B2/C2/D2) and implosive (tones A3/B3/C3/D3) classes of initial consonants in most languages. Ultimately, only a few Tai-Kadai languages possess nine distinct phonemic tones in a symmetrical fashion, a characteristic primarily found in certain dialects of the Kam language within the Kam-Sui branch. Table 4 illustrates pairs of examples, where the former represents a smooth syllable (tone A, B, or C) and the latter represents a checked syllable (tone DL or DS), except the last three tones (53, 453, and 33) which only occur in smooth syllable. Despite the symmetrical case of three consonant classes combined with three syllable types in Kam, the average number of tones across the Tai-Kadai family typically falls within the range of five to seven tones (see e.g. Brown 1985, Hudak 2008).

Table 4: Tones of common Kam (modified from Yang & Edmondson 2008: 514).

Tone value (contour)	Example	Gloss	Tone value (contour)	Example	Gloss
55	<i>ma</i> <sup>55</sup>	‘vegetable’	13	<i>no</i> <sup>13</sup>	‘rat’
	<i>jak</i> <sup>55</sup>	‘wet’		<i>p<sup>h</sup>at</i> <sup>13</sup>	‘blood’
35	<i>ma</i> <sup>35</sup>	‘to come’	31	<i>ma</i> <sup>31</sup>	‘horse’
	<i>mat</i> <sup>35</sup>	‘flea’		<i>mak</i> <sup>31</sup>	‘ink’
11	<i>ma</i> <sup>11</sup>	‘tongue’	53	<i>ja</i> <sup>53</sup>	‘paddy, wet field’
	<i>sak</i> <sup>11</sup>	‘thief’		<i>ma</i> <sup>453</sup>	‘to soak (rice)’
323	<i>ma</i> <sup>323</sup>	‘cloud, soft’	33	<i>pa</i> <sup>33</sup>	‘rice husk’
	<i>mak</i> <sup>323</sup>	‘big’			

The subgrouping of individual Tai languages can be discerned through both consonant and vowel inventories, as well as their tone paradigm. Concerning the latter, the distinctive homophone patterns across tone slots exhibit characteristics specific to individual language groups. This aspect of tone paradigms has significantly received scholarly attention within the field of Tai dialectology (see e.g. Brown 1985, Dockum 2019). Below, we present several characteristics unique to individual dialects, which are not commonly observed in other dialectal groups. These examples are drawn from dialectal data in Brown (1985) and Hartmann (2008) and specifically focus on the fundamental paradigmatic tones A, B, and C. Further explanations regarding individual dialectal groups are provided below following the paradigmatic summary in Table 5.

Table 5: Some characteristics of the tone paradigms in the selected Tai dialectal groups.

Tai Lue & White Tai			Tai Yuan / Northern Thai			Lao		
A1	B1	C1	A1	B1	C1	A1	B1	C1
A2	B2	C2	A2	B2	C2	A2	B2	C2
A3	B3	C3	A3	B3	C3	A3	B3	C3
A4	B4	C4	A4	B4	C4	A4	B4	C4

Shan & Central Thai			Southern Thai		
A1	B1	C1	A1	B1	C1
A2	B2	C2	A2	B2	C2
A3	B3	C3	A3	B3	C3
A4	B4	C4	A4	B4	C4

1. Tai Lue and White Tai dialects – a symmetrical split of all the paradigmatic tones A, B and C for smooth syllables between the original unvoiced and voiced initials
  - A1, A2, A3 ≠ A4
  - B1, B2, B3 ≠ B4
  - C1, C2, C3 ≠ C4

2. Tai Yuan/Northern Thai dialects – a symmetrical split as observed in Tai Lue dialects, but with a distinction between the original aspirated/plain and implosive/voiced initials within the paradigmatic tone A
  - A1, A2 ≠ A3, A4
  - B1, B2, B3 ≠ B4
  - C1, C2, C3 ≠ C4
3. Lao dialects – a merger of all tone slots within the paradigmatic tone B
  - B1 = B2 = B3 = B4
4. Shan and Central Thai dialects – a diagonal alignment of the original aspirated/plain/implosive and voiced initials within paradigmatic tones B and C
  - C1, C2, C3 = B4
5. Southern Thai dialects – a symmetrical split of the paradigmatic tones within tone A with distinctions for the original aspirated, plain/implosive and voiced initials, as well as a symmetrical split of the paradigmatic tones A, B, and C with the original voiced initials
  - A1 ≠ A2, A3 ≠ A4
  - A4 ≠ B4 ≠ C4

Next, we shift our focus to the primary data of the present study and the computer-aided methods and processes employed therein.

### **3 Data and methods**

#### **3.1 Data coverage**

In the present study, we conduct a comprehensive investigation using 315 datapoints obtained from data of speakers of Tai dialects, primarily from regions outside China. Our focus is on analysing patterns observed within the tone paradigm. The data is collected from a range of grammatical and phonological descriptions of Tai dialects in Mainland Southeast Asia. These descriptions have been published during the late-20th and early-21st centuries. Most, if not all, of the sources provide empirical data of individual dialect speakers, presented in the form of word lists covering all the tone slots outlined in Table 2. Alignment

patterns observed among different tone slots are also often discussed in a tone box similar to that of Table 5. The majority of these sources usually have their goal in describing how the observed tone paradigms correspond to or deviate from the common pattern of the specific Tai language. Some works, most notably Akharawatthanakun (2003), further delve into sociolinguistic factors which may have significantly influenced the evolution of the tone paradigm when such data is available and feasible from speech communities (see Section 4 for a similar attempt in the present study).

Out of the 315 datapoints, the data covers dialects from a total of 17 different Tai languages as listed below.<sup>1</sup>

- |                                  |                 |
|----------------------------------|-----------------|
| 1. Tai Lue                       | 10. Saek        |
| 2. Tai Yuan/Northern Thai        | 11. Shan        |
| 3. Phu Thai                      | 12. White Tai   |
| 4. Northern Lao (Luang Phrabang) | 13. Black Tai   |
| 5. Central Lao (Vientiane)       | 14. Yo          |
| 6. Southern Lao (Champasak)      | 15. Yoy         |
| 7. Western Lao/Northeastern Thai | 16. Phuan       |
| 8. Central Thai                  | 17. Khorat Thai |
| 9. Southern Thai                 |                 |

In the 21st century, dialects of various languages are spoken in diaspora areas, particularly among Lao dialect speakers who have relocated from the dialectal core regions in Luang Phrabang, Vientiane, or Champasak to various areas of Thailand. The geographic locations of the speakers included in the dataset are cartographically illustrated in Figure 4, using ArcGIS tools. As can be seen from the map in Figure 4, a noteworthy proportion of dialect speakers are not located within the core area of their respective languages, particularly the Lao diaspora dialects due to their migration history discussed by Piyabhan (1998, mentioned in Section 1). This phenomenon raises an important issue potentially contributing to the restructuring of tone paradigms among certain dialect speakers, a matter which will be discussed further in Section 4.

<sup>1</sup>The full details and sources for each datapoint can be found in the supplementary material: [https://researchportal.helsinki.fi/files/244396698/Convergence\\_and\\_divergence\\_of\\_tone\\_paradigms\\_across\\_Tai\\_dialects\\_in\\_the\\_21st\\_century\\_supplementum\\_.xlsx](https://researchportal.helsinki.fi/files/244396698/Convergence_and_divergence_of_tone_paradigms_across_Tai_dialects_in_the_21st_century_supplementum_.xlsx).

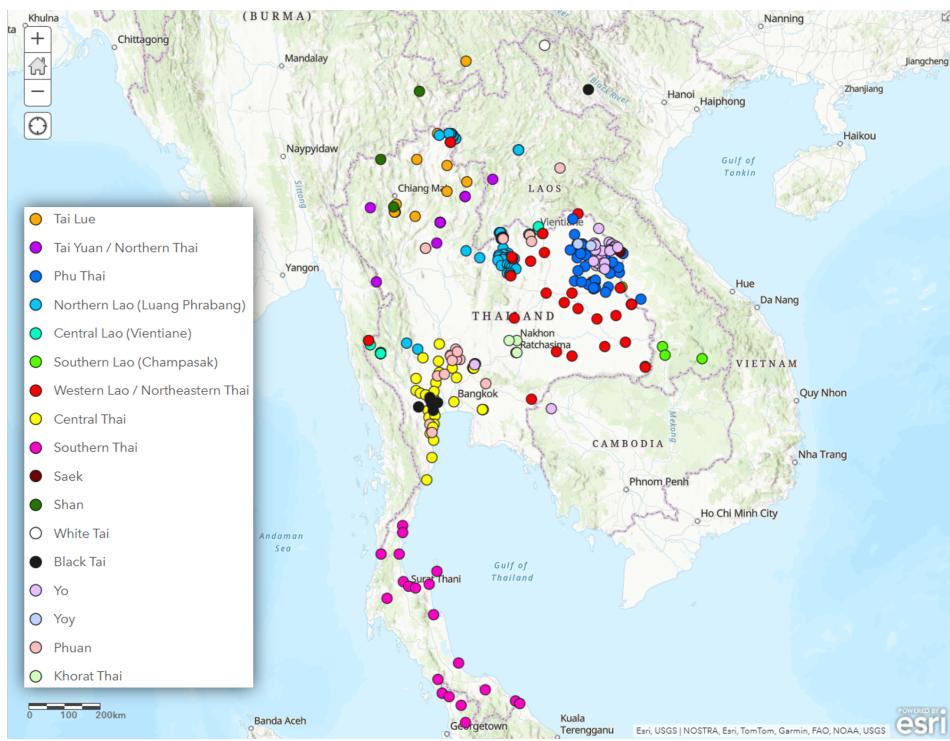


Figure 4: Datapoints of the Tai dialect speakers under investigation.

### 3.2 Methods for data illustration and analysis

As a general principle for organising the data, the information collected from various grammatical and phonological descriptions of Tai dialects is arranged based on whether tones DL and DS align with tones A, B, and/or C within individual dialects. This compiled information is then processed by a Neighbor-Net algorithm (Bryant & Moulton 2004). The outcome of this processing is a network diagram visualising the clustering and distances among typological profiles of each dialect under investigation. Importantly, this diagram is generated without exhibiting any bias towards the genealogical relationship or distances across dialects. This approach has been effectively applied to studies of several language families across Eurasia (e.g. Grünthal & Nichols 2016, Szeto et al. 2018, Nichols 2020, Yurayong & Szeto 2020, Szeto & Yurayong 2021). The data processing involves four different steps, which are elaborated upon in the following points below.

### 3.2.1 Step 1: From language description to tone paradigm

It is a common practice in Tai-Kadai linguistics for the phonological descriptions of Tai dialects to include details about tone contours (1 = lowest pitch vs. 5 = highest pitch) and paradigms presented in the form of tone box (as previously illustrated in Table 2 within Section 2). Utilising these tone boxes, the homophones between paradigmatic tones DL and DS, juxtaposed with tones A, B, and C, can be identified, as exemplified in Table 6. The tone paradigm in Table 6 reveals that the speaker of the Central Lao dialect spoken in Suan Pan, Nakhon Pathom province of Thailand, has six distinct phonemic tones. Among these, four alignment patterns of tones DL and DS are identified: (i) DL123 = C1, (ii) DL4 = C234, (iii) DS123 = A1, and (iv) DS4 = B1234. The alignment patterns, identified from the tone paradigms of individual dialect speakers, serve as the focal points of comparison in the quantitative analysis conducted in the present study.

Table 6: The tone paradigm of the Central Lao dialect speaker (LC24) in Suan Pan, Nakhon Pathom province, Thailand.

Initial consonant class at time of tonal splits	A	B	C	DL	DS	→	Alignment pattern
Aspirated	24			21			DL123 = C1
Plain		33			24		DL4 = C234
Implosive							DS123 = A1
Voiced	353		41		33		DS4 = B1234

### 3.2.2 Step 2: From tone paradigm to binary data

The homophones aligned between tones DL and DS and tones A, B, and C are then converted into binary values: 0 = no alignment vs. 1 = alignment present, as displayed in Table 7. This binary value representation characterises the tonological profile of each dialect speaker, drawing an analogy to chromosomes and their sequencing within a species. This is a concept commonly employed in evolutionary studies, in which the original development of this method finds its root (see general principles for the application of Neighbor-Net diagrams in evolutionary studies in Maddison et al. 1997, Huson & Bryant 2006). Subsequently, this dataset of tonological profiles will serve as a foundational component for data illustration model generated by the Neighbor-Net algorithm in Step 4 described in Section 3.2.4.

Table 7: The tonological profile of the Central Lao dialect speaker (LC24) in Suan Pan, Nakhon Pathom province, Thailand.

DL1												DL2											
A1	A2	A3	A4	B1	B2	B3	B4	C1	C2	C3	C4	A1	A2	A3	A4	B1	B2	B3	B4	C1	C2	C3	C4
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
DL3												DL4											
A1	A2	A3	A4	B1	B2	B3	B4	C1	C2	C3	C4	A1	A2	A3	A4	B1	B2	B3	B4	C1	C2	C3	C4
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
DS1												DS2											
A1	A2	A3	A4	B1	B2	B3	B4	C1	C2	C3	C4	A1	A2	A3	A4	B1	B2	B3	B4	C1	C2	C3	C4
1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
DS3												DS4											
A1	A2	A3	A4	B1	B2	B3	B4	C1	C2	C3	C4	A1	A2	A3	A4	B1	B2	B3	B4	C1	C2	C3	C4
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1

### 3.2.3 Step 3: From binary data to nexus format

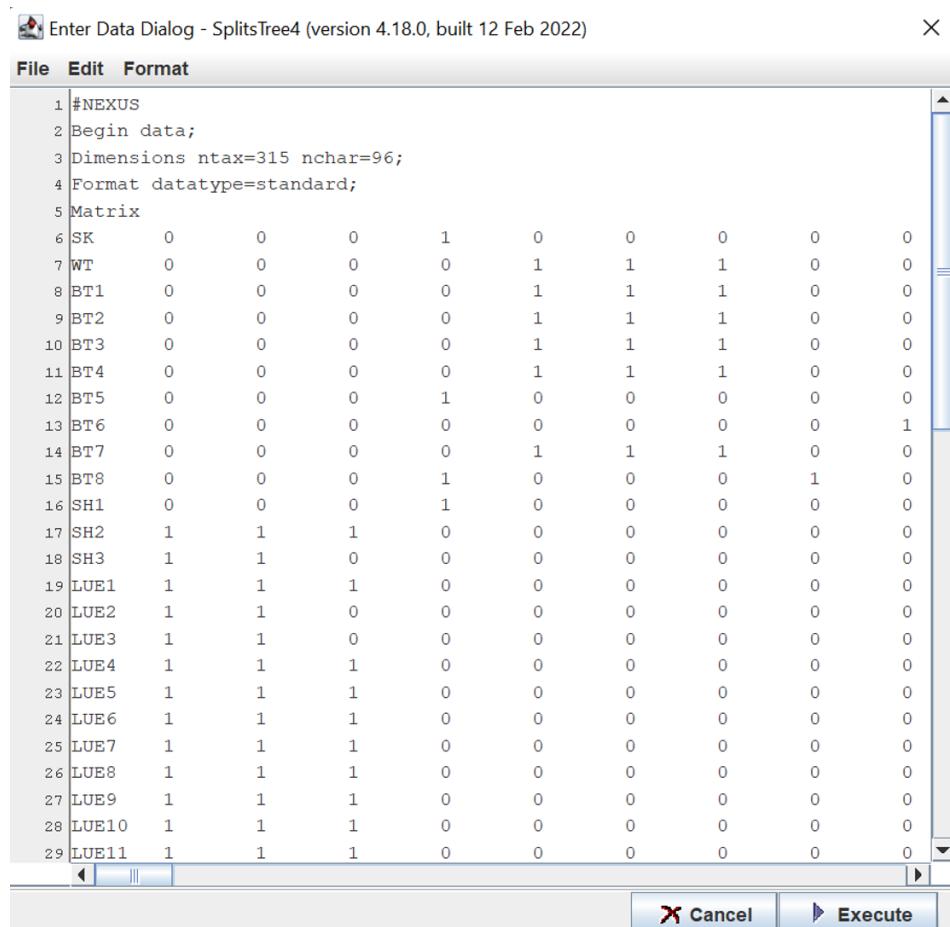
The binary data is converted into the nexus format (Maddison et al. 1997) which is a kind of structured and systematic data format for facilitating computational modelling tasks, such as the implementation of the Neighbor-Net algorithm, as demonstrated in Figure 5. The instructions for conducting the computational modelling process indicate several key parameters: (i) the total number of data-points (`ntax=315`) from 315 dialect speakers included, (ii) the total number of tone slots (`nchar=96`) from 12 tone slots within tones A, B and C  $\times$  8 tone slots within tones DL and DS, and (iii) a standard type of data analysis (`datatype=standard`) which is predicated on binary values (0 = no alignment vs. 1 = alignment present). These parameters collectively guide the computational modelling process, enabling the extraction of meaningful insights from the dataset.

### 3.2.4 Step 4: From nexus format to Neighbor-Net diagram

The final step of the data illustration process involves feeding the organised data, previously formatted according to the nexus format, into the software `Splitstree4` (version 4.18.0, built on 12 February 2022). This software operates using the Neighbor-Net algorithm, and by executing the imported data (as exemplified in Figure 5), the software performs calculations and generates visual representations of the distances across tonological profiles of each dialect. The calculations

also generate clustering of datapoints with similar profiles beneath the same nodes creating a clear illustration of the relationships, as illustrated in Figure 6.

At this point, the computer-aided tools have fulfilled their role on the quantitative front, presenting the data in a visually interpretable fashion. The subsequent work phase involves a qualitative analysis based on the Tai-Kadai linguistics scholarship. This qualitative account seeks to identify significant signals which may be indicative of language changes.



The screenshot shows the 'Enter Data Dialog - SplitsTree4 (version 4.18.0, built 12 Feb 2022)' window. The data is in NEXUS format, starting with '#NEXUS' and 'Begin data;'. The 'Dimensions ntax=315 nchar=96;' line is present. The 'Format datatype=standard;' line follows. The 'Matrix' section contains 29 rows of data, each representing a taxon (SK, WT, BT1, BT2, BT3, BT4, BT5, BT6, BT7, BT8, SH1, SH2, SH3, LUE1, LUE2, LUE3, LUE4, LUE5, LUE6, LUE7, LUE8, LUE9, LUE10, LUE11) and 10 columns of binary data (0 or 1). The data shows various patterns of 1s and 0s across the taxa. The window has a standard Windows-style interface with a menu bar (File, Edit, Format), scroll bars, and buttons for 'Cancel' and 'Execute'.

```

1 #NEXUS
2 Begin data;
3 Dimensions ntax=315 nchar=96;
4 Format datatype=standard;
5 Matrix
6 SK 0 0 0 1 0 0 0 0 0
7 WT 0 0 0 0 1 1 1 0 0
8 BT1 0 0 0 0 1 1 1 0 0
9 BT2 0 0 0 0 1 1 1 0 0
10 BT3 0 0 0 0 1 1 1 0 0
11 BT4 0 0 0 0 1 1 1 0 0
12 BT5 0 0 0 1 0 0 0 0 0
13 BT6 0 0 0 0 0 0 0 0 1
14 BT7 0 0 0 0 1 1 1 0 0
15 BT8 0 0 0 1 0 0 0 0 0
16 SH1 0 0 0 1 0 0 0 0 0
17 SH2 1 1 1 0 0 0 0 0 0
18 SH3 1 1 0 0 0 0 0 0 0
19 LUE1 1 1 1 0 0 0 0 0 0
20 LUE2 1 1 0 0 0 0 0 0 0
21 LUE3 1 1 0 0 0 0 0 0 0
22 LUE4 1 1 1 0 0 0 0 0 0
23 LUE5 1 1 1 0 0 0 0 0 0
24 LUE6 1 1 1 0 0 0 0 0 0
25 LUE7 1 1 1 0 0 0 0 0 0
26 LUE8 1 1 1 0 0 0 0 0 0
27 LUE9 1 1 1 0 0 0 0 0 0
28 LUE10 1 1 1 0 0 0 0 0 0
29 LUE11 1 1 1 0 0 0 0 0 0

```

Figure 5: The nexus-formatted data for computational modelling.

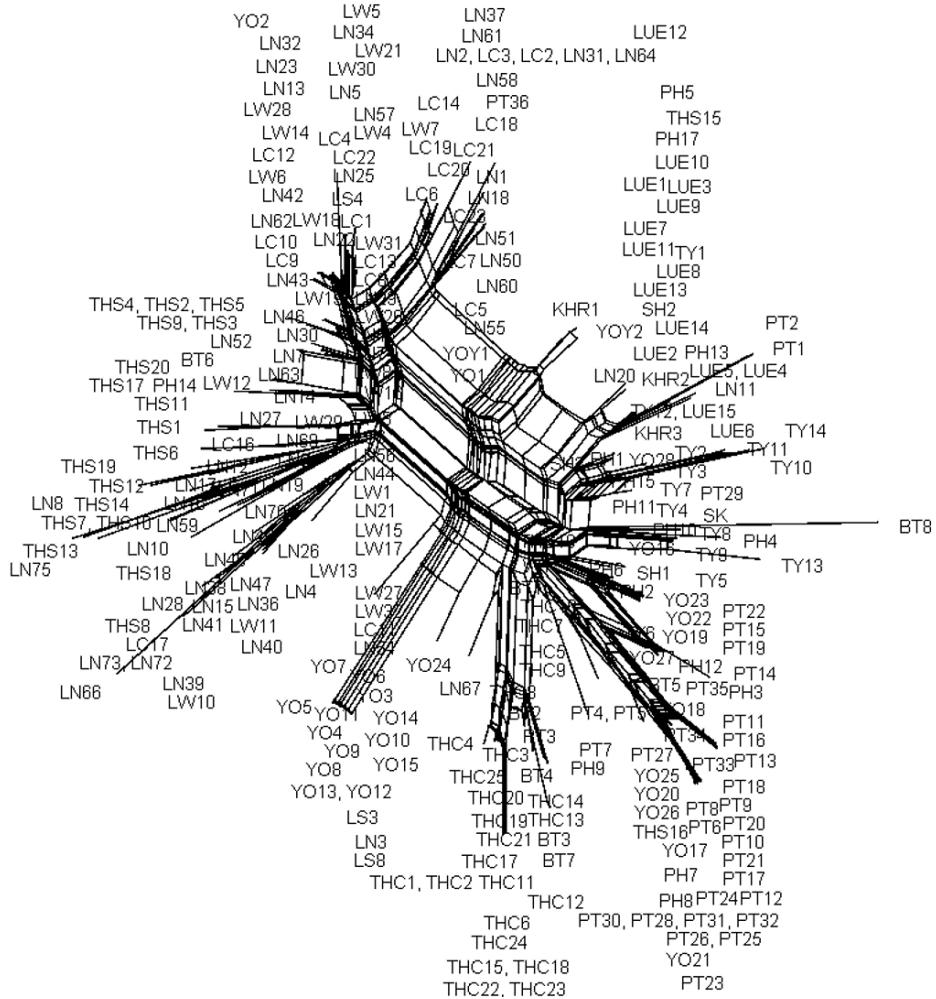


Figure 6: Neighbor-Net diagram for the tonological profiles of Tai dialects (n=315): LUE = Tai Lue; TY = Tai Yuan/Northern Thai; PT = Phu Thai; LN = Northern Lao; LC = Central Lao; LS = Southern Lao; LW = Western Lao/Northeastern Thai; THC = Central Thai; HIS = Southern Thai; SK = Saek; SH = Shan; WT = White Tai; BT = Black Tai; YO = Yo; YOY = Yoy; PH = Phuan; KHR = Khorat Thai.

## 4 Data interpretation and discussion

Through the Neighbor-Net diagram generated by the SplitsTree software (see Figure 6), several distinct dialect clusters can be identified. These clusters are discernible by the presence of dominant dialects indicated in the list below, and as annotated with different colours and arranged in an anti-clockwise direction within Figure 7.

- |                           |                      |
|---------------------------|----------------------|
| 1. Lao proper: LC, LS, LW | 5. Central Thai: THC |
| 2. Southern Thai: THS     | 6. Phu Thai: PT      |
| 3. Northern Lao: LN       | 7. Tai Yuan: TY      |
| 4. Yo: YO                 | 8. Tai Lue: LUE      |

In general, the results largely agree with the genealogical classification proposed in the previous studies, as discussed in Section 1. By identifying dialects which do not belong to their respective genealogical cluster based on the histori-cal-comparative basis, we look further into their current speaking areas on the map and migration history of their respective speech communities. In this context, two specific cases will be discussed as illustrative examples.

First, the Central Lao dialect speaker LN11 is situated within a convergence of the Tai Yuan and Lue clusters, as noted in Figure 8. As highlighted on the map using a red circle, the Northern Lao dialect LN11 is presently spoken within the major Tai Yuan and Lue speaking areas.

Second, the Central Lao dialect speaker LN67 aligns with the Central Thai cluster, as indicated by a red circle in Figure 9. By examining the map, the location of the Northern Lao dialect LN67 is currently situated adjacent to the major speaking areas of Central Thai dialects.

Both of the aforementioned cases highlight a scenario where the examined dialects have undergone or are undergoing a shift in their homophonous tone pattern of DL and DS. This shift reflects a convergence towards the tonal profile of a regional dialect within their newly settled area, in line with their sociocultural assimilation. This phenomenon can be effectively demonstrated by comparing their tone paradigms with our reconstructed common Lao tone system in Table 8, serving as a baseline.

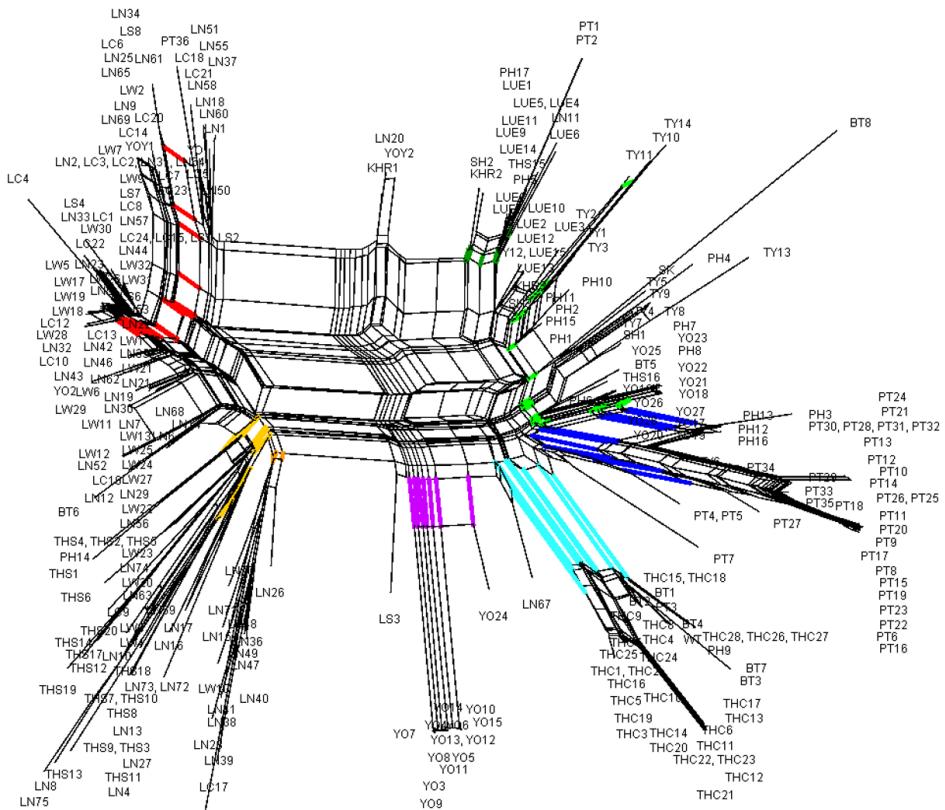


Figure 7: A Neighbor-Net diagram for the tonological profiles of Tai dialects (n=315) with identified clusters: LUE = Tai Lue; TY = Tai Yuan/Northern Thai; PT = Phu Thai; LN = Northern Lao; LC = Central Lao; LS = Southern Lao; LW = Western Lao/Northeastern Thai; THC = Central Thai; THS = Southern Thai; SK = Saek; SH = Shan; WT = White Tai; BT = Black Tai; YO = Yo; YOY = Yoy; PH = Phuan; KHR = Khorat Thai.

Table 8: A reconstructed tone paradigm of common Lao.

Initial consonant class at time of tonal splits	A	B	C	DL	DS
Aspirated	1		5		
Plain		4			1
Implosive	2				
Voiced	3		6		4

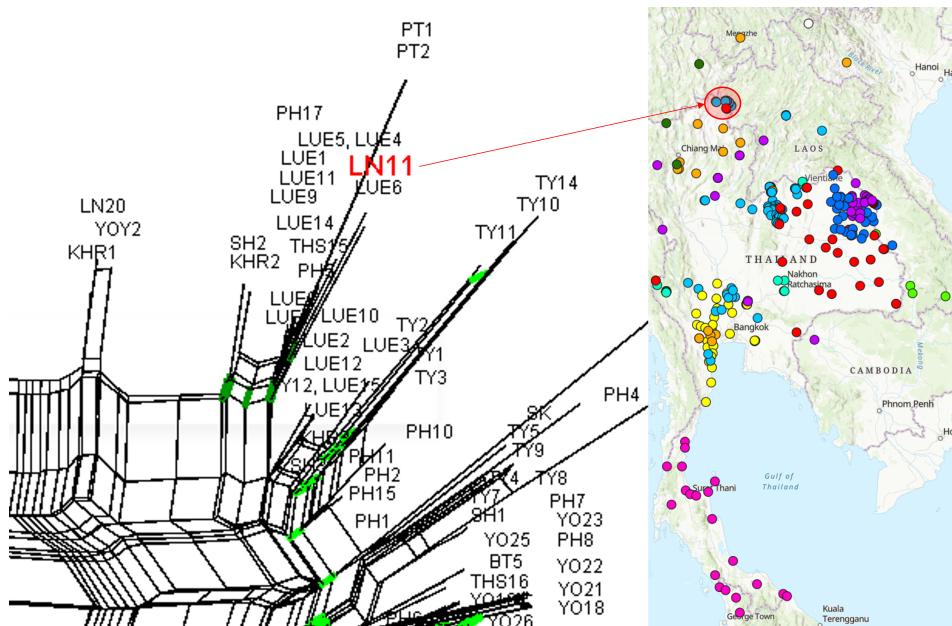


Figure 8: The Northern Lao dialect speaker (LN11) in Chiang Khong, Chiang Rai province, Thailand.

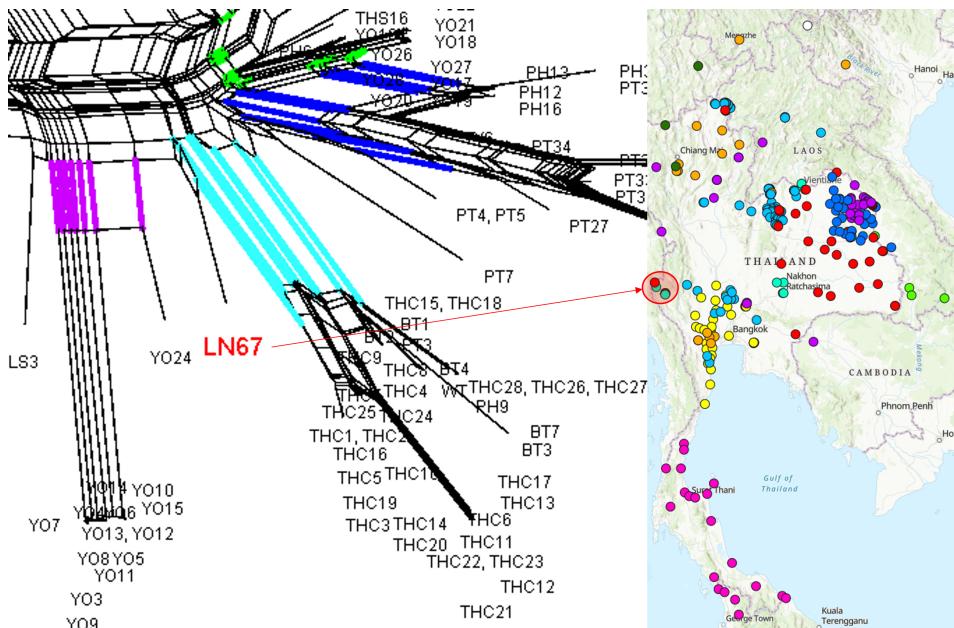


Figure 9: The Northern Lao dialect speaker (LN67) in Sangkhlaburi, Kanchanaburi province, Thailand.

There are two major characteristics of Lao tone paradigms in general and across dialects which have been, respectively, retained and undergone changes in the two dialects under investigation. The first feature is found within tone B where there is no distinction between the original initial consonant classes (as discussed in Section 2, Table 5). The second hallmark of the Lao tone paradigm is what is labelled in Tai-Kadai linguistics as the Lao ladders, which concerns the paradigmatic distribution of tones within tones C and DL, as marked in bold in Table 8.

On the one hand, the primary attribute of tone B remains largely intact across the majority of contemporary Lao dialects, including the two diaspora dialects under investigation (see Tables 9 and 10), irrespective of the geographical locations of their current speaking areas. Accordingly, Akharawatthanakun (2003: 337) observes that the split of tone B4 from tones B123 in any Lao dialects can be considered a strong indication of deviation from the common Lao pattern, as is evident in the case of Khorat Thai dialects regarded as hybrid dialects of Western Lao and Central Thai.

On the other hand, the second homophonous pattern between tones C and DL is gradually fading away within the two Northern Lao dialects under investigation. These dialects are spoken outside their core regions around Luang Phrabang in Northern Laos, as can be observed from the tone paradigms given in Tables 9 and 10.

Table 9: Tone paradigm of the Northern Lao dialect speaker (LN11) in Chiang Khong, Chiang Rai province, Thailand.

Initial consonant class at time of tonal splits	A	B	C	DL	DS
Aspirated			4		
Plain	1	3	5	3	1
Implosive					
Voiced	2		4	6	3

Our hypothesis is that the occurrence of a paradigmatic shift is ongoing, indicating a trend converging towards the tonal profiles prevalent in other Tai dialects, which are predominantly spoken by the majority populations in the areas. In order to assess the validity of our hypothesis, a comparative analysis is conducted using two additional baselines: Tai Yuan (Table 11) and Central Thai (Table 12).

Table 10: Tone paradigm of the Northern Lao dialect speaker (LN67) in Sangkhlaburi, Kanchanaburi province, Thailand.

Initial consonant class at time of tonal splits	A	B	C	DL	DS
Aspirated	1				
Plain			4	3	6
Implosive	2	3			5
Voiced			5	4	7

Table 11: Tone paradigm of common Tai Yuan.

Initial consonant class at time of tonal splits	A	B	C	DL	DS
Aspirated	1				
Plain		3	5	3	1
Implosive	2				
Voiced		4	6	4	5

Table 12: Tone paradigm of common Central Thai.

Initial consonant class at time of tonal splits	A	B	C	DL	DS
Aspirated	5				
Plain	1	2	3		2
Implosive					
Voiced	(6)	3	4	3	4

Upon comparing the paradigms of common Tai Yuan and common Central Thai with those present in the two Northern Lao dialects, it is obvious that the Lao ladders have been decomposed within the two diaspora dialects under investigation. Another observation pertains to tone A, wherein the Northern Lao dialect speaker in Sangkhlaburi (LN67) seems to have adopted a splitting pattern akin to that of Central Thai. This adaptation results in a differentiation of tone A1 (originating from aspirated initials) from the other tones, A234, within the paradigm.

Furthermore, we assess the significance of language contact by considering the sociolinguistic context of the two Northern Lao dialect speakers within the diaspora: LN11 and LN67. Although Akharawatthanakun (2003) does not explicitly report the extent of contact intensity and the specific domains of usage for both dialects, it is highly likely that bilingualism has resulted in the reorganisation and convergence of homophonous patterns within the Northern Lao tone paradigm towards the dominant regional dialects in their respective locations. In the case of LN11, this particular Northern Lao dialect speaker has primarily resided in the border region between northwestern Laos and northern Thailand on the Thai side, where the dominant regional language is Tai Yuan (Akharawatthanakun 2003: 150, 450). Similarly, the living environment of the Northern Lao dialect speaker LN67 involves contact with Central Thai, the dominant regional language in western Thailand, as well as with other diasporic Lao dialects (Akharawatthanakun 2003: 150, 448).

Based on our tonological profile data (as shown in Tables 9 and 10), exposure to Tai Yuan and Central Thai speaking environments, respectively, emerges as a highly plausible factor contributing to the decomposition of the Lao ladders within tones C and DL. Additionally, the Central Thai tone paradigm also provides a model for the reorganisation of tone slots within tone A for the Northern Lao dialect speaker LN67 (as discussed above). At the same time, it appears that homogeneity within the common Lao tones B1234 (as highlighted in Tables 5 and 8) remains stable and resistant to contact-induced changes in the tone paradigm of these two speakers.

As a supplementary remark, we may also consider a language-internal perspective and posit a hypothesis that tones DL and DS paradigmatically and cognitively lie at a deeper level of prominence and speaker's awareness. This heightened prominence might stem from their vague status within the tone paradigm and the variations observed among individual dialect speakers, due to their probabilistic status as a low-frequency syllable type and a low-probable tone type in the language system (see e.g. the case of Chinese dialects in Wiener & Ito 2015). Consequently, these distinctive characteristics may make them more prone to

alternation and change when compared to the salient tones A, B, and C, which are often emphasised in the process of acquisition, more systematically taught in school and acquired by children from their elementary education. This scenario is usually observed in the teaching of standard languages, Lao and Thai, as well as Vietnamese, Cantonese and Mandarin (see e.g. Bar-Lev 1991). However, the method used in the current study is not specifically designed to test this particular claim. Nevertheless, this aspect can offer an interesting direction for future research which could bridge the domains of dialectology and cognitive sciences. Such exploration could provide valuable insights into the dynamics of tone paradigms and their evolution.

## 5 Conclusions

In the present study, we have examined and discussed instances of language shift occurring in various regions where Tai dialect speakers are shifting their language whose tone paradigm structure subsequently also converges with a dominant regional dialect. From the perspective of tone paradigm, the convergence has sometimes resulted in the emergence of transitional dialect systems, in which a protosystem has undergone restructuring, aligning itself more closely with the model provided by the dominant regional dialect. These transitional dialect systems of tone paradigm stand out when examined through a quantitative approach in combination with the conventional comparative method. It enables identification of the protostructure of tone paradigm from which the transitional dialect systems have diverged. In any case, our results do not post significant challenges to the genealogical classification proposed in the previous studies, as the signals of change observed in the present study primarily pertain to contact-induced changes occurring subsequent to the dispersal stages of individual Tai subbranches.

At a methodological level, we have also demonstrated the utility of Neighbor-Net algorithm as an effective tool for identifying such instances from a big pool of data. Our method employed to gather, organise and analyse the data can potentially offer a preliminary model for scholars engaged in the studies of Tai dialects as well as for those researching dialects of other Mainland Southeast Asian languages with tones. As more data, particularly relating to tone paradigms, have been continuously collected from field during the recent decades, this methodological model should also facilitate scholars and dialectological studies with a focus on tonal aspects in embracing an emerging trend within digital humanities and big data studies.

Lastly, our vital message to scholars engaged in language documentation and description is that the collection of sociolinguistic data stands on equal footing with the description of language features. From the present study, we see that insufficient description of sociolinguistic context regarding informants may pose a challenge when attempting to establish a contact-based explanation for a language change. With the inclusion of such comprehensive sociolinguistic information about the speakers, their speech communities and their linguistic repertoires, numerous finely-tuned analyses centred around cross-factor correlations can yield significantly more insightful understanding of the language situation. Such analyses will contribute to the discussion of language change and the diversification of dialects, a dynamic process which continues to evolve as we advance into the 21st century.

## Abbreviations

### Tones

- A Tone A, smooth syllable
- B Tone B, smooth syllable
- C Tone C, smooth syllable
- DL Tone D, long vowel, checked syllable
- DS Tone D, short vowel, checked syllable

### Languages

BT	Black Tai	SH	Shan
KHR	Khorat Thai	SK	Saek
LC	Central Lao	THC	Central Thai
LN	Northern Lao	THS	Southern Thai
LS	Southern Lao	TY	Tai Yuan/Northern Thai
LUE	Tai Lue	WT	White Tai
LW	Western Lao/Northeastern Thai	YO	Yo
PH	Phuan	YOY	Yoy
PT	Phu Thai		

### Acknowledgements

This research was supported by the Grant for Graduate Student 2021 from the King Prajadhipok and Queen Rambhai Barni Memorial Foundation, awarded to

Saknarin Pimvunkum for the research project “Classification of Vientiane Lao dialects in Western Thailand by tone paradigm” and was part of the project “North-east and Southeast Asian Studies Network in Finland and Thailand (NSEANET)” funded by the Finnish National Agency for Education (EDUFI) during 2020–2023. This study was exempted from ethical review by the Committee for Research Ethics (Social Sciences), Faculty of Social Sciences and Humanities, Mahidol University (2022/001.1701, 17 January 2022).

## References

- Akharawatthanakun, Phinnarat. 2003. *Tone change: A case study of the Lao language*. Chulalongkorn University. (Doctoral dissertation).
- Akharawatthanakun, Phinnarat. 2020. Tonal diversity and tone Sandhi in Lue. *Journal of Liberal Arts, Thammasat University* 20(2). 513–548. DOI: 10.14456/lartstu.2020.31.
- Bar-Lev, Zev. 1991. Two innovations for teaching tones. *Journal of the Chinese Language Teachers Association* 26(3). 1–24. <https://eric.ed.gov/?id=EJ447341>.
- Brown, J. Marvin. 1985. *From ancient Thai to modern dialects*. Bangkok: White Lotus.
- Bryant, David & Vincent Moulton. 2004. Neighbor-net: An agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution* 21(2). 255–265. DOI: 10.1093/molbev/msh018.
- Bunyasathit, Wannaporn, Chantas Pientam & Theptida Silapabanleng. 2016. The historical development of LaoWiang people in U Thong district, Suphanburi province. *Journal of Nakhonratchasima College* 11(1). 26–38. [http://journal.nmc.ac.th/th/admin/Journal/2560Vol11No1\\_702.pdf](http://journal.nmc.ac.th/th/admin/Journal/2560Vol11No1_702.pdf).
- Burusphat, Somsonge. 2012. Tones of Thai Song varieties. *Journal of the Southeast Asian Linguistics Society* 5. 32–48. <http://hdl.handle.net/1885/9118>.
- Canilao, Kritsana. 2010. *Tonal geography of the provinces of Central Thailand*. Mahidol University. (Doctoral dissertation). <http://mulinet11.li.mahidol.ac.th/e-thesis/2553/cd446.1/4836363.pdf>.
- Chamberlain, James R. 1975. A new look at the history and classification of the Tai languages. In Jimmy G. Harris & James R. Chamberlain (eds.), *Studies in Tai linguistics in honor of William J. Gedney*, 49–66. Bangkok: Central Institute of English Language. <http://sealang.net/sala/archives/pdf8/chamberlain1975new.pdf>.

- Damanhuri, Umayah. 2004. The classification of some Thai dialects spoken in Kedah. In Somsonge Burusphat (ed.), *Papers from the Eleventh Annual Meeting of the Southeast Asian Linguistics Society 2001*, 167–182. Tempe, AZ: Arizona State University Programme for Southeast Asian Studies Monograph Series Press. <http://sealang.net/sala/archives/pdf4/damanhuri2004classification.pdf>.
- Dockum, Rikker. 2019. *The Tonal Comparative Method: Tai tone in historical perspective*. Yale University. (Doctoral dissertation).
- Edmondson, Jerold A. 1990. Kam tone splits and the variation of breathiness. In Jerold A. Edmondson, Crawford Feagin & Peter Mühlhäusler (eds.), *Development and diversity: Language variation across time and space (a Festschrift for Charles-James N. Bailey)*, 187–202. Arlington, TX: Summer Institute of Linguistics. <https://www.sil.org/resources/archives/8425>.
- Edmondson, Jerold A. & David B. Solnit. 1997. Introduction. In Jerold A. Edmondson & David B. Solnit (eds.), *Comparative Kadai: The Tai branch*, 1–32. Arlington, TX: Summer Institute of Linguistics.
- Ferlus, Michel. 2004. The origin of tones in Viet-Muong. In Somsonge Burusphat (ed.), *Papers from the Eleventh Annual Meeting of the Southeast Asian Linguistics Society 2001*, 297–313. Tempe, AZ: Arizona State University Programme for Southeast Asian Studies Monograph Series Press. <http://sealang.net/sala/archives/pdf8/ferlus2004origin.pdf>.
- Gedney, William J. 1972. A checklist for determining tones in Tai dialects. In M. Estellie Smith (ed.), *Studies in linguistics in honor of George L. Trager*, 423–437. The Hague: de Gruyter Mouton.
- Grünthal, Riho & Johanna Nichols. 2016. Transitivizing-detransitivizing typology and language family history. *Lingua Posnaniensis* 58(2). 11–31. DOI: 10.1515/linpo-2016-0008.
- Handel, Zev. 2014. Historical phonology of Chinese. In C.-T. James Huang, Y.-H. Audrey Li & Andrew Simpson (eds.), *The handbook of Chinese linguistics*, 576–598. Oxford: John Wiley & Sons. DOI: 10.1002/9781118584552.ch22.
- Hartmann, John F. 2008. The Lue language. In Anthony V. N. Diller, Jerold A. Edmondson & Yongxian Luo (eds.), *The Tai-Kadai languages*, 254–297. London: Routledge.
- Haudricourt, André-Georges. 1954. De l'origine des tons en vietnamien. *Journal Asiatique* 242. 69–82. [https://lacito.hypotheses.org/files/2015/12/Haudricourt\\_1954\\_Origine-Tons-Vietnamien\\_scan.pdf](https://lacito.hypotheses.org/files/2015/12/Haudricourt_1954_Origine-Tons-Vietnamien_scan.pdf).
- Hill, Nathan. 2019. *The historical phonology of Tibetan, Burmese, and Chinese*. Cambridge: Cambridge University Press.

- Hudak, Thomas John. 2008. *William J. Gedney's comparative Tai source book* (Oceanic Linguistics Special Publications 34). Honolulu, HI: University of Hawai'i Press. <https://www.jstor.org/stable/20532978>.
- Huson, Daniel H. & David Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23(2). 254–267. DOI: 10.1093/molbev/msj030.
- Kingston, John. 2011. Tonogenesis. In Marc van Oostendorp, Colin J. Ewen, Elizabeth Hume & Keren Rice (eds.), *The Blackwell companion to phonology*, 2304–2333. Oxford: Wiley-Blackwell. DOI: 10.1002/9781444335262.wbctp0097.
- Koowatthanasiri, Kanjana. 1981. *The tones of Nyo*. Chulalongkorn University. (MA thesis). <http://cuir.car.chula.ac.th/handle/123456789/35801>.
- Li, Fang-Kuei. 1977. *A handbook of comparative Tai* (Oceanic Linguistics Special Publications 15). Manoa, HI: University Press of Hawai'i. <https://www.jstor.org/stable/20006684>.
- Liang, Min & Junru Zhang. 1996. *Dòngtái yǔ Gailùn [An introduction to the Kam-Tai languages]*. Beijing: China Social Sciences Publishing House.
- Liao, Hanbo. 2016. *Tonal development of Tai languages*. Payap University. (MA thesis). [https://www.academia.edu/download/49907598/Final-TONAL\\_DEVELOPMENT\\_OF\\_TAI\\_LANGUAGES-0714.pdf](https://www.academia.edu/download/49907598/Final-TONAL_DEVELOPMENT_OF_TAI_LANGUAGES-0714.pdf).
- Liao, Hanbo. 2023a. An integrated tone box scheme for determining tones in Tai varieties beyond Southwestern Tai: Diachronic and synchronic concerns. *Folia Linguistica* 57(s44–s1). 199–244. DOI: 10.1515/flin-2022-2048.
- Liao, Hanbo. 2023b. *The formation of Lingnan linguistic traits: Typological structures and diachronic issues*. The University of Hong Kong. (Doctoral dissertation).
- Luo, Meizhen. 1988. Dǎi Tàï cíhuì bìjiào [A comparison of Dai and Thai vocabularies]. *Mínzú Yǔwén [Minority Languages of China]* (2). 26–34.
- Luo, Yongxian. 1997. *The subgroup structure of the Tai languages: A historical-comparative study* (Journal of Chinese Linguistics Monograph Series 12). Hong Kong: The Chinese University of Hong Kong Press. <https://www.jstor.org/stable/23887080>.
- Maddieson, Ian. 1984. The effects on F0 of a voicing distinction in sonorants and their implications for a theory of tonogenesis. *Journal of Phonetics* 12(1). 9–15. DOI: 10.1016/S0095-4470(19)30845-9.
- Maddison, David R., David L. Swofford & Wayne P. Maddison. 1997. Nexus: An extensible file format for systematic information. *Systematic Biology* 46(4). 590–621. DOI: 10.1093/sysbio/46.4.590.

- Michaud, Alexis & Bonny Sands. 2020. Tonogenesis. In Mark Aronoff (ed.), *Oxford research encyclopedia of linguistics*. Oxford: Oxford University Press. DOI: 10.1093/acrefore/9780199384655.013.748.
- Mitani, Yasuyuki. 1977. Tai-Kadai shogo no gengonendaigaku-teki kōsatsu [Linguistic chronology of Tai-Kadai languages]. *Tōnan'ajia kenkyū [Southeast Asian Studies]* 15(3). 421–429. <https://kyoto-seas.org/pdf/15/3/150309.pdf>.
- Nichols, Johanna. 2020. Dispersal patterns shape areal typology. In Mily Crevels & Pieter Muysken (eds.), *Language dispersal, diversification, and contact: A global perspective*, 25–43. Oxford: Oxford University Press. DOI: 10.1093/oso/9780198723813.003.0002.
- Ostapirat, Weera. 2005. Kra-dai and Austronesian: Notes on phonological correspondences and vocabulary distribution. In Roger Blench, Laurent Sagart & Alicia Sanchez-Mazas (eds.), *The peopling of East Asia: Putting together archaeology, linguistics and genetics*, 107–131. London: Routledge. DOI: 10.4324/9780203343685.
- Pittayaporn, Pittayawat. 2009. *The phonology of Proto-Tai*. Cornell University. (Doctoral dissertation). <https://ecommons.cornell.edu/bitstream/handle/1813/13855/Pittayaporn,%20Pittayawat.pdf>.
- Pittayaporn, Pittayawat. 2014. Layer of Chinese loanwords in Proto-Southwestern Tai as evidence for the dating of the spread of Southwestern Tai. *Manusya: Journal of Humanities, Special Issue No 20* 17(3). 47–68. DOI: 10.1163/26659077-01703004.
- Piyabhan, Bung-on. 1998. *Laaw nai krung rāttanákoosin* [Lao people in the Rattanakosin kingdom]. Bangkok: The Thailand Research Fund.
- Ratliff, Martha. 2010. *Hmong-Mien language history*. Canberra: Pacific Linguistics. DOI: 10.15144/PL-613.
- Ratliff, Martha. 2015. Tonoexodus, tonogenesis, and tone change. In Patrick Honneybone & Joseph Salmons (eds.), *The Oxford handbook of historical phonology*, 245–261. Oxford: Oxford University Press. DOI: 10.1093/oxfordhb/9780199232819.013.021.
- Srithonrat, Potjanee. 1983. *A tonal comparison of Phuthai dialect in three provinces*. Mahidol University. (MA thesis).
- Szeto, Pui Yiu, Umberto Ansaldi & Stephen Matthews. 2018. Typological variation across Mandarin dialects: An areal perspective with a quantitative approach. *Linguistic Typology* 22(2). 233–275. DOI: 10.1515/lingty-2018-0009.
- Szeto, Pui Yiu & Chingduang Yurayong. 2021. Sinitic as a typological sandwich: Revisiting the notions of altaicization and taicization. *Linguistic Typology* 25(3). 551–599. DOI: 10.1515/lingty-2021-2074.

- Thurgood, Graham. 2002. Vietnamese and tonogenesis: Revising the model and the analysis. *Diachronica* 19(2). 333–363. DOI: 10.1075/dia.19.2.04thu.
- Wiener, Seth & Kiwako Ito. 2015. Do syllable-specific tonal probabilities guide lexical access? Evidence from Mandarin, Shanghai and Cantonese speakers. *Language, Cognition and Neuroscience* 30(9). 1048–1060. DOI: 10.3798/2014.946934.
- Wulff, Kurt. 1934. *Chinesisch und Tai: Sprachvergleichende Untersuchungen*. Copenhagen: Levin & Munksgaard.
- Yang, Tongyin & Jerold A. Edmondson. 2008. Kam. In Anthony V. N. Diller, Jerold A. Edmondson & Yongxian Luo (eds.), *The Tai-Kadai languages*, 509–584. London: Routledge.
- Yurayong, Chingduang & Pui Yiu Szeto. 2020. Altaicization and de-altaicization of Japonic and Koreanic. *International Journal of Eurasian Linguistics* 2(1). 108–148. DOI: 10.8833-12340026.
- Zhang, Junru. 1980. Yuánshǐ tái yǔ shēngmǔ lèibíé tànsuǒ [An exploration of the classification of initial consonants of proto-Tai]. *Mínzú Yǔwén [Minority Languages of China]* (2). 31–40.
- Zhu, Xiaonong, Mingying Wei & Junfang Wang. 2016. Shíwǔ tiáohé qìtiáo: Dòngyǔ róngjiāngxiàn kǒuzhài fāngyán ànlì [Fifteen tones and breathy tones: A case of the Kouzhai Dong dialect of Rongjiang county]. *Mínzú Yǔwén [Minority Languages of China]* 5. 12–24.

# Name index

- Akharawatthanakun, Phinnarat, 405, 407, 415, 425, 427  
Al'muxamedova, Zel'fa, 246  
Allen, Harold, 204, 206  
Ammon, Ulrich, 294, 302, 335  
Anders, Christina Ada, 321  
Anderson, Kay, 78, 83  
Andreeva, Bistra, 252, 261  
Andrews, David R., 242, 249, 261  
Antieau, Lamont, viii  
Antieau, Lamont D., 204, 214  
Arabie, Phipps, 377  
Arbatzat, Hartmut, 175  
Auer, Peter, 225, 242, 297, 299, 302, 305, 321  
Avanesov, Ruben, 244  
Ayres-Bennett, Wendy, 288, 290  
Bachelier, Veronique, 40  
Bailey, Charles-James N., 79  
Bailey, Guy, 156  
Baker, Philip, 204  
Baltaretu, Adriana, 85  
Bar-Lev, Zev, 428  
Baranowski, Maciej, 130  
Barnes, Jonathan, 243, 244, 247, 257  
Barreda, Santiago, 131  
Bates, Douglas, 90, 135  
Baugh, John, 6  
Bauman, Kurt, 8, 33  
Baur, Gerhard W., 335, 336  
Baxter, Kimberley, vii, 4, 9, 15, 16, 32  
Bayer, Josef, 62  
Beal, Joan, 80, 288, 290  
Beaman, Karen V., 156  
Beider, Alexander, 125, 144  
Beijering, Karin, 345  
Belkin, Dmitrij, 223  
Bellamy, John, 288, 290  
Bergenholtz, Henning, 353  
Berthele, Raphael, 353  
Besch, Werner, 309  
Bethin, Christina Y., 245, 246  
Birnbaum, Solomon Asher, 125, 127  
Blake, Renée, 5, 13, 15, 19  
Blaßnigg, Julian, vii  
Bleaman, Isaac L., 124, 131  
Blodgett, Sue Lin, 10, 13, 42  
Bloemhoff, Henk, 85, 86, 175, 177, 178, 181, 192  
Blommaert, Jan, 288  
Blondeau, Hélène, 156  
Boberg, Charles, 298  
Boersma, Paul, 252, 253  
Bogner, Alexander, 224  
Bondarko, Lija, 243, 244, 247, 249, 257, 262  
Borg, Ingwer, 378, 385  
Borges Völker, Emanuel, 183  
Bouma, Gosse, 183  
Britain, David, 78, 81–83, 107, 108, 110, 118  
Brown, J. Marvin, 407, 412, 413  
Brückner, Dominik, 353

## *Name index*

- Bryant, David, 416, 417  
Buchstaller, Isabelle, 156  
Bunin Benor, Sarah, 219–222, 224, 228  
Bunyasathit, Wannaporn, 405  
Burkette, Allison, viii, 214  
Burrell, Courtney M., 296  
Burusphat, Somsonge, 407  
Busch, Brigitta, 226, 230  
Buschfeld, Sarah, 303  
Callies, Marcus, 309  
Canilao, Kritsana, 407  
Carmichael, Katie, 165  
Castro, Andy, 361, 363–367, 383  
Cedergren, Henrietta, 294  
Čekmonas, Valerij, 246  
Chamberlain, James R., 405  
Chambers, Jack K., 95, 272, 288, 289, 291, 302, 362  
Chang, Charles B., 140  
Chao, Yuen-Ren, 362, 364, 365, 367, 371, 375  
Chapman, Don, 288  
Charmaz, Kathy, 226  
Chasteen, Alison L., 157, 167  
Chen, Hailun, 375  
Chen, Xiaojan, 375  
Chen, Xiaojin, 375  
Cheng, Chin-Chuan, 363, 367, 368  
Cheung, Yat-Shing, 368, 375  
Chinese Academy of Social Sciences, 372, 373, 377  
Cieri, Christopher, 156  
Clyne, Michael G., 287, 290, 302  
Cohen, Steven M., 228  
Collins, Peter C., 110  
Comrie, Bernard, 11, 12, 27, 242  
Cooper, Levi, 128, 130  
Corbett, Greville G., 242  
Costa, James, 289  
Coulmas, Florian, 291  
Council of Local Authorities for International Relations, 109  
Coupland, Nikolas, 108  
Cramer, Jennifer, 210, 226  
Crosswhite, Katherine M., 243, 244  
Cukor-Avila, Patricia, 156  
D’Arcy, Alexandra, 110  
D’jačenko, Svetlana, 246  
Damanhuri, Umayyah, 407  
Darwin, Clayton, 204  
Davies, Winifred V., 288  
Day, R. A., 84, 95  
De Cillia, Rudolf, 288, 301, 304, 305, 308  
De Ridder, Reglindis, 307  
De Schutter, Georges, 85  
de Sousa, Hilário, 391  
de Vriend, Folkert, 298  
Demirci, Mahide, 96  
Denz, Josef, 320  
Deumert, Ana, 288  
Diercks, Willy, 321  
Dieth, Eugen, 34  
DiÖ, 306  
Do, Youngah, 366  
Dockum, Rikker, 413  
Dollinger, Stefan, viii, 288–291, 293, 294, 296, 298–300, 302, 304–310  
Dollmayr, Viktor, 320  
Dubina, Andrei, 243, 244, 260  
Dudenredaktion, 335  
Durgasingh, Ryan, 306  
Dürscheid, Christa, 294, 301, 304, 306

- Duryagin, Pavel V., 258  
Dwoskin, Elizabeth, 9, 27  
Eckert, Penelope, 222, 223  
Edmondson, Jerold A., 404, 409, 412  
Eichhoff, Jürgen, 60  
Eickmans, Heinz, 320  
Eisenstein, Jacob, 36, 50  
Elmentaler, Michael, 321  
Elspaß, Stephan, 60, 61, 294, 300, 301,  
    304–306, 309  
Ericsdotter, Christine, 249  
Ernst, Peter, 300  
Erofeeva, Elena, 242, 247–250, 257,  
    258, 261, 262  
Erofeeva, Tamara, 242, 250  
Evans, Betsy E., 206  
Evans, Bronwen G., 119  
Faninger, Kurt, 291  
Ferlus, Michel, 409  
Fisher, Linda, 288  
Fisher, Sabriya, 31, 32, 34–36, 50  
Fishman, Joshua A., 221, 288  
Flick, Uwe, 229  
Francis, Winthrop Nelson, 375  
Gal, Susan, 226, 230  
Gandour, Jackson, 369, 370, 375, 376,  
    378, 383, 388, 390  
Gardner, Matt Hunt, 166  
Garellek, Marc, 142  
Gedney, William J., 409, 410  
George, Crissandra, 206  
Giles, Howard, 108  
Glasser, Paul, 127  
Glauninger, Manfred, 294, 308  
Goebl, Hans, 70, 362, 368  
Goeman, Ton, 85  
Gold, David L., 221, 222  
Gooskens, Charlotte, 82, 85, 345, 376  
Goossens, Jan, 79, 177  
Gopal, Deepthi, 39  
Gould, Peter, 80  
Grammatčikova, E. V., 242, 244, 247,  
    248, 257, 258, 262, 263  
Grave, Edouard, 184  
Green, Lisa J., 4, 6, 10, 11, 13, 14  
Gregersen, Frans, 156  
Grieve, Jack, 4, 10  
Grimm, Jacob, 293  
Groenen, Patrick J. F., 378, 385  
Grondelaers, Stefan, 288  
Grootaers, Willem A., 362  
Grünthal, Riho, 416  
Hall-Lew, Lauren, 136, 165  
Hamilton-Brehm, Anne Marie, 204  
Handel, Zev, 375, 409, 411  
Hanzawa, Yasushi, 275, 280  
Harris, Wendell A., 35  
Harshman, Richard A, 369, 370, 375,  
    376, 378, 383, 388, 390  
Hartmann, John F., 413  
Hary, Benjamin H., 219, 220, 222, 224  
Haß-Zumkehr, Ulrike, 353  
Haudricourt, André-Georges, 409  
Haugen, Einar, 291  
Havinga, Anna D., 291  
Hay, Jennifer, 136  
Heeringa, Wilbert, 88, 94, 95, 345,  
    361, 362, 364, 376, 378  
Hehner, Stefanie, 309  
Heisler, Troy, 272  
Herman, Matt, 42  
Herrgen, Joachim, 300, 301, 305, 336,  
    345  
Herrmann-Winter, Renate, 175

## *Name index*

- Hertz, Birgitte, 244  
Herzog, Marvin I., 125  
Hettler, Ivonne, 321  
Hickey, Raymond, 288  
Hiestermann, Heike, 175  
Hill, Nathan, 409, 411  
Hinskens, Frans, 88  
Hirano, Keiko, viii, 108, 110  
Hock, Hans Henrich, 280  
Höck, Leonhard, 64  
Hofer, Lorenz, 320, 341  
Hofer, Silvia, 303  
Hölscher, Christoph, 85  
Horvath, Barbara, 9, 10, 35  
Howe, Darin, 31  
Howren, Robert, 204  
Huang, Qun, 375  
Hubert, Lawrence, 377  
Hudak, Thomas John, 407, 412  
Hudley, Anne H. Charity, 289  
Hughes, Adam, 9  
Hundt, Markus, 321  
Huson, Daniel H., 417  
Hutton, Christopher M., 289, 295, 296, 300, 309, 310  
Hyman, Larry M., 363  
Inoue, Fumio, 275, 280  
Iosad, Pavel, 242–244, 246  
Irvine, Judith T., 206, 222, 223, 226, 230  
Ito, Kiwako, 427  
Iverson, Paul, 119  
Jacobs, Neil G., 125, 127, 142, 221, 228  
Jahns, Esther, viii, 220, 222, 223, 225, 229, 230, 233  
Jandl, Marco, 293  
Jankowski, Bridget L., 110  
Jelinek, Yeshayahu A., 128–130  
Jellinghaus, Hermann, 175  
Jenkins, John Michael, 84  
Jeszenszky, Péter, 82, 83, 96  
Johnson, Ellen, 206  
Johnson, Sasha Rosena, 6  
Johnstone, Barbara, 222  
Jones, Mari C., 36  
Jones, Taylor, 4, 10, 50  
Jongenburger, Willy, 85  
Jørgensen, Anna, 6, 7, 13, 35  
Joseph, John E., 288  
Kahl, Heinrich, 175, 176  
Kahle, David, 128  
Kahn, Lily, 219  
Kalenčuk, Marija, 242  
Kapferer, Stefanie, 352  
Kasatkina, Rozalija, 243, 246, 249, 256  
Kathrein, Yvonne, viii  
Katz, Dovid, 127  
Kautzsch, Alexander, 31, 32, 34, 35, 303  
Keren-Kratz, Menachem, 128, 130  
Kerswill, Paul, 119  
King, Sharese, 5  
Kingston, John, 409  
Kircher, Ruth, 288  
Klagsbrun Lebenswerd, Patric Joshua, 228  
Klein, Karl Kurt, 330, 338  
Kleiweg, Peter, 376  
Knjazev, Sergej, 243, 258  
Koch, Peter, 231  
Kocharov, Daniil, 244  
Kochetov, Alexei, 246  
Kodzasov, Sandro, 243  
Komoróczy, Sonja Rahel, 128

- Konen-Witzel, Katrin, 175  
Koowatthanasiri, Kanjana, 407  
Koppensteiner, Wolfgang, 301, 309  
Kostadinova, Viktorija, 166  
Krämer, Philipp, 287, 288  
Kranzmayer, Eberhard, 294, 297, 320,  
    328, 331, 333  
Krassnigg, Albert, 292  
Krause, Marion, 242  
Kremer, Ludger, 298  
Kroch, Anthony S., 109, 110  
Krogh, Steffen, 128, 129  
Kronsteiner, Otto, 293  
Kroskrity, Paul V., 231  
Krysin, Leonid, 242  
Kubozono, Haruo, 276  
Kurath, Hans, 203, 206  
Kürschner, Sebastian, 85  
Kuznecov, Vladimir, 243, 244, 246  
Kuznetsova, Alexandra, 135  
  
Labov, William, 5, 6, 34–36, 79, 130,  
    132, 155, 249, 261, 307  
Lai, Ryan Ka Yau, 366  
Lameli, Alfred, 186, 188, 193, 321, 345  
Lämmert, Eberhard, 293, 297, 309  
Landesstatistik Salzburg, 58  
Landolt, Christoph, 320  
Lanehart, Sonja, 4  
Langer, Nils, 288, 300, 306, 309  
Lapteva, Ol'ga, 249, 260  
Lasch, Agathe, 179  
Lawton, Carol A, 84  
Leemann, Adrian, 60, 165  
Leinonen, Therese, 386, 387  
Lenz, Alexandra N., 225, 301, 309,  
    321, 353  
Leopold, Johan A., 182  
Leopold, Lubbertus, 182  
  
Lewi, Hermann, 292, 308  
Lewin, Kurt, 306  
Li, Fang-Kuei, 405, 407  
Li, Lianjin, 373  
Liang, Jinrong, 373  
Liang, Min, 373, 409  
Liao, Hanbo, 404, 405, 409, 410  
Lin, Yi, 375  
Lindow, Wolfgang, 179–181  
Liu, Caifeng, 375  
Liu, Cunhan, 375  
Lorenz, David, 109  
Lücht, Wilko, 179  
Łukaszewicz, Beata, 260  
Luo, Meizhen, 404  
Luo, Yongxian, 405  
Lynch, Kevin, 80  
  
Maas, Utz, 291  
MacEachren, Alan M., 84  
Maddieson, Ian, 409  
Maddison, David R., 417, 418  
Madsen, Michael, 204  
Maegaard, Marie, 288  
Mair, Friedrich, 66  
Malmasi, Shervin, 186  
Martin, Katie, 11, 12, 14  
Masis, Tessa, 4  
Masor, Alyssa, 125, 144  
Mauser, Peter, 60  
Maxwell, Alexander, 288  
McAuliffe, Michael, 131  
Mechler, Johanna, 156  
Meer, Philipp, 306  
Meier, Stefanie, 320, 341  
Menz, Wolfgang, 224  
Meuser, Michael, 224  
Michaud, Alexis, 409  
Mitani, Yasuyuki, 404

## Name index

- Mitchell, Travis, 33  
Molczanow, Janina, 244, 260  
Möller, Robert, 60  
Montello, Daniel R., 77, 83, 84, 95  
Montgomery, Chris, 80, 81, 96  
Moody, Simanique Davette, 4, 32  
Moschonas, Spiros A., 288  
Moulton, Vincent, 416  
Muhr, Rudolf, 290, 297, 301, 302, 309  
Myhill, John, 221, 222
- Nadler, Josef, 295, 300  
Nagel, Ulrike, 224  
Naro, Anthony Julius, 156  
National Language Research Institute, 269–271, 273  
National Statistics Center of Japan, 109, 118  
Nerbonne, John, 79, 85, 91, 186, 345, 376, 386, 387  
Nesbitt, Monica, 165  
Nguyen, Dong, 33  
Nichols, Johanna, 416  
Niebaum, Hermann, 175  
Niedzielski, Nancy, 210  
Niehaus, Konstantin, 62, 74, 300, 301, 309  
Nikolaeva, Tatjana, 246  
Nove, Chaya R., 124, 131  
Nove, Chaya R., viii, 124, 130–132, 136, 142, 145  
Nycz, Jennifer R., 119, 136
- Oakes, Leigh, 308  
Oesterreicher, Wulf, 231  
Okumura, Ayako, 280  
Omdal, Helge, 119  
Orton, Harold, 34  
Ostapirat, Weera, 403
- Pabst, Katharina, viii  
Padgett, Jaye, 244, 246, 247  
Palliwoda, Nicole, 321  
Panov, Mixail, 242  
Passarelli, Nicholas, 206  
Paufošima, Rozalija, 244, 246  
Payne, Arvilla Chapin, 280  
Pederson, Lee, 203, 204, 206, 209  
Pedregosa, Fabian, 187  
Peltz, Rakhmiel, 221  
Pennington, Jeffrey, 184  
Pennycook, Alastair, 290  
Peters, Friedrich Ernst, 181  
Petersen, Julius, 293  
Pfalz, Anton, 298  
Pheiff, Jeffrey, 177, 178, 189  
Piller, Ingrid, 291  
Pittayaporn, Pittayawat, 404, 405, 407, 409–412  
Piyabhan, Bung-on, 405  
Plank, Barbara, 7  
Pohl, Heinz-Dieter, 293  
Polonsky, Antony, 127, 128  
Popkema, Jan, 181  
Portugali, Juval, 80  
Post, Margje, viii, 242, 252, 254, 256, 261  
Potebnja, Aleksandr, 243  
Powersland, Peter F., 108  
Preston, Dennis R., 78, 80–82, 205, 206, 208, 210, 211, 226, 227, 301, 321, 341  
Prilutski, Noah, 126  
Prokić, Jelena, 391  
Pröll, Simon, 70  
Purschke, Christoph, 321
- Qi, Cuihong, 84  
Qi, Peng, 184

- Qualtrics, 88
- Rabanus, Stefan, 78, 79
- Rabin, Chaim, 221
- Ran, Qibin, 364
- Rand, William M., 377
- Ransmayr, Jutta, 288, 301, 304, 305, 308
- Ranzmaier, Irene, 294
- Ratliff, Martha, 409, 411
- Rawlins, Jacob D., 288
- Reiffenstein, Ingo, 60
- Reker, Siemon, 86
- Retti, Gregor, 320, 341
- Rickford, John R., 4, 6, 17, 36, 37
- Roeper, Thomas, 13
- Roth, Tobias, 320
- Rozanova, Nina, 249
- Rubin, Aaron D., 219
- Rubin, Israel, 129
- Ruck, Julia, 301, 309
- Ryan, Camille L., 8, 33
- Sadock, Benjamin, viii, 125, 131, 144
- Saltveit, Laurits, 180, 181
- Sammon, John W., 385
- Sands, Bonny, 409
- Sankoff, David, 9, 10, 35
- Sankoff, Gillian, 156, 160, 167
- Savinov, Dmitrij, 246
- Schäfer, Lea, 130
- Schallert, Oliver, 182
- Schatz, Josef, 328, 330, 339, 340
- Scherre, Maria M. P., 156
- Scheuringer, Hermann, 298
- Scheutz, Hannes, 60
- Schilling-Estes, Natalie, 36
- Schmid, Tanja, 181
- Schmidt, Jürgen E., 336, 345
- Schnabel, Michael, 320
- Schneider, Edgar W., 290, 303, 306
- Schröder, Ingrid, 174
- Schröder, Saskia, 321, 352
- Schwaiger, Alois, 64
- Schweizerisches Idiotikon, 339
- Séguy, Jean, 361
- Sekeres, Hedwig G., vii
- Shao, Huijun, 375
- Shi, Rihai, 375
- Shockey, Linda, 119
- Shu, Hong, 84
- Sibata, Takesi, 80, 270
- Siewert, Janine, viii, 182
- Silverstein, Michael, 231
- Simpson, Adrian P., 249
- Smits, Tom, 177
- Sneller, Betsy, 165
- Solnit, David B., 404
- Spitzmüller, Jürgen, 226, 230
- Spolsky, Bernard, 221
- Spruit, Marco René, 186
- Srithonrat, Potjanee, 407
- Stanford, James N., 82, 119, 364
- Stanley, Joseph A., 132
- Stellmacher, Dieter, 174, 175
- Stevenson, Jonathan, vii, 4, 9, 10, 15, 16, 34, 36
- Stöckle, Philipp, 320
- Strelluf, Christopher, 34, 36
- Sugitō, Miyoko, 280
- Sundgren, Eva, 156
- Sung, Ho Wang Matthew, viii, 375, 386, 391
- Švorc, Peter, 130
- Szeto, Pui Yiu, 416
- Szmrecsanyi, Benedikt, 186
- Tabain, Marija, 244, 246, 247

## *Name index*

- Tagliamonte, Sali A., 109, 110, 157, 164, 165
- Takemura, Akiko, viii, 273, 276, 280
- Tamir, Christine, 36
- Tan, Yuanxiong, 373, 375
- Tang, Chaoju, 361, 363, 364, 366–370, 383
- Tenbrink, Thora, 85
- Terry, J. Michael, 4, 11–13
- Thies, Heinrich, 175, 176, 178
- Thomas, Erik R., 31
- Thurgood, Graham, 412
- Tobler, Waldo, 376
- Tomaschek, Fabian, 156
- Trudgill, Peter, 75, 79, 95, 107, 108, 118, 280, 288, 289, 291, 302, 362
- USC Shoah Foundation Visual History Archive, 131
- Uwano, Zendō, 271
- van Bree, Cor, 86, 178, 181
- van der Vliet, Goaitsen, 182
- van Gemert, Ilse, 82
- van Hout, Roeland, 288
- van Noord, Gertjan, 183
- Van Rooy, Raf, 288
- Vandenbussche, Wim, 288
- Vardøy, Benedikte Fjellanger, 242, 252
- Verbickaja, Ljudmila, 246, 249, 262
- Versloot, Arjen, 85
- von Polenz, Peter, 302
- Vysotskij, Sergej, 244, 245, 249, 251, 257, 258, 260–262
- Wagner, Suzanne E., 156, 165
- Waldner, Gernot, 291
- Walker, Kyle, 42
- Walmsley, Dennis James, 84
- Watts, Akiah, 165
- Watts, Richard, 288, 289
- Weenink, David, 252, 253
- Wei, S., 373
- Weijnen, Antonius Angelus, 80
- Wein, Martin J., 222
- Weinberg, Werner, 220
- Weinreich, Max, 125–127, 221, 291
- Weinreich, Uriel, 125, 127–130
- Weinryb, Bernard D., 127, 128
- Weldon, Tracey, 31, 32, 34
- Wells, John C., 368
- Weng, Zewen, 375
- Wexler, Paul, 221
- White, Rodney, 80
- Wichmann, Søren, 364
- Wickham, Hadley, 42, 128
- Wieling, Martijn, 85, 186, 387
- Wiener, Seth, 427
- Wiese, Heike, 229
- Wiesinger, Peter, 327, 328, 346
- Willis, David, 10, 13, 36, 39, 50
- Wilson, August, 11
- Wisser, Wilhelm, 178, 181
- Wittgenstein, Ludwig, 292, 308
- Wojcik, Stefan, 9
- Wolf, Nobert, 309
- Wolf, Norbert, 294
- Wolfram, Walt, 3–6, 31, 32, 36, 294
- Wolk, Christoph, 186
- Wright, Laura, 288
- Wu, Wei, 371, 373
- Wulff, Kurt, 409
- Xie, Jianyou, 375
- Yaeger-Dror, Malcah, 156

- Yang, Cathryn, 361, 363–367, 383  
Yang, Shiwen, 375  
Yang, Tongyin, 412  
Yip, Moira, 361, 363, 389  
Yoshida, Norio, 270, 271  
You, R., 368  
Yuan, Jiahua, 373  
Yue-Hashimoto, Anne, 372  
Yurayong, Chingduang, viii, 416
- Zampieri, Marcos, 186  
Zemskaja, Elena, 249  
Zhan, Bohui, 362, 364, 368, 372, 373,  
    375  
Zhang, Junru, 373, 409  
Zhong, Ziqiang, 375  
Zhu, Xiaonong, 409  
Zlatoustova, Ljubov', 243, 244





# (Dia)lects in the 21st century

This book offers an in-depth exploration of contemporary issues and methodologies in the fields of dialectology and sociolinguistics. Readers will find a diverse collection of studies that examine how language varies and changes across different regions, communities, and social contexts. The book covers a wide range of languages, including German, English, Yiddish, Russian, and Japanese, providing a global perspective on linguistic diversity.

Key themes include the use of modern data sources, such as social media, to study language patterns and the impact of digital communication on regional dialects. The book also addresses the dynamics of language contact in expatriate communities, revealing how speakers adapt and merge linguistic features from different dialects.

Several chapters focus on the evolution of dialectological research, offering critiques and new approaches to studying regional language variations. Readers will also encounter innovative methods, such as cognitive geography, which uses mental representations of space to understand dialect variation, and tone distance measures, which are crucial for studying tonal languages.

Additionally, the book presents case studies on how non-experts perceive and categorize dialects, providing insights into the public's understanding of linguistic diversity. It also tackles challenges in selecting dialect speakers for research, especially in urban environments, where traditional criteria may no longer apply.

Overall, this book is a valuable resource for linguists, researchers, and anyone interested in the complex and ever-changing landscape of human language. It highlights the importance of adapting research methods to keep pace with the evolving nature of language and offers fresh perspectives on how we study and understand dialects and language variation.