

Statistics in Corpus Linguistics

Do you use language corpora in your research or study, but find that you struggle with statistics? This practical introduction will equip you to understand the key principles of statistical thinking and apply these concepts to your own research, without the need for prior statistical knowledge. The book gives step-by-step guidance through the process of statistical analysis and provides multiple examples of how statistical techniques can be used to analyse and visualize linguistic data. It also includes a useful selection of discussion questions and exercises which you can use to check your understanding. The book comes with a companion website, which provides additional materials (including answers to exercises, datasets, advanced materials, teaching slides etc.) and Lancaster Stats Tools online (<http://corpora.lancs.ac.uk/stats>), a free click-and-analyse statistical tool for easy calculation of the statistical measures discussed in the book.

VACLAV BREZINA is a senior lecturer at the Department of Linguistics and English Language, Lancaster University. He specializes in corpus linguistics, statistics and applied linguistics, and has designed a number of different tools for corpus analysis.

Statistics in Corpus Linguistics

A Practical Guide

VACLAV BREZINA

Lancaster University



CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of
education, learning, and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/9781107125704

DOI: 10.1017/9781316410899

© Vaclav Brezina 2018

This publication is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without the written
permission of Cambridge University Press.

First published 2018

Printed in the United Kingdom by TJ International Ltd. Padstow Cornwall

A catalogue record for this publication is available from the British Library.

Library of Congress Cataloging-in-Publication Data

Names: Brezina, Vaclav, 1979– author.

Title: Statistics in corpus linguistics : a practical guide / Vaclav Brezina, Lancaster University.

Description: Cambridge ; New York : Cambridge University Press, 2018. |

Includes bibliographical references and index.

Identifiers: LCCN 2018007010 | ISBN 9781107125704 (alk. paper)

Subjects: LCSH: Corpora (Linguistics) | Linguistics – Statistical methods.

Classification: LCC P128.C68 B76 2018 | DDC 410.1/88–dc23

LC record available at <https://lccn.loc.gov/2018007010>

ISBN 978-1-107-12570-4 Hardback

ISBN 978-1-107-56524-1 Paperback

Cambridge University Press has no responsibility for the persistence or accuracy of
URLs for external or third-party internet websites referred to in this publication
and does not guarantee that any content on such websites is, or will remain,
accurate or appropriate.

To Anna, Olinka and Jan, who share my passion for numbers.

Contents

<i>List of Figures</i>	<i>page</i> x
<i>List of Tables</i>	xiv
<i>About This Book</i>	xvii
<i>Acknowledgements</i>	xix
1 Introduction: Statistics Meets Corpus Linguistics	1
1.1 What Is This Chapter About?	1
1.2 What Is Statistics? Science, Corpus Linguistics and Statistics	1
1.3 Basic Statistical Terminology	5
1.4 Building of Corpora and Research Design	15
1.5 Exploring Data and Data Visualization	22
1.6 Application and Further Examples: Do Fiction Writers Use More Adjectives than Academics?	30
1.7 Exercises	32
Things to Remember	36
Advanced Reading	36
2 Vocabulary: Frequency, Dispersion and Diversity	38
2.1 What Is This Chapter About?	38
2.2 Tokens, Types, Lemmas and Lexemes	38
2.3 Words in a Frequency List	42
2.4 The Whelk Problem: Dispersion	46
2.5 Which Words Are Important? Average Reduced Frequency	53
2.6 Lexical Diversity: Type/Token Ratio (TTR), STTR and MATTR	57
2.7 Application and Further Examples: Do the British Talk about Weather All the Time?	59
2.8 Exercises	62
Things to Remember	64
Advanced Reading	65
3 Semantics and Discourse: Collocations, Keywords and Reliability of Manual Coding	66
3.1 What Is This Chapter About?	66
3.2 Collocations and Association Measures	66
3.3 Collocation Graphs and Networks: Exploring Cross-associations	75
3.4 Keywords and Lockwords	79
3.5 Inter-rater Agreement Measures	87

3.6 Application and Further Examples: What Do Readers of British Newspapers Think about Immigration?	92
3.7 Exercises	96
Things to Remember	100
Advanced Reading	101
4 Lexico-grammar: From Simple Counts to Complex Models	102
4.1 What Is This Chapter About?	102
4.2 Analysing a Lexico-grammatical Feature	103
4.3 Cross-tabulation, Percentages and Chi-squared Test	108
4.4 Logistic Regression	117
4.5 Application: <i>That</i> or <i>Which</i> ?	130
4.6 Exercises	134
Things to Remember	137
Advanced Reading	138
5 Register Variation: Correlation, Clusters and Factors	139
5.1 What Is This Chapter About?	139
5.2 Relationships between Variables: Correlations	139
5.3 Classification: Hierarchical Agglomerative Cluster Analysis	151
5.4 Multidimensional Analysis (MD)	160
5.5 Application: Registers in New Zealand English	170
5.6 Exercises	177
Things to Remember	181
Advanced Reading	182
6 Sociolinguistics and Stylistics: Individual and Social Variation	183
6.1 What Is This Chapter About?	183
6.2 Individual Style and Social Variation: Where Does a Sociolinguistic Variable Start?	183
6.3 Group Comparison: T-Test, ANOVA, Mann–Whitney <i>U</i> Test, Kruskal–Wallis Test	186
6.4 Individual Style: Correspondence Analysis	199
6.5 Linguistic Context: Mixed-Effects Models	207
6.6 Application: Who Is This Person from the White House?	211
6.7 Exercises	215
Things to Remember	217
Advanced Reading	218
7 Change over Time: Working Diachronic Data	219
7.1 What Is This Chapter About?	219
7.2 Time as a Variable: Measuring and Visualizing Time	219

7.3 Finding and Interpreting Differences: Percentage Change and the Bootstrap Test	229
7.4 Grouping Time Periods: Neighbouring Cluster Analysis	235
7.5 Modelling Changes in Discourse: Peaks and Troughs and UFA	241
7.6 Application: Colours in the Seventeenth Century	247
7.7 Exercises	251
Things to Remember	255
Advanced Reading	256
8 Bringing Everything Together: Ten Principles of Statistical Thinking, Meta-analysis and Effect Sizes 257	
8.1 What Is This Chapter About?	257
8.2 Ten Principles of Statistical Thinking	257
8.3 Meta-analysis: Statistical Synthesis of Research Results	267
8.4 Effect Sizes: A Guide for Meaningful Use	275
8.5 Exercises	280
Things to Remember	282
Advanced Reading	282
<i>Final Remarks</i>	283
<i>References</i>	285
<i>Index</i>	294

Figures

1.1	The relationship between the relative frequency of adjectives and verbs	page 4
1.2	Process of statistical analysis	6
1.3	Example of a dataset	7
1.4	The distribution of the first-person pronoun in the <i>Trinity Lancaster Corpus</i>	9
1.5	Standard normal distribution	9
1.6	Dispersion of adjective frequencies in 11 corpus files	11
1.7	Confidence intervals: two situations	14
1.8	Research designs in corpus linguistics	21
1.9	Bar chart: variable x in three corpora	24
1.10	Boxplot: variable x in three corpora	24
1.11	Error bars: variable x in three corpora	25
1.12	Histogram: the definite article in BE06	26
1.13	Histogram: the f-word in BNC64	26
1.14	Scatterplot: <i>the</i> and <i>I</i> in BNC64	27
1.15	Scatterplot: <i>the</i> , <i>I</i> and <i>you</i> in BNC64	28
1.16	Top ten places connected with ‘going’ or ‘travelling’ in the BNC	28
1.17	Other types of visualizations	29
1.18	The use of adjectives by fiction and academic writers: boxplot	31
1.19	The use of adjectives by fiction and academic writers: error bars	32
1.20	Great Britain: main island	33
2.1	Distribution of word frequencies in the BNC	45
2.2	Example corpus: calculation of SD	49
2.3	Distribution of words w_1 and w_2	55
3.1	Frequency and exclusivity scale	74
3.2	Collocation graph: ‘love’ in BE06 (10a – log Dice (7), L3–R3, C5–NC5)	76
3.3	Collocation networks: concept demonstration	77
3.4	Third-order collocates of time in LOB (3a–MI(5), R5–L5, C4–NC4; no filter applied)	78
3.5	Collocation network of ‘university’ based on BE06 (3b–MI(3), L5–R5, C8–NC8)	79

3.6	Collocation networks around ‘immigrants’ in the <i>Guardian</i> (3a–MI(6), R5–L5, C10–NC10; no filter applied)	94
3.7	Collocation networks around ‘immigrants’ in the <i>Daily Mail</i> (3a–MI(6), R5–L5, C20–NC20; no filter applied)	94
3.8	Selected collocation networks	97
4.1	The definite and indefinite articles in BNC subcorpora	104
4.2	<i>The</i> vs <i>a(n)</i> dataset: linguistic feature design (an excerpt)	105
4.3	A mosaic plot: article type by contextual determination	109
4.4	Logistic regression: a basic schema	119
4.5	Article use in English: a dataset (an excerpt)	122
4.6	A sentence from this book corrected for ‘grammar’	130
4.7	Visualization of the relationship between <i>which</i> and <i>that</i> and a separator	132
4.8	<i>Must</i> , <i>have to</i> and <i>need to</i> in British English (BE06)	135
5.1	Nouns and adjectives in BE06	140
5.2	Verbs and adjectives in BE06	140
5.3	Pronouns and coordinators in BE06	141
5.4	Correlation: five data points	143
5.5	Correlation: covariance	143
5.6	Statistically significant ($p < 0.05$) Pearson’s correlations in relation to the number of observations	145
5.7	Multi-panel scatterplot: nouns, adjectives, verbs, pronouns and coordinators	149
5.8	Correlation matrix: nouns, adjectives, verbs, pronouns and coordinators	150
5.9	Colour terms in the BNC	152
5.10	Creating clusters: Steps 1–4	155
5.11	Creating clusters: final result	156
5.12	Colour terms: a tree plot (dendrogram) – z-score ₂ normalized, Euclidean distance, SLINK method	156
5.13	Tree plot: SLINK method	157
5.14	Tree plot: CLINK method	157
5.15	Tree plot: average linkage method	158
5.16	Tree plot: Ward’s method	159
5.17	A dataset for multidimensional analysis (a small extract)	164
5.18	Data reduction: ten variables into two factors	165
5.19	Promax factor rotation	166
5.20	Factor extraction: scree plot	167
5.21	Mean scores of registers placed on Dimension 1: Involved vs Informational	169
5.22	Correlation matrix: 44 variables	173
5.23	Correlation between mean word length and contractions: register clusters	174

5.24 Cluster plot: registers in New Zealand English	175
5.25 Dimension 1: New Zealand English – full MD analysis	177
5.26 Dimension 2: New Zealand English – full MD analysis	177
5.27 Relationship between mean word length (number of characters) and mean sentence length (number of words) in BNC	178
5.28 Relationship between the use of the past and the present tense in BE06	178
5.29 Relationship between the use of adjectives and colour terms in BE06	179
5.30 Relationship between text length (tokens) and type–token ratio (TTR) in BNC	179
5.31 Dimension 3	181
5.32 Dimension 4	181
6.1 Distribution of personal pronouns in BNC64 female speakers	188
6.2 ANOVA calculation: between-group variance (top), within-group variance (bottom)	193
6.3 Dataset from BNC64 – relative frequencies and ranks: use of personal pronouns	195
6.4 Distribution of <i>ain't</i> in BNC64 speakers: social-class effect	198
6.5 <i>Ain't</i> in BNC64: 95% CI	198
6.6 A correspondence plot: word classes in the speech of individual speakers	201
6.7 Speaker (row) profiles: Euclidean distance	204
6.8 Speaker (row) profiles: chi-squared distance	205
6.9 Sociolinguistic dataset: internal and external factors (an excerpt)	208
6.10 Mixed-effects models: output	209
6.11 Correspondence analysis: use of word classes by White House press secretaries	214
6.12 Correspondence analysis: use of epistemic markers in BNC64	216
7.1 Modals in the Brown family corpora	220
7.2 Modals in the Brown family corpora: an alternative interpretation	223
7.3 Google n-gram viewer: ‘man’ and ‘woman’	224
7.4 Modals in the Brown family corpora: original (top) and rescaled (bottom)	225
7.5 Modals in British English: (a) boxplots; (b) 95% CI error bars	227
7.6 Candlestick plot: the development of individual modals 1931–2006	228
7.7 Bootstrapping: demonstration of the concept	231
7.8 Example of a dataset for the bootstrap test: <i>its</i> in EEBO	233
7.9 Data points over time: an invented example	235

7.10	Two clustering principles: (a) hierarchical agglomerative clustering; (b) variability-based neighbour clustering	237
7.11	Dendograms: (a) hierarchical agglomerative clustering; (b) variability-based neighbour clustering	238
7.12	Dendrogram: use of the possessive pronoun <i>its</i> in the seventeenth century	239
7.13	Scree plot: use of the possessive pronoun <i>its</i> in the seventeenth century	240
7.14	Resulting peaks and troughs graphs: settings as indicated	244
7.15	Results of UFA for <i>war</i> 1940–2009 (3a–MI(3), L5–R5, C10relative–NC10relative; AC1)	246
7.16	Frequency of colour terms in the seventeenth century	248
7.17	Candlestick plot: colours in the seventeenth century	249
7.18	Results of UFA for <i>red</i> 1600–99 (3a–MI(3), L5–R5, C10relative–NC10relative; AC1)	250
7.19	VNC: <i>red</i> in the seventeenth century	251
7.20	Number of tweets related to an episode of the UK <i>X-Factor</i> (16/11/2014, 7–11pm)	252
7.21	Development of frequencies of <i>handsome</i> , <i>pretty</i> and <i>beautiful</i> followed by a male (M) or female (F) person in the seventeenth century	252
7.22	Development of frequencies of the possessive pronoun <i>its</i> in the seventeenth century	253
7.23	Four frequency change scenarios	254
7.24	<i>Handsome</i> in the seventeenth century	254
7.25	<i>Pretty</i> in the seventeenth century	255
8.1	Overview of genres in BE06 (Baker 2009)	260
8.2	Past tense in different written genres of BE06	260
8.3	Past tense (a) and present tense (b) in different written genres of BE06: boxplot rendition	265
8.4	Finding the Globe	268
8.5	Forest plot: meta-analysis of four studies	274
8.6	Comparison of two subcorpora	278
8.7	Forest plot: example 1	281
8.8	Forest plot: example 2	281

Tables

1.1	The effect size r and its standard interpretation	page 14
1.2	Brown family sampling frame	16
1.3	Frequencies of selected words and expressions in three English corpora	19
1.4	Different levels of analysis in corpus linguistics	20
1.5	Subcorpora in mini-research	30
2.1	Type, lemma and lexeme: advantages and disadvantages	41
2.2	Top ten words in the BNC	42
2.3	Example corpus: one million tokens	47
2.4	Calculation of DP with the example corpus	53
2.5	BE06	60
2.6	Weather-related lemmas in BE06	61
2.7	Ranks of weather-related lemmas in BE06	62
2.8	BNC: distribution of four selected words	64
3.1	Observed frequencies	70
3.2	Expected frequencies: random occurrence baseline	71
3.3	Association measures: overview	72
3.4	Ranking of collocates of ‘new’ in BE06 (L3–R3)	73
3.5	Collocation parameters notation (CPN)	75
3.6	AmE06: American English keywords	80
3.7	Decisions about keywords: BASIC options	81
3.8	Comparison of selected lexical items in BE06 and AmE06	83
3.9	American English keywords: different keyword identification procedures	86
3.10	BE06: selected concordances for ‘religion’	88
3.11	Double coding: concordances from the ‘Think about’ task	89
3.12	Overview of inter-rater agreement measures	91
3.13	Keywords	93
3.14	Evaluations of ‘immigrant(s)’ in the GU and DM corpora	95
3.15	Collocates of <i>issue</i> in BE06	96
3.16	Keywords	98
3.17	Examples for rating	99
4.1	Examples of lexico-grammatical variables with a grammatical frame	107
4.2	Cross-tabulation: article type by contextual determination	109
4.3	Percentage options in cross-tabulation	111

4.4	Strong (semi-)modals in different genres of British and American English: cross-tabulation	112
4.5	Expected frequencies: article type by contextual determination	114
4.6	Interpretation of Cramer's <i>V</i>	115
4.7	Probabilities: article type by contextual determination	116
4.8	Probabilities: article type by noun type	121
4.9	Models: an overview	124
4.10	A part of the logistic regression output: large standard errors	124
4.11	Cross-tabulation: separator use with <i>which</i> and <i>that</i> relativizers	131
4.12	<i>Which</i> and <i>that</i> in different contextual situations: cross-tabulation	133
4.13	<i>Which</i> or <i>that</i> : logistic regression estimates	134
4.14	Appropriate research design	134
5.1	Ranks of nouns and adjectives in five texts from BE06	147
5.2	Correlation table (Pearson's correlations): nouns, adjectives, verbs, pronouns and coordinators	148
5.3	The full set of Biber's (1988) features based on Conrad & Biber (2001: 18–19)	162
5.4	Factor 1: loadings of individual variables	168
5.5	Registers in ICE-NZ	171
5.6	Results of factor analysis of NZ English: factor loadings	176
5.7	Results of factor analysis of NZ English: factor loadings of Factors 3 and 4	180
6.1	Cross-tabulation table: word classes in the speech of individual speakers	201
6.2	Cross-tabulation table: verbs and articles in the speech of two speakers	202
6.3	Row (speaker) profiles	203
6.4	WH corpus	213
6.5	Swearing and gender: BNC64	217
7.1	Comparison of two periods in the EEBO corpus: Commonwealth & Protectorate and Restoration	230
7.2	Comparison of two periods in the EEBO corpus: results of the bootstrap test	233
7.3	Final evaluation of the results: <i>its</i> , <i>must</i> and <i>pestilence</i>	234
7.4	Relative frequency (per million) of the possessive pronoun <i>its</i> in the seventeenth century	239
7.5	Relative frequencies (per million) of <i>war</i>	243
7.6	Differences between relative frequencies of <i>war</i>	243
7.7	Log transformed relative frequency (per million) of <i>war</i>	243
7.8	Collocate profiles of <i>war</i>	245
8.1	The use of the past and the present tense in different registers (original dataset)	258

8.2	The use of the past and the present tense in different registers (research report)	258
8.3	Examples of use of the past tense in academic writing and mystery fiction	261
8.4	Examples of texts: academic writing and mystery fiction	263
8.5	Overview of answers	267
8.6	Studies reviewed for the meta-analysis	272
8.7	Input data for a simple meta-analysis	273
8.8	Effect size measures introduced in this book	276
8.9	Effect size transformation and extrapolation	277
8.10	Effect size: standard interpretation	278
8.11	Effect size measures: BNC validation	279
8.12	Effect size transformations	281

About This Book

What Is This Book About?

This book is a practical introduction to statistical procedures in corpus linguistics, a discipline that uses computers to analyse language, organized according to linguistic topics. These range from vocabulary and grammar to sociolinguistics, discourse analysis and historical investigations of language. The book offers an overview of the state-of-the-art methodologies of language analysis using corpora and introduces new techniques that have not previously been used in the field. No prior knowledge of statistics is assumed; instead, all necessary concepts and methods are explained in non-technical language. In addition, all procedures described in the book can be easily carried out using Lancaster Stats Tools online (see ‘How Should You Use This Book?’ below). Throughout the book, many examples (case studies) of the application of corpus statistics are provided and standard reporting of statistics is shown. The emphasis of the book on the practical aspects of statistical analysis of language is also reflected in its focus on research design and the implications of different ‘shapes’ of data for statistical analysis – for this reason, the companion website offers complete datasets used in this book for easy replication of the analyses. Corpus linguistics is an extremely versatile methodology of language analysis applicable in a wide range of contexts, in linguistics, social science, digital humanities and elsewhere – the book thus aims to facilitate meaningful use of corpora for as wide a range of users as possible.

Who Is This Book For?

The book is intended for anyone interested in corpus linguistics and quantitative analysis of language. This includes students and researchers in the field of linguistics, sociology, history, psychology, education etc. The main goal of the book is to help readers understand key principles of statistical thinking in order to be able to make informed decisions about the applications of particular statistical techniques. To facilitate this, in addition to the expository parts, the book also includes discussion questions ('Think about...') and exercises which the readers can use to better engage with the material and to check their comprehension of the subject matter; answers to the exercises are provided at the companion website (<http://corpora.lancs.ac.uk/stats/materials.php>).

How Should You Use This Book?

The book reflects the needs of students and researchers who are looking for a practical guidebook on corpus statistics, which is grounded in the current literature in the field and reflects the best practice. The book can be used as a course book or for independent study. After reviewing general statistical principles in Chapter 1, readers can follow their own path through the book according to the linguistic topics of their interest. Statistical techniques introduced in the book are cross-referenced and included in the Index at the end of the book.

The book comes with a companion website – Lancaster Stats Tools online – (<http://corpora.lancs.ac.uk/stats>), which not only provides additional examples, datasets, video tutorials and PowerPoint slides but also, more importantly, includes easy-to-use tools for calculating statistics and producing graphs discussed in the book. In fact, all procedures described in the book can be performed by the reader using Lancaster Stats Tools online. Readers thus don't need to rely on commercial statistical packages such as *IBM SPSS* which are not easily affordable for users without institutional subscriptions. Neither will readers be required to learn the complex syntax of free statistical packages such as *R*. Instead, Lancaster Stats Tools online offers access to powerful statistical tools through a simple user interface, into which the data can be directly copy-pasted from a spreadsheet (e.g. Excel or Calc). I believe that statistics shouldn't be a hurdle in our research – computers can and should do all the hard work of number crunching for us; statistics, instead, can be used as a very effective analytical tool – all that is needed is to understand the basic principles of statistical thinking and their application to language analysis. Let's explore them together!

Acknowledgements

I would like to thank Tony McEnery and Dana Gablasova for their continued support and encouragement throughout the process of writing this book as well as for their detailed comments on each of the chapters. The work has also greatly benefited from the insightful points raised by Gabriele Pallotti who has read a large part of the manuscript. In addition, I would also like to thank the following colleagues for their helpful comments on different parts of the manuscript: Peter Diggle, Michael Gauthier, Andrew Hardie and Irene Marin Cervantes. I thank Gill Smith for her help with the formatting of the manuscript and Irene Marin Cervantes for preparing index entries. Thanks are also due to two anonymous reviewers for their very helpful and encouraging comments.

The writing of this book was supported by ESRC grants nos. EP/P001559/1 and ES/K002155/1.

1 Introduction

Statistics Meets Corpus Linguistics

1.1 What Is This Chapter About?

This chapter introduces basic principles of statistical thinking that are necessary for informed application of statistical procedures to corpus data. It starts with an explanation of the role of statistics in scientific research in general and corpus linguistics in particular. After that, more specific topics such as the creation of corpora, types of research design, basic statistical terminology, as well as data exploration and visualization are discussed. The chapter ends with a case study demonstrating the use of statistics in corpus research.

In particular, we'll be exploring answers to five questions:

- What is the role of statistics in science and corpus research?
(Section 1.2)
- What are the key terms in corpus statistics? (Section 1.3)
- How do we build and analyse corpora? (Section 1.4)
- How can we explore and visualize data? (Section 1.5)
- How can statistics be used in corpus research? (Section 1.6)

1.2 What Is Statistics? Science, Corpus Linguistics and Statistics

Think about ...

Before reading this section, think about the following questions:

1. What is science? What are the basic features of scientific enquiry?
2. Which of these statements about language are scientific statements?
 - (a) Women's speech seems in general to contain more instances of 'well', 'y'know', 'kinda', and so forth ...
 - (b) Words are easy, like the wind.
 - (c) Passives are most common by far in academic prose [compared to other registers], occurring about 18,500 times per million words.
 - (d) The faculty of language can reasonably be regarded as a 'language organ'.

(See next page for one more example)

- (e) Our results show that there were significant changes in at least one formant¹ for 10 of 11 vowel sounds and in both formants for 5 of 11 vowel sounds from the 1950s to the 1980s Christmas broadcasts . . . We conclude that the Queen no longer speaks the Queen's English of the 1950s . . .

Unlike other sources of information such as mythology, philosophy or art, **science** relies on the systematic collection of empirical data and testing of theories and hypotheses. One of the most influential theoreticians of science, Karl Popper, defined a scientific statement or theory as something that can in principle be falsified (Popper, 2005 [1935]). In other words, we can call a statement or theory scientific only if it can be tested empirically. This means that we need to collect data and evaluate if the data is consistent with our theory. If not, we can say that the available evidence contradicts the theory. When we look at the statements in question 2 of the 'Think about' task we can see that they vary considerably in whether they can be put to a test by collecting empirical evidence: clearly, statements (c) and (e) can be considered scientific.² Not only can they be empirically tested, these statements are already accompanied by empirical evidence. On the other hand, the poetic statement (b) expresses a metaphor, which despite its power would be difficult to test by collecting data. Statement (a), which is taken from Lakoff's (1975) book *Language and Woman's Place*, can be empirically tested (indeed numerous researchers have tested it), yet the author herself offers little empirical evidence in her book beyond anecdotes. Statement (d), which comes from Chomsky (2000), proposes a view of language that relies more on philosophical (pre-empirical) understanding of the human language faculty and does not necessarily seek empirical confirmation. In sum, statements about language which provide direct reference to systematically³ collected empirical evidence can be considered scientific.

Corpus linguistics is a scientific method of language analysis. It requires the analyst to provide empirical evidence in the form of data drawn from language corpora in support of any statement made about language. Another scientific requirement corpus linguists follow in principle is replicability of results. This means that researchers need to be able to confirm the findings of one study in follow-up studies (see Section 8.3). In order for the results to be replicable, corpus linguists need to make their choice of corpora and analytical techniques transparent. It is also good practice in corpus linguistics to make corpora available to other researchers who can explore the same dataset further and thus advance knowledge in the field.⁴

¹ Formant is a component of a vowel sound in phonetic research.

² Sources of statements: (a) Lakoff 1975: 53; (b) Shakespeare? 1992 [1599]: 269; (c) Biber et al. 1999: 476; (d) Chomsky 2000: 4; (e) Harrington et al. 2000: 927.

³ Empirical research can be both qualitative (descriptive and interpretational) and quantitative (using numbers). These areas of research are complementary.

⁴ Unfortunately, sometimes corpora are 'locked' behind corporate walls with unclear principles of how these corpora were constructed. This makes their use difficult for any serious scientific

In essence, corpus linguistics is a quantitative methodology; this means that corpus linguistics typically works with numbers which reflect the frequencies of words and phrases in corpora (McEnery & Hardie 2011.) For this reason, statistics is crucial for corpus linguists because it helps us work effectively with quantitative information. There are many different understandings of what statistics is. In this book, we will be working with the following definition: **statistics** is a discipline which helps us make sense of quantitative data; in other words, statistics is a ‘science of collecting and interpreting data’ (Diggle & Chetwynd 2011: vii) that can be counted, measured or quantified in some way. One of the important tools in statistics is the use of mathematical expressions – that’s why we’ll be looking at various equations in this book. Mathematical expressions help us understand complex and fuzzy reality through capturing important features of the data by means of numbers and symbols, making it possible to handle the data easily in the process of analysis.

Let us have a look at two examples that illustrate this point. First, imagine that we are interested in the number of adjectives different British fiction writers use in their texts. We might hypothesize that using more adjectives leads to more colourful descriptions in novels. We have randomly selected 11 fiction texts by different authors from the *British National Corpus* (BNC) and counted the number of adjectives in each text; this is their absolute frequency (see Section 2.3). We have then normalized the absolute frequency for comparability.⁵ In statistics, we call our 11 texts a **sample**. The following are the relative frequencies of adjectives per 10,000 words:

508, 542, 552, 553, 565, 567, 570, 599, 656, 695, 699

However, showing a long list of results is not a very efficient way of dealing with quantitative data – imagine what would happen if we had to list 100 or 1,000 different results. Instead, we can use a very simple statistical measure to summarize our findings. This measure is called the **mean** and gives us an average value which represents a whole range of values. The mean for the numbers above is 591.45.

The mean is calculated in the following way:

$$\text{mean} = \frac{\text{sum of all values}}{\text{number of cases}}$$

exploration and must lead to doubt being cast on claims made using such corpora. If corpus linguistics wants to retain its scientific status, it should not be content with statements such as ‘this feature was found in a large corpus that is, however, not available’.

⁵ Because the texts are of different length, we have taken the relative frequencies per 10,000 words to show how many adjectives on average each author uses in 10,000 words (see Section 2.3 for the explanation of relative frequency). The relative frequencies have been rounded to the nearest integer.

Applied to the dataset above:

$$\text{mean} = \frac{508 + 542 + 552 + 553 + 565 + 567 + 570 + 599 + 656 + 695 + 699}{11}$$

$$= 591.45 \quad (1.1)$$

Because the mean describes our sample, it is part of what we call **descriptive statistics**. Another example of a mathematical representation of complex linguistic reality is a line, in statistics called a **regression line** or **line of the best fit** (see Chapter 4 for an explanation of regression models). Imagine that we are interested in whether the authors that use more adjectives also use more verbs. We can list the frequencies of verbs⁶ just below the frequencies of adjectives to see whether there is any relationship between these two linguistic features:

508, 542, 552, 553, 565, 567, 570, 599, 656, 695, 699
 2339, 2089, 2056, 2276, 2233, 2056, 2241, 1995, 2043, 1976, 2062

However, a better way of finding out whether there is a relationship between the use of adjectives and verbs is to display these numbers in a graph (see Section 1.5 on how to create graphs).

The graph in Figure 1.1 shows a clear tendency marked by the regression line. The regression line points to the fact that the number of verbs and adjectives in the sample is in an inversely proportional relationship – the more adjectives

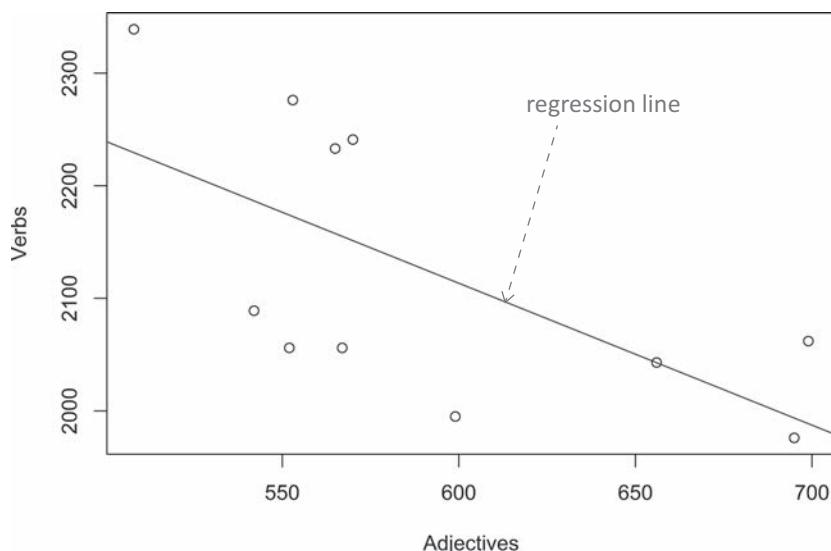


Figure 1.1 *The relationship between the relative frequency of adjectives and verbs*

⁶ Again, shown as relative frequencies per 10,000 words rounded to the nearest integer.

authors use, the fewer the verbs they employ and vice versa. The line is plotted in such a way as to find the best fit for all the individual data points (marked as dots in the graph). It is by sheer coincidence that one of the points actually lies exactly on the line; often the line does not go through any of the actual data points because it is a mathematical abstraction representing the dataset as a whole. The purpose of this mathematical model of reality is to tell us something interesting about the data that we wouldn't necessarily notice if we looked at the individual data points in isolation. These two examples demonstrate the main point of statistical thinking that will appear in various forms throughout the book: statistics in corpus linguistics is about mathematical modelling of a complex linguistic reality. It can help us discover and elucidate patterns and tendencies in the data that might otherwise remain hidden.

1.3 Basic Statistical Terminology

Think about ...

Before reading this section, think about the meaning of the following terms. Have you heard them before? If so, in what context? Would you be able to define them?

- assumption
- case
- confidence interval
- dataset
- dispersion
- distribution
- effect size
- normal distribution
- null hypothesis
- outlier
- p-value
- robust
- rogue value
- statistical measure
- statistical test
- standard deviation
- variable

The following is an overview of basic statistical terminology used in this book. It includes key terms with examples from corpus research and is ordered from basic concepts to more complex ones which rely on the understanding of the previous terminology. Mastering these terms will make reading of the rest of the book, and many papers in corpus linguistics, much easier.

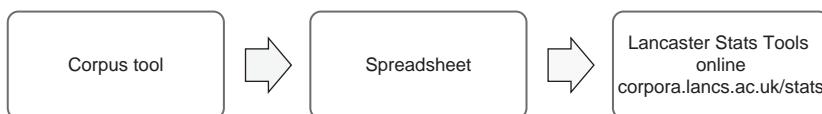


Figure 1.2 *Process of statistical analysis*

Corpus (pl. corpora) is a specific form of linguistic data. It is a collection of written texts or transcripts of spoken language that can be searched by a computer using specialized software. A corpus usually represents a **sample** of language, i.e. a (small) subset of the language production of interest; in some limited cases of very specialized corpora, a corpus can include the whole **population**, i.e. all language of interest to the researcher (see Section 1.4). The software that is used to search a corpus usually implements basic types of statistical analysis such as the statistical identification of collocations and keywords (see Chapter 3). For more sophisticated statistical analyses, however, we usually need to use appropriate statistical packages. This book uses Lancaster Stats Tools online, free statistical tools available from the companion website. Figure 1.2 outlines the process of analysis with Lancaster Stats Tools online.

Note that preparing the spreadsheet in the appropriate format is as important as the statistical analysis that follows. The book offers many examples of datasets based on different corpora, which are suitable for different types of analysis. It is always useful to compare your data to the model examples provided (full datasets are available from the companion website) to see if your data is in the appropriate format.

Dataset is a series of corpus-based findings that can be statistically analysed. It is a systematic collection of individual results that can be stored in the form of a table in a spreadsheet program (e.g. Excel, Calc etc.), each line representing an individual **data point** or **case** and each column representing a separate **variable**. Figure 1.3 provides an example of a dataset with five variables and multiple cases, each case representing one speaker. Note that example datasets used in this book are available at the companion website. It is important to study them for the particular ‘shape’ of data that lends itself to particular types of statistical analyses.

Variable, as the name suggests, is something that can vary and take on different values. For example, speaker’s age is a variable that can take on different **values** from about one year (when children typically learn their first words) to over 100. Much corpus research can be characterized as searching for variables in corpora and analysing the relationship between them. An important distinction needs to be made between linguistic variables and explanatory variables. **Linguistic variables** capture frequencies of linguistic features of interest in the corpus. **Explanatory variables** (sometimes called ‘independent variables’) capture contexts in which the linguistic features appear. For instance, an

		explanatory variables		linguistic variables		
1	speaker_id	gender	proficiency	I	you	pers_pronouns_all
2	6_SP_51	0	1	38.75969	9.302326	80.62015504
3	6_SL_7	0	1	33.46856	19.26978	100.4056795
4	8_ME_24	1	2	39.10112	38.65169	129.8876404
5	8_IT_28	1	2	51.98181	11.04613	122.8070175
6	8_IT_14	0	2	33.41584	8.663366	108.9108911
7	IT_65	1	3	37.127	19.43095	100.9715475
8	7_CH_17	0	2	58.64198	23.91975	100.308642
9	7_ME_6	1	2	42.48573	10.14585	119.2136969
10	6_CH_15	0	1	56.12245	22.95918	145.4081633
11	IT_54	1	3	25.81369	19.01969	101.010101
12	6_ME_2	1	1	34.90401	33.15881	108.2024433
13	6_CH_25	1	1	47.82147	11.68969	145.5897981
14	CH_6	1	3	52.44601	25.1212	121.6394888
15	6_IN_3	1	1	29.8359	26.74807	131.6872428

gender: 0... male, 1... female; English proficiency: 1...pre-intermediate, 2...intermediate, 3... advanced

Figure 1.3 Example of a dataset

explanatory variable can be the genre/register or date of publication of a text as well as speaker's age, gender and language proficiency, to name only a few. The dataset in Figure 1.3, which comes from the *Trinity Lancaster Corpus* of spoken L2 production (Gablasova et al. 2017), contains two explanatory (gender and language proficiency) and three linguistic variables (relative frequencies of *I*, *you* and all personal pronouns together).

Variables (both linguistic and explanatory) can be either nominal, ordinal or scale variables. A **nominal variable** has values that represent different categories into which the cases in a dataset can be grouped; there is no order or hierarchy between the categories. Speaker's gender is an example of a nominal variable because we can assign speakers in the dataset to one of two groups: (1) male speakers and (2) female speakers. There is no hierarchy in this classification. For convenience, we often use numbers to indicate the group membership. In the dataset in Figure 1.3, 0 stands for 'male speaker' and 1 for 'female speaker' but these numbers have no inherent value; they are just a shorthand for longer labels. We could just as well have used 1 (or any unique number) for indicating the male speakers and 0 (or any unique number) for indicating the female speakers. An **ordinal variable** is similar to the nominal variable in that it groups cases into distinct categories; the categories, however, can be ordered according to some inherent hierarchy. For example, speaker's proficiency in a foreign language is an ordinal variable because we can rank speakers according to their proficiency from beginners to advanced speakers. In the dataset in Figure 1.3, 1 indicates a lower proficiency than 2 and 2 indicates a lower proficiency than 3. Finally, a **scale variable** is a quantitative variable because it can take on any value on a scale showing the quantity of a particular feature; this also means that values taken on such

scales can be added, subtracted, multiplied and divided, because they represent measurable quantities, not just rank orders.⁷ In the case of linguistic variables, the scale shows the relative frequencies of a linguistic feature in different texts, speakers or subcorpora in a corpus. For example, the numbers indicating the relative frequencies per 1,000 words of the first-person pronoun *I* in Figure 1.3 are values of a scale variable. In fact, all three linguistic variables in the dataset in Figure 1.3 are of this type.

The dataset in Figure 1.3 can be used to explore different types of research questions. For instance:

- Is there a relationship between speaker's gender (a nominal explanatory variable) and the use of personal pronouns (a scale linguistic variable)?
- Does a speaker's English proficiency (an ordinal explanatory variable) have an effect on the use of the first-person pronoun (a scale linguistic variable)?
- Is there a relationship between the use of the first-person and the second-person pronouns (both of which are scale linguistic variables)?

The **frequency distribution** of a variable provides information about the values a variable takes and their frequencies. Distributions of scale variables can be shown in a histogram (see Section 1.5). Figure 1.4 displays the distribution of the first-person pronoun from the dataset in Figure 1.3. The x-axis lists different frequency bands of the linguistic variable, in this case the first-person pronoun, per 1,000 words, while the y-axis shows the number of cases in the dataset for each frequency band. Thus, for example, the graph shows that in the corpus there were 19 texts (speakers) where the first-person pronoun was used 10 times or less per 1,000 words (this information is indicated by the first bar from the left), 88 texts where it appeared 11–20 times (second bar from the left), 214 where it occurred between 21 and 30 times (third bar from the left) etc.

As a benchmark in statistics, one of the common distributions – the **normal distribution**⁸ – is often used. The shape of the normal distribution is a symmetrical bell as shown in Figure 1.5.

Although a lot of data in the natural and social world follows the normal distribution, most linguistic data is positively skewed (↗), i.e. there is more data to the left of the distribution than the right, as we saw, for example, in Figure 1.4. Distributions in statistics are crucial because they provide an overview of the data, which indicates what statistical techniques are appropriate to use. The shape of the distribution thus plays an important role in the assumptions of different statistical procedures (see 'assumptions' below).

⁷ The label 'scale variable' subsumes interval (without a meaningful zero point) and ratio (with a meaningful zero point) variables, which are distinguished for some purposes; the distinction is not essential for corpus analysis.

⁸ 'Normal' here is used in a technical sense as a label introduced by Pearson (1920: 25) for a specific statistically important distribution; there is nothing abnormal about other types of distributions.

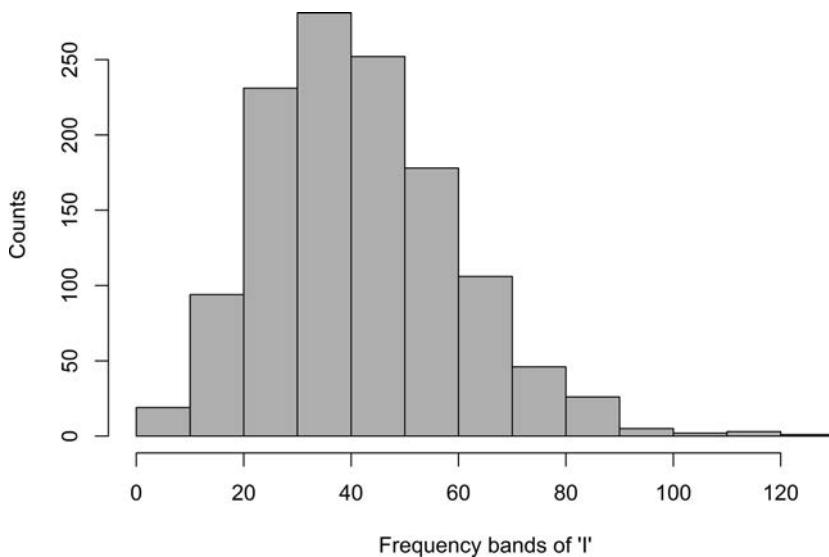


Figure 1.4 *The distribution of the first-person pronoun in the Trinity Lancaster Corpus*

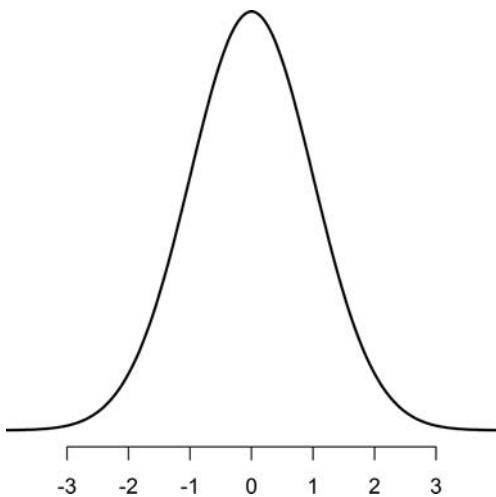


Figure 1.5 *Standard normal distribution*

Outlier or rogue value? When we look at distributions we often check for outliers. **Outliers** are extreme values, i.e. values that are very far from the other values. Section 1.5 will introduce boxplots, a useful means of identifying outliers. When we find an outlier we need to check if the outlier is a genuine value or a measurement error – a so-called **rogue value**. A rogue value can be caused, for instance, by mistyping data in a spreadsheet or by a tagging error in the corpus.

An outlier, instead, is a valid data point, which for some reason stands out from others. While outliers are not in themselves ‘errors’, they present problems for statistical models because they may obscure the general tendency (see ‘measure of central tendency’ below) in the data and the researcher must decide how to go about the analysis of data which includes outliers. If there is a good reason, outliers can be excluded (bracketed out) from part of the analysis that focuses on the central tendency in the data.

The **measure of central tendency** or ‘average’ provides one summary value for a series of values of a scale variable. It is a simple statistical model that is usefully paired with dispersion (see below) to complete the summary description of the data. Different types of average can be used. In corpus linguistics the most useful ones are: mean, median and 20% trimmed mean. **Mean (M or \bar{x})**, as we have already seen, is the sum of all values divided by the number of cases (see Section 1.2). The mean is a useful measure in distributions which do not have extreme values (outliers) that sway the mean towards them; in distributions with outliers, the mean might represent the outlier more than the rest of the values, which leads to the mean failing to be a useful model. Take for instance the frequency of adjectives in 11 fiction texts taken from the *British National Corpus* (BNC) used as an example to calculate the mean in Section 1.2.

508, 542, 552, 553, 565, 567, 570, 599, 656, 695, 699

The mean of these 11 values is 591.45. However, imagine what would happen if the last value in the series was 6,990 instead of 699. In this case, the mean would be pulled towards the extreme value and we would get 1,163.36; this number is a poor model for the data because only one out of 11 values is above 1,000. One way around this problem of the sensitivity of the mean to outliers is to use the median instead. The **median (mdn)** is the middle value in a series of values ordered from the smallest to the largest. For our 11 values the median is 567, as can be seen from the illustration below.

508, 542, 552, 553, 565, **567**, 570, 599, 656, 695, 699

The median will always stay in the middle of the distribution regardless of what happens at the periphery i.e. whether we have 699 or 6,990 or even 69,900 as the maximum.

If we had ten instead of eleven values, which is an even number, the median would lie half way between the two central values 565 and 567, as demonstrated below.

508, 542, 552, 553, **565**, **566**, 567, 570, 599, 656, 695

The general rule for the median is this: the median is the middle value in the case of an odd number of values; in the case of an even number of values, the

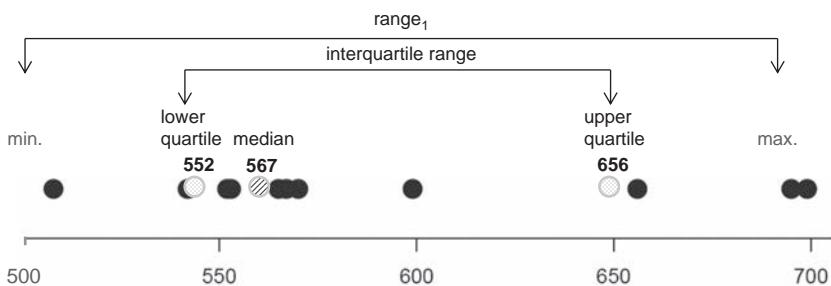


Figure 1.6 *Dispersion of adjective frequencies in 11 corpus files*

median is the mean value of the two central values (i.e. we first add the two central values and then divide them by two).

Dispersion is the spread of values of a variable in a dataset. Take again the adjective counts analysed in Section 1.2 sorted from the lowest (508) to the highest (699) value:

508, 542, 552, 553, 565, 567, 570, 599, 656, 695, 699

When we plot these values on a line we can see their dispersion, i.e. how far apart the individual values are. The distance between the smallest and the largest value is called the **range₁**. The subscript indicates that there is another concept of ‘range’ (range_2) used in corpus linguistics which is explained in detail in Chapter 2. Range₁ gives us an indication of the dispersion of a scale variable. However, the range is affected by extreme values (outliers) and therefore the **interquartile range** is usually preferred. The interquartile range is the interval between the lower and the upper quartile – the lower and upper boundary of the ‘middle bulk’ of the values sorted from the lowest to the highest – and represents 50% of the values from the distribution excluding the median and the borders (quartile values). The concepts of range₁ and interquartile range are demonstrated in Figure 1.6. Here we have divided the whole distribution into four quarters and identified the values that form the borders of the quarters (lower and upper quartile); the interquartile range is the interval containing all values between the lower and the upper quartile.

An alternative dispersion measure to the range₁ and the interquartile range is standard deviation. **Standard deviation (SD)** is the square root of the sums of squared distances of the individual values from the mean. This gives us an indication of the overall distance of individual values from the mean (see Chapter 2 for more detail).

Statistical measure is a general term for any statistic we calculate. It can be as simple as the mean or it can involve complex statistical modelling such as mixed-effects models (Chapter 6). Other examples of statistical measures discussed in this book are: *SD*, Cohen’s *d*, MI-score, Delta P, t-score, F-score.

A **statistical test** is a procedure in a branch of **inferential statistics**, i.e. statistics that goes beyond the sample (corpus) to infer something about the population (language use as such). Its underlying logic rests on ‘null-hypothesis significance testing’ (**NHST**) – see below. The most visible sign of a statistical test is a p-value that the test produces (e.g. in research reports, you can often see mentions such as ‘ $p < .05$ ’ or ‘ $p < .01$ ’). Based on the p-value we can decide whether we have enough evidence in the corpus (sample) to reject the null hypothesis. The procedure follows these steps:

1. We start with the hypothesis we want to test called the alternative hypothesis or H_1 . For example, a sociolinguistic H_1 can claim that *men and women differ in the use of swearwords*.
2. We formulate the null hypothesis (H_0) that is the reverse of H_1 . To put it very simply, the **null hypothesis** states that there is nothing special going on in the corpus or corpora we analyse, e.g. there is no difference between two (sub)corpora. In our example, the H_0 would therefore claim that *there is no difference between men and women when it comes to the use of swearwords*.
3. We test the null hypothesis using a statistical test such as the independent samples t-test. Before doing this, however, we need to check that our data satisfies the assumptions of the selected test (see ‘assumptions’ below).
4. We usually get two important values from a statistical test: (a) the test statistic and (b) the p-value. Based on the p-value (i.e. the probability value of the observation in the corpus by chance alone) we decide whether to reject the null hypothesis. If the p-value is small enough, usually smaller than 0.05, i.e. 5%, we reject the null hypothesis and conclude that the observed difference is unlikely to be due to chance and therefore the result is **statistically significant**. This means that the difference observed in the corpus (sample) is likely to be a true difference in the population (all language use). If the p-value is equal to or is larger than 0.05 (or 5%) we conclude that there is not enough evidence in the corpus to reject the null hypothesis. We need to be careful when interpreting a result like this, which should not be taken to mean that the alternative hypothesis (H_1) is false or that the null hypothesis (H_0) is true: there is simply not enough evidence to reject H_0 ; if we collect more data the statistical test might turn out significant. Note that 0.05 or 5% is the conventional cut-off point which can be imagined as the risk we are willing to take when inferring from the sample to the population (see p-value below). If we are willing to take only a smaller risk than 5%, we can decide on the p-value cut-off point 0.01 (1%) or even 0.001 (0.01%).

A **p-value** is often the most visible sign of a statistical test (see above). However, it would be misleading to reduce all statistics to p-values. A p-value

is a probability value (p stands for probability) and is one of the outcomes of a statistical test. P-value can be defined as the probability that the data would be at least as extreme as that observed if the null hypothesis were true. In the example of a sociolinguistic research looking at the use of swearwords by men and women we would get a p-value that would give us the probability of seeing the observed or even more extreme difference between the two groups if the null hypothesis were true, that is, if there really was no difference in the population and the difference observed in the corpus (sample) was merely due to chance as the result of a sampling error.

Assumptions of a statistical test, as traditionally understood, are conditions that should be met for the statistical test to produce valid results. One of the typical assumptions of a number of statistical tests called **parametric tests** (e.g. the t-test or ANOVA) is the **normality assumption**. This assumption presupposes that the frequency distribution of the linguistic variable does not deviate considerably from the normal distribution. This is because parametric tests such as the t-test typically compare the means which may not be good models of the values of linguistic variables if the distributions are too skewed (see ‘measures of central tendency’). In such cases, non-parametric tests such as the Mann–Whitney *U* test (non-parametric version of the t-test) can be used. These non-parametric tests typically compare sums of ranks of values rather than the means of the actual values (see Section 6.3). However, statistical research shows (e.g. Boneau 1960, Lumley et al. 2002, Schmider et al. 2010) that many parametric tests such as t-test or ANOVA are actually **robust** to the violation of the normality assumption. This means that they produce valid results even if the normality assumption is violated and the data is skewed. Other examples of common assumptions of statistical tests are:

Homogeneity of variances: the (sub)corpora we compare have similar variance, i.e. dispersion of the individual values.

Independence: most statistical tests used in this book (such as the independent samples t-test or the chi-squared test) presuppose the independence of observations in different (sub)corpora. This means, for example, that the linguistic features in one text are assumed not to be influenced by the linguistic features in another.

Linearity: the relationship between some variables can be modelled by a straight line (see Figure 1.1). However, in other cases a different model such as a curve can be more appropriate. In this situation, the linearity assumption would be violated.

When discussing individual statistical tests in this book, the assumptions of these tests will be listed and discussed in more detail.

The **confidence interval (CI)** in inferential statistics is an attempt to move away from the dichotomous thinking that is often connected with NHST, statistical tests and p-values. Rather than a yes/no decision about statistical significance, the confidence interval provides an estimation of the true value of a statistical measure (such as the mean) or of a difference between two statistical

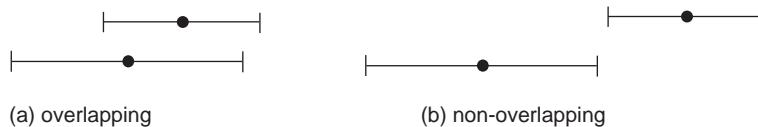


Figure 1.7 *Confidence intervals: two situations*

Table 1.1 *The effect size r and its standard interpretation*

Effect size (r)	Interpretation
0.1	Small effect
0.3	Medium effect
0.5	Large effect

measures (such as the difference between two means) in the population. A confidence interval, as the name suggests, is not a single value but a range of values (that can be visualized as error bars – see Figure 1.7). In corpus linguistics, we often construct 95% confidence intervals. A 95% confidence interval is an interval that is constructed around a statistical measure based on our sample (corpus) in such a way that the true value of this measure lies within this interval for 95% of the samples (corpora) taken from the same population. In practice, we often use CIs to compare two or more (sub)corpora. Here we distinguish two prototypical situations: (a) largely overlapping and (b) non-overlapping CIs (see Figure 1.7). If the CIs largely overlap, the (sub)corpora in question very likely represent the same population, so there is no difference between the groups. On the other hand, non-overlapping CIs signify that the (sub)corpora very likely come from different populations (for more detailed discussion see Cumming et al. 2007).

Effect size in descriptive statistics is a standardized measure, that is a measure comparable across different studies (see Section 8.3 for the discussion of meta-analysis), that expresses the practical importance of the effect observed in the corpus or corpora. For example, if we establish by a statistical test (see above) that two groups of speakers (e.g. men and women) differ from each other in the use of a particular linguistic variable, i.e. there is a statistically significant difference between these two groups, we still need to see how large this difference is and whether it is practically important. To help us with this judgement, effect size measures such as r , odds ratio or Cohen's d can be used. Table 1.1 provides standard interpretations of three cut-off points of the r effect size measure (Cohen 1988). It should be noted that this interpretation provides only general guidance and shouldn't be taken as God's truth (see Section 8.4)

1.4 Building of Corpora and Research Design

Think about ...

Before reading this section, think about the following questions:

1. How many texts do we need to collect to create a corpus?
2. What does it mean to say that a corpus is representative?
3. Are large corpora always better than small corpora?

A corpus is a collection of texts⁹ (or transcripts of speech) that can be analysed using a computer. Corpora can be either *general*, reflecting the language as such, or *specific*, focusing on a particular genre, author or area of language use. A general language corpus is a sample of language taken from a very large population – in the case of a general corpus the population consists of all of the language that people produce during a certain period of time. A *synchronic* corpus samples language around one point in time (e.g. produced during one year), whereas a *diachronic* corpus includes language across different periods. It has been estimated (Mehl et al. 2007) that on average, a person utters 16,000 words per day. With about 400 million people who speak English as their first language and an additional hundreds of millions who speak English as their second language (Crystal 2003), the daily spoken production of English alone can be estimated to be in the order of trillions of words. However, people also write text messages, emails, blog posts, shopping lists, business reports, essays, poems etc., all of which add to the daily linguistic production. In contrast, corpora, even the largest ones, are relatively small when compared to the total of all language production. Yet if well constructed, corpora can tell us something useful about the **population**, i.e. the language as such. But what does ‘well constructed’ mean?

In corpus linguistics the term ‘representative’ is often used to describe a corpus. **Representative** is a descriptor of a sample when it has similar characteristics to the population it is drawn from. This allows us to draw conclusions about the population from the sample. Ideally, this would be achieved by truly **random sampling**¹⁰ where each text ever produced and each spoken interaction that has ever taken place would have the same chance of appearing in the sample. Random sampling is, however, impracticable because there is no catalogue of all language production that we could refer to when doing the sampling; besides, not all language produced is recorded in some form. To make the sampling manageable, corpus designers often start with a set of categories within which

⁹ In general linguistics, the word ‘text’ is often used as a general term to represent both written and oral texts. For corpus analysis, all texts need to exist in a machine-readable format.

¹⁰ In statistics, a difference is often made between probability sampling and non-probability (convenience) sampling. Random sampling is a type of probability sampling.

Table 1.2 *Brown family sampling frame*

Text categories		Number of texts in each category
A	Press: reportage	44
B	Press: editorial	27
C	Press: reviews	17
D	Religion	17
E	Skills, trades and hobbies	36
F	Popular lore	48
G	Belles lettres, biography, essays	75
H	Miscellaneous (government documents, foundation reports, industry reports, college catalogue, industry house organ)	30
J	Learned and scientific writings	80
K	General fiction	29
L	Mystery and detective fiction	24
M	Science fiction	6
N	Adventure and western fiction	29
P	Romance and love story	29
R	Humour	9
Total		500

they aim to collect an unbiased sample. These categories are called the **sampling frame**. Table 1.2 provides an example of a sampling frame. This is the Brown family sampling frame, one of the most well-known sampling frames in corpus linguistics (Francis & Kučera 1979).

The Brown family consists of 15 genre-based categories according to which the total of 500 2,000-word text samples are selected. Each Brown family corpus thus consists of approximately one million words of written English ($500 \times 2,000$). In this traditional corpus design, the aim of the corpus creators is to achieve an unbiased sample of texts in the categories from the sampling frame. Again, we would ideally use the random sampling procedure within the categories. Yet in practice, the selection is guided by text selection principles (see below) to avoid bias in the selection process. By **bias** we mean a systematic but often hidden deviation of the sample from the population.

The following is a list of the most common types of bias and related text selection principles to avoid this bias:

- **Text sample bias:** different sections of texts (e.g. beginning, middle and end) have different linguistic properties. For example, ends of texts tend to include language that summarizes the main point of the text (*in sum, to conclude, they lived happily ever after*). If a corpus samples only certain sections, e.g. by taking the first or the last 2,000 words of each text, these sections will be

overrepresented. In corpora that do not include whole texts but only text samples (like the Brown family) it is therefore important to achieve a balance in terms of the different sections of texts represented.

- **Topic bias:** topic bias is created when many texts on the same topic get included in the corpus (unless the corpus is deliberately constructed as a specialized corpus on the topic). These texts usually contain a number of specific topic-related vocabulary items that are repeated multiple times in each text. This is connected to a so-called ‘whelk problem’ (see Section 2.4) when some infrequent lexical items become dramatically overrepresented in the corpus. Topic bias is a problem especially in small corpora where each individual text forms a relatively large proportion of the corpus. In most cases, corpus designers should therefore consciously select texts on a range of topics.
- **Non-coverage bias:** some texts are more ‘visible’ than others because corpus designers see them as prototypical for different reasons. For example, published texts might be given a preference over private letters or emails. Corpus designers need to actively seek to cover as wide a range of texts as possible.
- **Traditional text type bias:** this type of bias is a specific case of non-coverage bias. When selecting the sampling frame we are often predisposed to see as salient the text types that have traditionally been included in language corpora such as those in the Brown family sampling frame (see Table 1.2). However, with new technologies and modes of communication new genres such as blog posts and tweets emerge. Corpus creators thus need to think carefully about what text types to sample and what the implications of these are for the representativeness of the corpus.
- **Legal considerations bias:** corpus designers often face the problem of copyright; this is especially the case when corpus creators want to share their corpora with other researchers. Legal considerations may thus lead to a selection of texts to which copyright does not apply (older out-of-copyright texts, texts under creative commons licences etc.), which, however, creates a problem with biased sampling. Currently, there is no clear solution to this issue as approaches may differ according to legal requirements in individual countries. Corpus designers, however, should be mindful of this problem.
- **Practicality bias:** some texts such as webpages are easier to obtain than others. This practicality consideration may lead to creating a corpus that overrepresents easily obtainable texts. Corpus creators need to resist this temptation and strive for a range of texts regardless of whether they can be obtained easily or not.
- **Self-selection bias:** this bias is created when contributors (i.e. authors of texts) are asked to provide texts on a voluntary basis. For instance, if we want to create a corpus of classroom writing and ask students to volunteer and contribute their texts, we may end up with texts from highly motivated students that will not reflect the written production of the class as such. Corpus designers therefore

need to reach out to different groups of contributors and use a range of incentives to obtain a representative sample.

So far, the traditional approach to corpus design has been considered. However, a different approach to corpus representativeness and sampling emerged from the **web as corpus** initiative (Baroni & Ueyama 2006; Baroni et al. 2009; Jakubiček et al. 2013). The leading principle of this approach is based on a simple observation that with the rise and the growing popularity of the internet, large quantities of interesting linguistic material have become available online. In addition, the online linguistic environment has expanded to cover a wide variety of features of our daily life – work, entertainment, social interaction, etc. By crawling the web, we can build corpora that are larger than ever before and sample the online language more systematically than we would be able to with the traditional approach based on a set of text-type categories (sampling frame). It is claimed that large online corpora thus have the potential to be more truly representative of current language use, especially because the online environment is becoming increasingly more comprehensive, mirroring to a large extent offline language production (e.g. printed books are also available as ebooks). However, even the internet, despite its enormous size, represents only a certain proportion of the total of linguistic production, excluding for instance face-to-face informal conversation, which forms a large part of everyday language production. So, as with any corpus, when analysing a web-based corpus we need to think carefully about the relationship between the corpus as a sample and the language production (population) which it represents.

As discussed, a corpus is usually a sample of language. However, in some specific cases, a corpus can include the whole population. For example, in studies of literature, corpora comprising all works by a particular author include the whole population. If we wish to compare the speech of Prince Hamlet with the speech of his friend Horatio in Shakespeare's famous tragedy, we will be working with all the evidence there is about the linguistic behaviour of these two characters. Similarly, if we collect all newspaper articles about a particular topic in a given period we will be looking at the whole population of articles that were written on the particular topic. Baker et al. (2013) with their 143-million-word corpus of British newspaper articles that contain words relating to Muslims and Islam come very close to this ideal. In these and similar cases, the statistical method appropriate for the analysis of **population-based corpora** is that of description (i.e. descriptive statistics). The process of statistical inference can be skipped (see the discussion of the role of statistics in corpus analysis below) because we will be able to observe the true state of affairs.

Having discussed different aspects of corpus building, one basic question still remains to be answered: how large should a corpus be? There is no universal answer to this query because **corpus size depends on the research question** and the kind of linguistic features we want to investigate. For the investigation of

Table 1.3 *Frequencies of selected words and expressions in three English corpora*

Expression	BE06 (1 million words)	BNC (100 million words)	EnTenTen13 (20 billion words)
co-pilot ^a (noun)	0	69	7,887
rater ^a (noun)	0	11	7,832
beautiful memories	0	1	2,552
somewhat humorously	0	0	90
uninhabitably	0	0	4

^a Frequencies include plural forms.

grammatical structures such as passives, which are generally fairly common, even a small corpus (e.g. one million words or less) can be sufficient. On the other hand, many lexical items and their combinations are fairly infrequent even in very large corpora and we may easily encounter the **data sparsity** problem. To illustrate this issue, Table 1.3 shows five expressions and their frequencies in a one-million, 100-million and 20-billion word corpus respectively. Although all of these expressions are straightforward and entirely transparent in terms of their meaning, they do not appear in the one-million-word corpus at all. Interestingly, the adverb *uninhabitably* (e.g. *Fumes from this single burn will eventually render a 44 km area of the lunar surface uninhabitably radioactive*) has only four occurrences in 20 billion words.

From this example, we can derive a general rule: unless the corpus represents the whole population, **the absence of evidence is not the evidence of absence**. In other words, if an expression does not appear in a corpus, this doesn't mean that this expression is non-existent. As corpus users we therefore need to think critically about the nature of the evidence that corpora provide in terms of their quality (representativeness and balance) as well as their quantity (corpus size).

Finally, a few words need to be said about the role of statistics in the analysis of corpora. Let us start by considering some general principles followed by the discussion of specific research designs. In the process of corpus analysis, there are four separate but interconnected dimensions (see Table 1.4); each of them needs to be interpreted properly in order to achieve a meaningful result.

We usually start our analysis with data exploration during which we look at frequencies and distributions of linguistic variables (see Section 1.3 for statistical terminology) and often produce graphs which capture the main patterns in the data (see Section 1.5 for data exploration and visualization). If our corpus represents a sample rather than the population (which is typically the case), we should consider the amount of evidence we have in the sample. In other words, we can use inferential statistics to enquire whether the observed effects and differences between (sub)corpora can be generalized to the population, i.e. all language that

Table 1.4 *Different levels of analysis in corpus linguistics*

Dimension	Key questions	Key terms
1. DATA EXPLORATION	<i>What are the main tendencies in the data?</i>	Graphs, means, SDs
2. INFERRENTIAL STATISTICS: AMOUNT OF EVIDENCE	<i>Do we have enough evidence to reject the null hypothesis? Is the effect that we see in the sample due to chance (sampling error) or does it reflect something true about the population?</i>	statistically significant p-values confidence intervals
	<p>The diagram illustrates the concept of statistical inference. At the top, the text 'statistical inference' is written above a downward-pointing arrow. Below the arrow is a horizontal double-headed arrow spanning three categories: 'one example (not enough evidence)' on the left, 'large enough sample' in the middle, and 'population (the whole truth)' on the right. A small triangle points towards the 'large enough sample' category.</p>	
3. EFFECT SIZE	<i>How large is the effect in the sample? (standardized measure)</i>	effect size e.g. Cohen's d , r
4. LINGUISTIC INTERPRETATION	<i>Is the effect linguistically/socially meaningful?</i>	

the corpus or corpora represent. Inferential statistics produces p-values or confidence intervals and we use words such as ‘statistically significant’ or ‘non-overlapping 95% confidence intervals’ to describe the inferences.

Currently, there is a debate in a number of disciplines such as psychology, sociology and applied linguistics about the place that inferential statistics, especially p-values, should have in the research process.¹¹ Unfortunately, many researchers are still trapped in dichotomous thinking about their results as statistically significant or not significant, often confusing statistical significance with practical importance or linguistic and social meaningfulness (see below). For more information see Cumming (2012), who offers convincing arguments in favour of what he calls ‘new statistics’ – statistical procedures based on effect sizes and estimation of confidence intervals. For our purposes it is sufficient to stress that ‘statistical significance’, ‘practical importance’ and ‘linguistic and social meaningfulness’ are three distinct concepts. **Statistical significance** tells us whether we have enough evidence in the corpus to reject the null hypothesis; in very large corpora even a small (hardly noticeable) difference between two groups of texts/speakers turns out to be statistically

¹¹ The journal *Basic and Applied Social Psychology* even goes as far as putting a ban on all inferential statistics in submissions to the journal. The editors of this journal claim that the procedure involved in inferential statistics – the so-called null-hypothesis significance testing (NHST) procedure – ‘is invalid’ (Trafimow & Marks, 2015: 1).

(a)			(b)		
Case(corpus)	Passives(AF)	Passives(RF)	Case(text)	Passives(AF)	Passives(RF)
BNC	1121436	11406.74	A00	50	72.5
			A01	81	99.8
			A02	24	70.0
			A03	369	184.8
			A04	464	117.2
			A05	580	137.1
			A06	280	76.5
			A07	424	106.2
			A08	205	51.1

Case(feature)	Short/Long Passive	Speech/Writing	Genre	Example		
1	0	1	0	Hedging plants are usually cut	back to half	
2	0	1	1	regions, but it has been deployed	under secto	
3	0	1	1	BBC's recordings aren't meant	for release c	
4	1	1	One-way system. It was caused	by the IRA, v		
5	0	1	Development projects are scheduled	for the forth		
6	0	1	3ety grew and laws were passed	for her proté		
7	0	1	Oberal policies will be implemented	in Peru at le		
8	1	1	0Romans, the Celts were dismissed	by contemp		
9	0	1	4 solar calendar by being placed	at the winte		
10	0	1	5 Final Invoice will be issued	as appropri		
11	0	1	6 he tissue samples are taken	from the foe		

Figure 1.8 *Research designs in corpus linguistics*

significant. **Practical importance** uses standardized statistical measures to express the size of the effect; here we are trying to evaluate the magnitude of the effect (e.g. how large the difference really is between two groups). Finally, we need to relate the observed effect back to what we know about language and society and interpret the results in the context of linguistic and social theory. This crucial step seeks to discover **linguistic and social meaningfulness** of what we observed in corpora.

When analysing corpora, we also need to think about the type and format of the data that needs to be obtained from the corpus in order to answer the research question. This is a so-called **research design**; research design is important because it has considerable implications for the specific statistical procedures that we can use with the data. In general, three main types of research design can be distinguished: (1) whole corpus design, (2) individual text/speaker design and (3) linguistic feature design. Figure 1.8 exemplifies the structure of datasets based on these three research designs.

In the **whole corpus design**, the unit of analysis is usually the whole corpus, sometimes also large subcorpora. In Figure 1.8a, the corpus in question is the BNC, in which we trace the frequency of the passive construction (e.g. *he was*

seen) as our main linguistic variable. AF stands for absolute frequency and RF signifies relative frequency per million words (see Chapter 2 for a fuller discussion). Whole corpus design is the most basic type of design that provides only very general information about frequencies of linguistic features in corpora without reference to internal variation inside the corpus (see Section 4.2 for a practical example of the whole corpus design). In contrast, **individual text/speaker design** enables us to trace the frequency of a linguistic feature in individual texts or speakers. In Figure 1.8b, we can see the frequencies of passives in the individual BNC files; the figure displays only the first nine files in alphabetical order (see Section 6.3 for a practical example of the individual-text/speaker design). Finally, the **linguistic feature design** focuses on the linguistic feature as a single observation (case). In Figure 1.8c, each line represents a single occurrence of the passive construction in the BNC, which is categorized according to different properties of the construction and the context of occurrence. For example, the passive construction can be short (e.g. *plants are usually cut*) – coded as 0 – or long with a *by*-phrase (e.g. *It was caused by the IRA*) – coded as 1; it can appear in speech (0) or writing (1) and in different genres (coded as 0–6). The linguistic feature design allows us to investigate different factors that play a role in the use of a particular type of linguistic feature. For example, from the dataset in Figure 1.8c, we would be able to research whether any particular genre favours long passives over short passives (see Sections 4.3 and 4.4 for a practical example of the linguistic feature design). Paying attention to the ‘shape’ of data and the research design is crucial for the success of the statistical analysis. Note that different research questions require different research designs, which in turn produce different ‘shapes’ of data. This book discusses the appropriate research design for different types of linguistic analyses and demonstrates the use of statistical techniques with multiple case studies; full datasets used in this book are available from the companion website.

In sum, selecting the right corpus, analytical procedure and corpus design is the first step in successful corpus analysis. As researchers, we need to think carefully about what the corpora we are working with represent and how they can reveal interesting findings about language and society.

1.5 Exploring Data and Data Visualization

Think about ...

Before reading this section, think about the following questions:

1. Why is looking critically at data before analysis important?
2. What types of errors can we encounter in a dataset?
3. What types of graphs do you know?

Bad data leads to bad results. No matter how sophisticated our statistical analysis is, if we come up with wrong numbers or make a mistake when copy-pasting the data from a spreadsheet into the statistical software, we will end up with results that are simply wrong. One way of preventing these accidental errors is to keep a research journal which records every step of the analysis so that the procedure can be easily checked or repeated. Another, and probably even more important, aspect of good data analysis is constant questioning of the ‘sanity’ of the data: *Is this the expected size of the corpus or have I counted also part-of-speech tags by mistake? Does this effect make sense given what I know about the distributions of words in texts? What is the reason for this unusual data point? etc.* Constantly questioning the data can help avoid making trivial errors and misinterpreting the results.

In addition, in order to understand the main trends in the data, data visualization is crucial. Effective data visualization summarizes patterns in data without hiding important features. The following example illustrates how effective visualization works. Figure 1.9 and Figure 1.10 display the same dataset. This dataset compares three language corpora in terms of the frequencies of a linguistic variable – as this is a made-up example simply for purposes of illustration, let us call it x . The **bar chart** in Figure 1.9 shows the frequencies of x in each corpus; from this display, it appears that the corpora clearly differ from each other, with Corpus 3 having the largest frequency of variable x .

However, Figure 1.10, based on exactly the same dataset, tells a very different story. Corpus 1 indeed appears to be different, but Corpora 2 and 3 are to a large extent similar, although Corpus 3 is more widely dispersed than Corpus 2. In particular, there is one value in Corpus 3 that stands out – this is what we call an **outlier**. In Figure 1.10, the data is presented in the form of a **boxplot** with added means and individual data points. The boxplot shows the distribution of the data in each corpus as well as the extreme values (minimum and maximum) and outliers, where applicable. The inside of the box represents the **interquartile range** (50% of values) and the thick horizontal line in the box represents the **median** (middle value). The ‘whiskers’ above and below the box show the minimum and maximum values but exclude outliers (extreme values), which are displayed outside the scope of the whiskers as separate data points. Individual values (frequencies of x in different texts) are also displayed in this boxplot together with the mean (short horizontal line).

In sum, a boxplot provides much more information about the data than a simple barchart. While the barchart in Figure 1.9 only displays three values (2.2–6.1–8.1), the boxplot in Figure 1.10 uses the same space to display 46 different pieces of information, including important summary statistics such as the mean, median, range and interquartile range.¹² This helps us interpret the data more meaningfully and avoid unfounded conclusions.

¹² When visually presenting data, we should aim at rich (informative) forms of display with a high information/ink ratio (Tufte 2001, 2006; Hudson 2015).

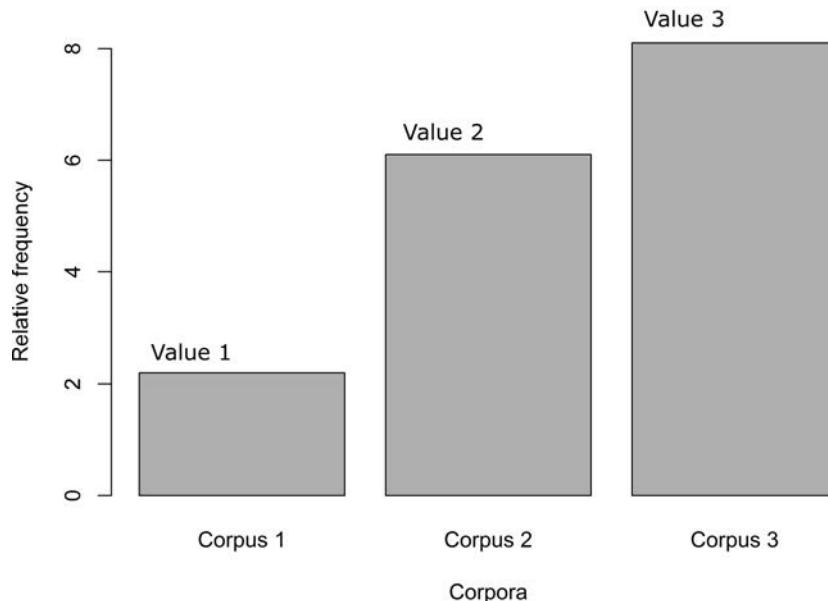


Figure 1.9 Bar chart: variable x in three corpora

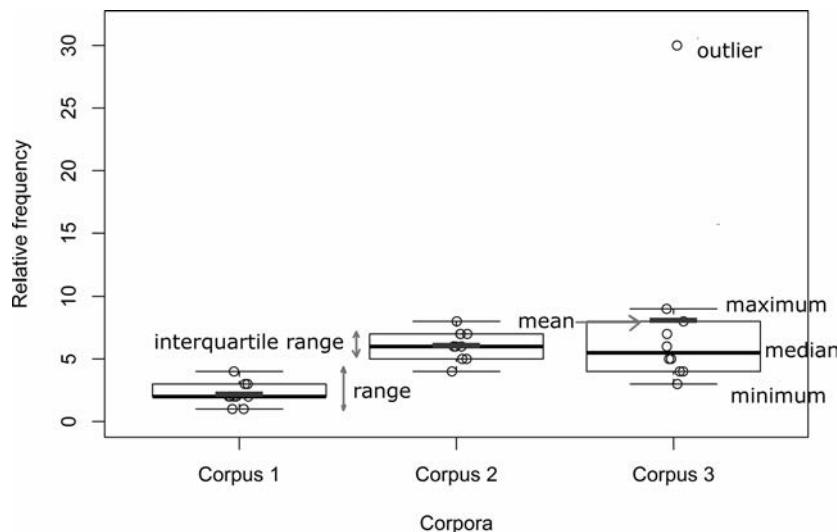


Figure 1.10 Boxplot: variable x in three corpora

If we want to go beyond the sample and generalize about the population we can calculate 95% confidence intervals for the mean value of variable x in the three corpora. The 95% confidence intervals are displayed as **error bars** in Figure 1.11.

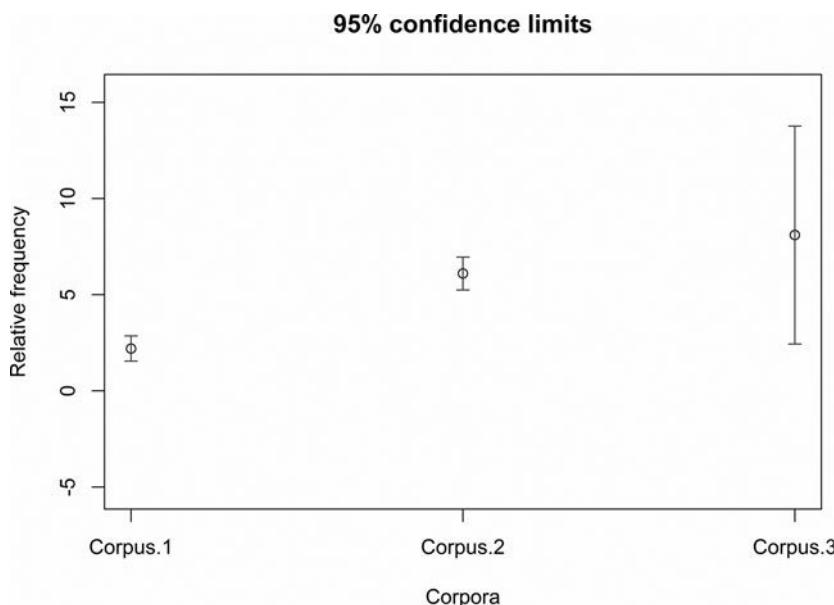


Figure 1.11 *Error bars: variable x in three corpora*

Another useful display of data is a histogram (see Figures 1.12 and 1.13). A **histogram** is a graph that shows the frequency distribution of a linguistic variable in the form of bars, each representing the frequency of values of the linguistic variable in a given interval or bin (e.g. 20–30, 31–40 etc. in Figure 1.12). In histograms, we are interested in the shape of the observed distribution of the data. The examples in Figures 1.12 and 1.13 show the relative frequency (per 1,000 words) distribution of the definite article in writing and the relative frequency (per 1,000 words) distribution of the f-word (*fuck, fucked, fucking*) in speech respectively. Here, we can see two positively skewed distributions, a typical shape for linguistic data. In the case of the f-word the distribution is extremely skewed because very few people in the corpus use the f-word at all.

So far, we have been dealing with a single linguistic variable. If we want to compare the relationship between multiple linguistic variables, a **scatterplot** or a series of scatterplots can be used. Below are scatterplots based on the use of the definite article, first-person pronoun and second-person pronoun in BNC64, a 1.5-million subsample of the spoken BNC. Looking at Figure 1.14, we can see that the use of the definite article *the* has been plotted against the first-person pronoun *I*. Each circle represents an individual data point – a speaker in this case. From the graph, we can thus clearly see how individual speakers use these two linguistic variables and what the relationship between the variables is. The

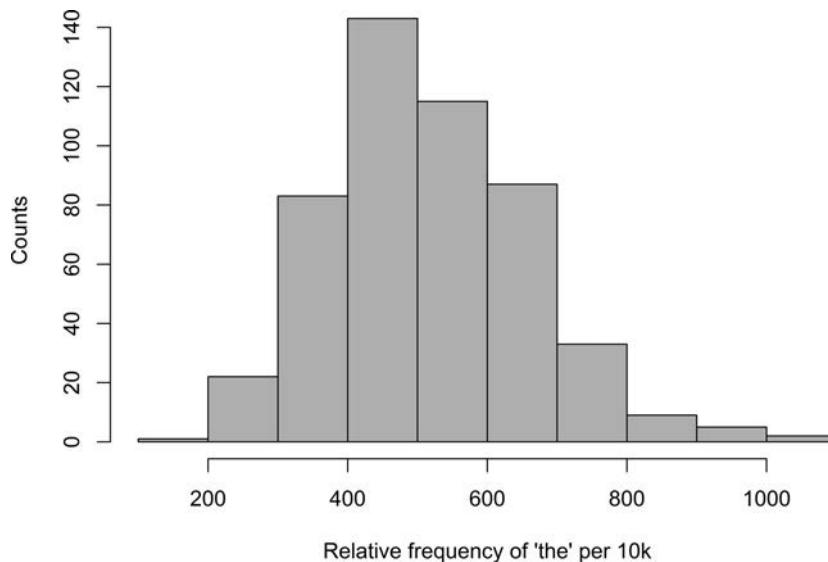


Figure 1.12 *Histogram: the definite article in BE06*

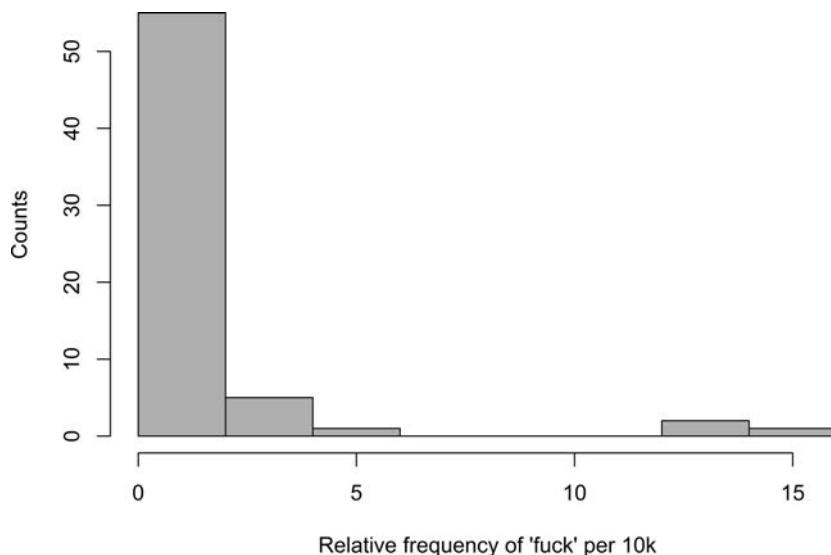


Figure 1.13 *Histogram: the f-word in BNC64*

regression line going through the middle of the ‘speaker cloud’ explicitly shows that in this dataset, the use of the definite article is inversely proportional to the use of the first-person pronoun. In Figure 1.15, multiple scatterplots are displayed as a **scatterplot matrix**, each presenting a combination of

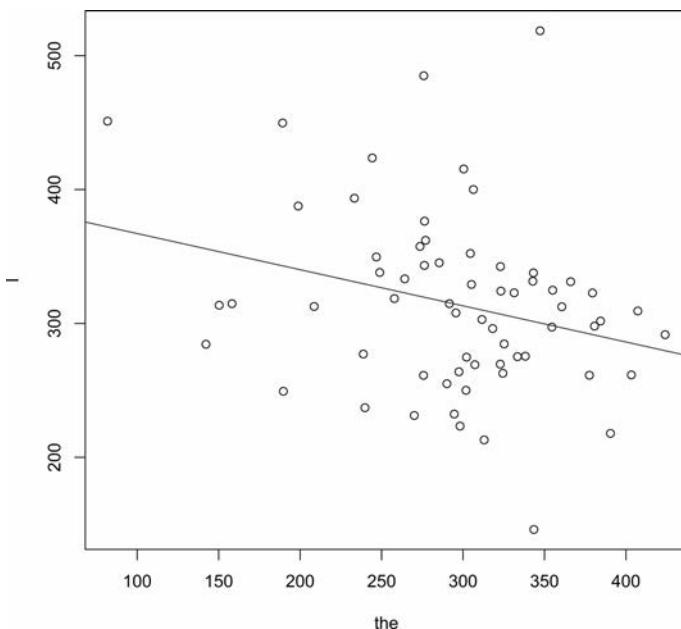


Figure 1.14 Scatterplot: the and I in BNC64

two of the three linguistic variables. For example, the middle square in the first column is a scatterplot of *the* and *I*, with *I* on the x-axis and *the* on the y-axis. The best way to read a scatterplot matrix is this: the label (word) in the column indicates what's on the x-axis, while the label (word) in the row indicates what's on the y-axis.

There are also more complex types of visualization that can be used with corpora. Geomapping, for instance, is an efficient way of showing mentions of different places in a corpus and their frequencies on a map. The map below displays the top ten places from the BNC that collocate with the verbs *to go* or *to travel*. These are (in the frequency order): London, Paris, Oxford, Rome, Cambridge, Manchester, New York, Leeds, Edinburgh and Liverpool.

The type of visualization we use should provide a means for us to better understand the general patterns in the corpus data. This largely depends on the research question and the type of study (research design) we are dealing with. Figures 1.17 provides an overview of different types of more specific graphs used in this book with references to the chapters where these are discussed in detail.

All graphs presented in this section can be easily produced using the Graph tool from Lancaster Stats Tools online.

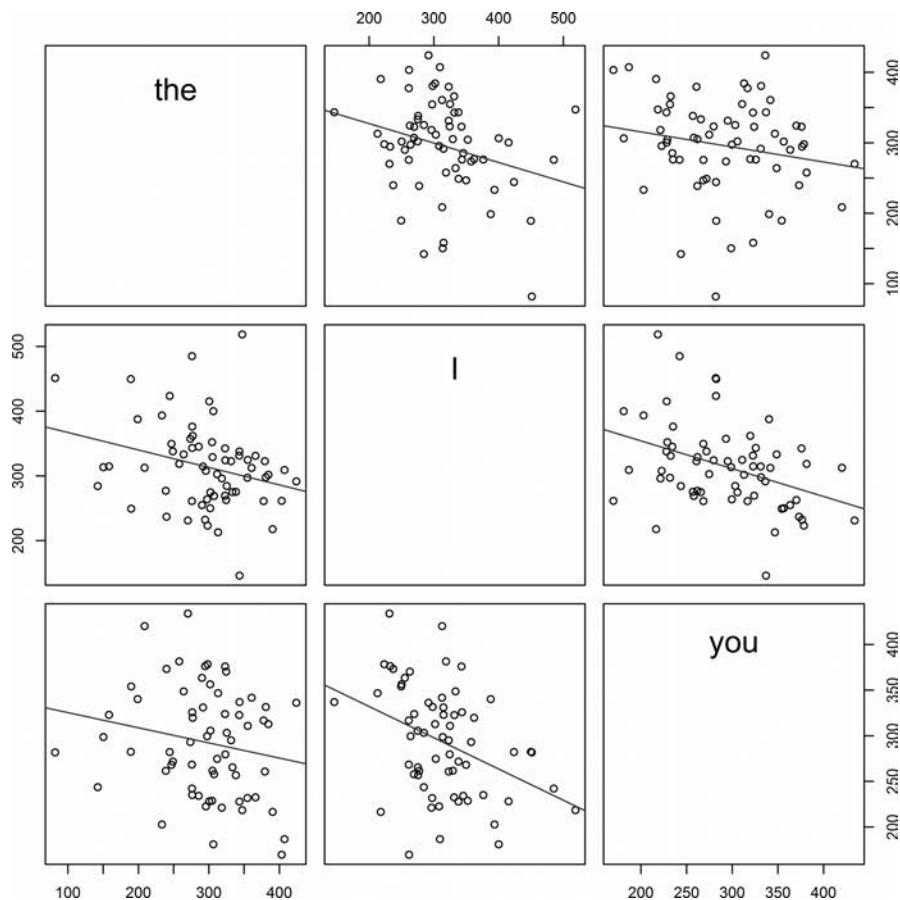


Figure 1.15 *Scatterplot: the, I and you in BNC64*



Figure 1.16 *Top ten places connected with 'going' or 'travelling' in the BNC*

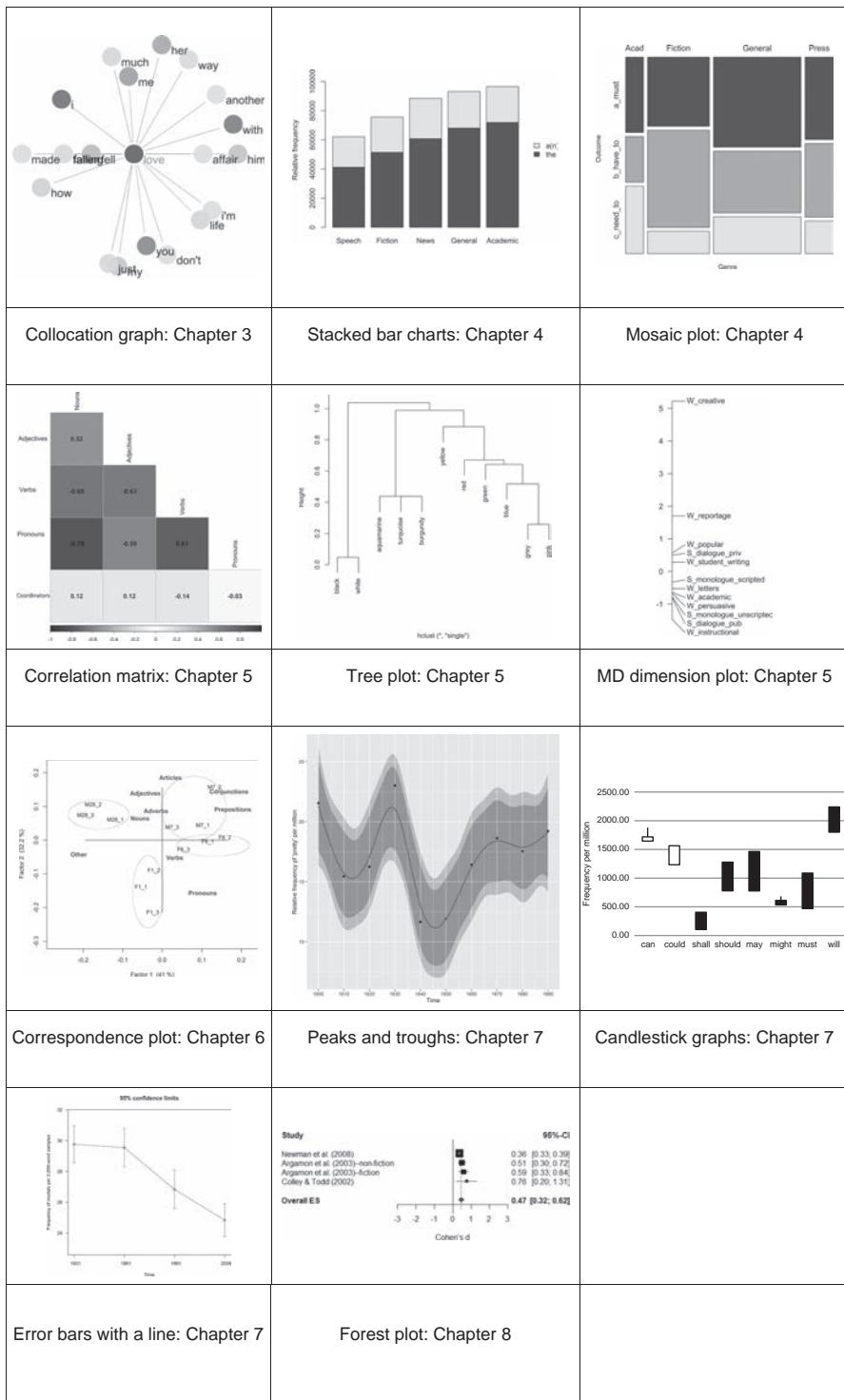


Figure 1.17 Other types of visualizations

1.6 Application and Further Examples: Do Fiction Writers Use More Adjectives than Academics?

A friend of mine who is an excellent novel writer once told me that she thinks academic writing is dry because academics use very few adjectives. I thought about this for a moment and then, instead of producing a witty reply (which I couldn't think of anyway), I said: 'that's an empirical question'. Next, I offered to check her hypothesis in the BNC, a corpus that samples both academic writing and fiction. The following is a short report I prepared for my friend. Because my friend has very little knowledge about corpora and statistics I used the margins of the report for explanatory notes. Both the report and the notes are reproduced below.

This study is based on the BNC. In particular, two subcorpora of the BNC were extracted to investigate the difference in the use of adjectives in fiction and academic writing. Table 1.5 shows the two subcorpora.

Table 1.5 *Subcorpora in mini-research*

Subcorpus	BNC_fiction	BNC_academic
Words (tokens)	16,075,667	15,619,286
Files	432	501

BNC represents the use of British English in the early 1990s. So the corpus is already fairly old. However, the use of adjectives is a stable enough linguistic variable, so there is no need to worry about the representativeness too much. If you are worried about this, though, this study can be replicated with some more recent data.

In this study, each file representing a text written by a single author was considered as a separate observation. The following hypothesis was tested:

- Hypothesis (H_1): Academics use fewer adjectives than fiction writers.

This just says that I looked at individual differences between writers, not only at group averages.

This is your hypothesis, remember?

The data was first explored using a boxplot. Then, 95% confidence intervals were calculated for the two subcorpora and r was used as a standard effect size measure. Finally, the independent samples t-test was used to test the null hypothesis that there is no difference between the two groups of writers:

Null hypothesis (H_0): There is no difference between academics' and fiction writers' use of adjectives.

These are different statistical techniques. You don't need to understand the details.

Null hypothesis is part of the formal statistical procedure. It is a negation of your hypothesis.

The results suggest that there is indeed a difference between academics and fiction writers in terms of their use of adjectives. However, as the boxplots in Figure 1.18 show, the difference is in favour of academic writers, i.e. they appear to use more adjectives than fiction writers. We can also observe a larger homogeneity among fiction writers than academic writers.

Moving beyond the sample, 95% confidence intervals can be calculated. Figure 1.19 reveals non-overlapping confidence intervals which are far apart for the two compared groups. This can be interpreted as a clear, statistically significant difference between the groups with academic writers using far more adjectives than fiction writers.

Independent samples t-test confirmed that there is a statistically significant difference between fiction writers and academic

Here we go;)

Larger homogeneity is signalled by a smaller box and shorter ‘whiskers’.

This is a statistical technique to test whether the difference is true also for fiction writers and academics in general and not only for the fiction writers and academics sampled in the BNC (provided the BNC is an unbiased representative sample).

More ‘stats magic’ just to confirm the same point.

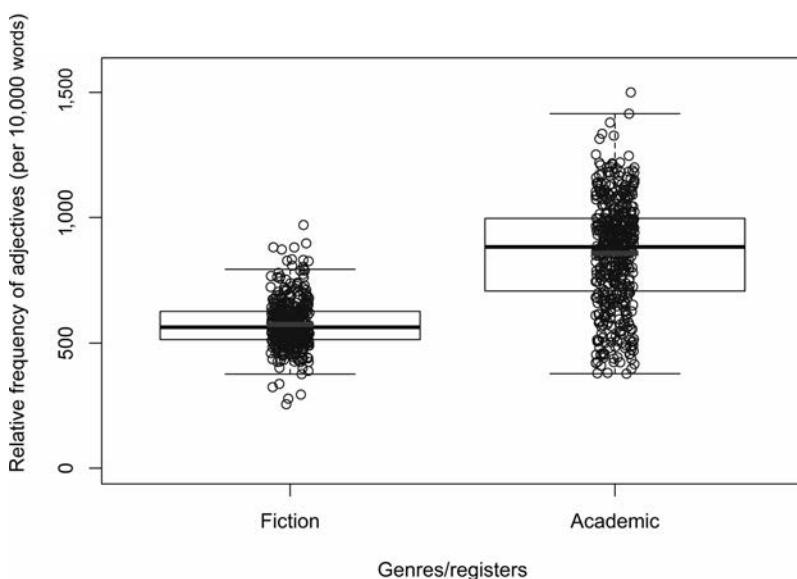


Figure 1.18 *The use of adjectives by fiction and academic writers: boxplot*

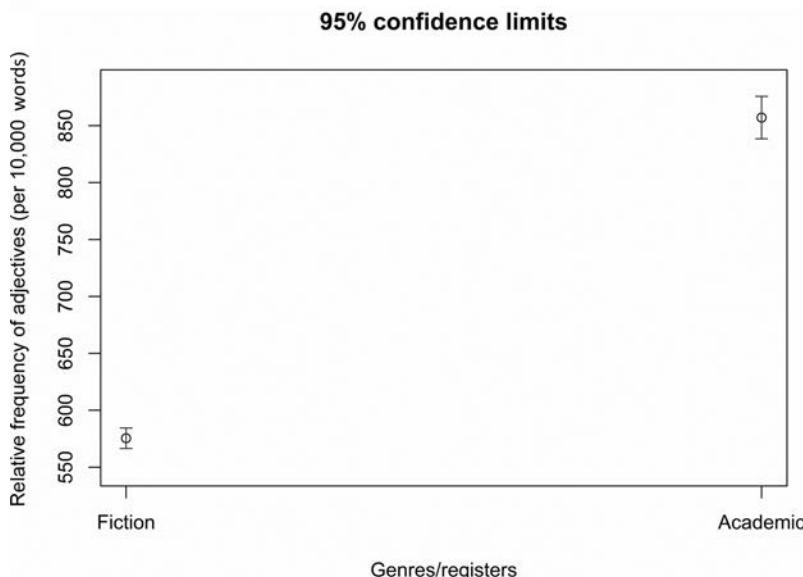


Figure 1.19 *The use of adjectives by fiction and academic writers: error bars*

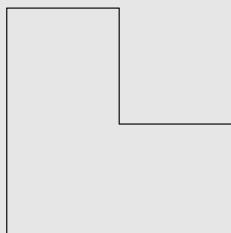
writers ($t(714.67) = -26.78, p < .001$). The effect size was large ($r = .708$), which can be related back to the linguistic reality as the difference between 857 adjectives per 10,000 words (academics) and 575 adjectives per 10,000 words (fiction writers). This means that academic writers on average use over 280 adjectives more in 10,000 words than fiction writers; this seems to be a stylistically meaningful difference, i.e. a difference noticeable by the reader.

The difference is not only statistically significant, it also appears to be linguistically meaningful.

With a big smile on my face I presented the report to my friend. ‘Hm, interesting . . .’ she said. ‘But I still think that the adjectives fiction writers use are somehow richer.’ ‘That is a different research question!’ I exclaimed in exasperation. Only much later did I add ‘Maybe you are right, though . . . Why don’t we find out?’

1.7 Exercises

- As a warm-up exercise (with a twist), divide the following shape that represents three quarters of a square into *four* identical shapes. Feel free to skip this exercise if you want to focus on statistical techniques immediately.



After you have done this, take a whole square, but this time divide it into *five* identical shapes.



2. Calculate the mean for the following numbers: 2339, 2089, 2056, 2276, 2233, 2056, 2241, 1995, 2043, 1976, 2062. These are the frequencies of verbs in fiction texts by British writers discussed in this chapter (Section 1.2).
3. What is a model in scientific thinking?
4. Select the best-fitting geometrical model for the area of Great Britain (see Figure 1.20) that would help you investigate the area of the island.
 - (a) rectangle
 - (b) circle
 - (c) triangle

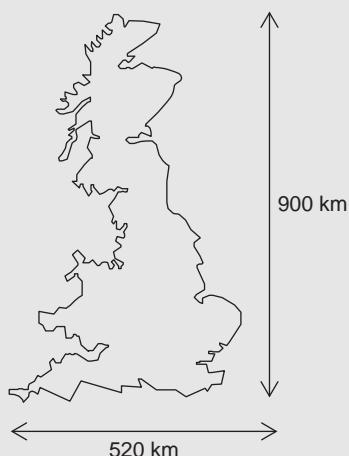


Figure 1.20 Great Britain: main island

5. Use the model and the dimensions given in Figure 1.20 to calculate the area of Great Britain (the large island only).
6. Test your knowledge about the basic statistical terminology in a short quiz.
 - i. **What is the difference between the *average* and the *mean*?**
 - (a) There is no difference; these terms are synonyms.
 - (b) Mean is a type of average; so are the *median* and *trimmed mean*.
 - (c) Mean is usually larger than the average.
 - ii. **What is the *mean* of the following values: 5, 10, 15, 20, 25?**
 - (a) 15.0
 - (b) 17.32
 - (c) 25.3
 - iii. **What is the *median* of the following values: 5, 10, 15, 20, 25?**
 - (a) 10
 - (b) 15
 - (c) 20
 - iv. **What type of variable is the rank of words in a frequency list?**
 - (a) nominal
 - (b) ordinal
 - (c) scale
 - v. **What is dispersion in a corpus?**
 - (a) The distribution of a (linguistic) variable in different parts of a corpus.
 - (b) Another term for standard deviation.
 - (c) Spread of a sample in a population.
 - vi. **If you plot a normally distributed set of data, what will the shape of the graph be?**
 - (a) Flat
 - (b) J-curve
 - (c) Bell-shaped
 - vii. **What is a 95% confidence interval?**
 - (a) An interval that shows that we can be 95% confident in the correctness of the result within this interval.
 - (b) The measure of objectivity of our findings.
 - (c) An interval constructed around a particular measure in a sample in such a way that the true value of the measure in the population will fall within this interval for 95% of samples.
 - viii. **What is a p-value?**
 - (a) The probability that the null hypothesis is true.
 - (b) The probability of seeing values at least as extreme as observed if the null hypothesis were true.
 - (c) The probability of seeing a unicorn in Lancaster.
 7. Imagine you have 500 texts (the population of interest), 250 written by a male and 250 written by a female author. However, you don't know which one is which. For the purposes of your study you want to select an unbiased sample of

40 texts that would represent the population. Use the Random number generator from the Lancaster Stats Tools online to create a list of 40 random numbers between 1 and 500 and note these down.

8. In the Answer section at the companion website, you can find which texts were written by a male speaker and which by a female speaker. Check the answers in Exercise 7 and calculate the number of male and female speakers that were selected.

Male speakers in the sample: _____

Female speakers in the sample: _____

Did you get an approximately equal representation of male and female speakers in the sample?

9. Later, you also find out that half of the 500 texts were written by a young speaker and another half by an older speaker (see the Answer section at the companion website to find out which texts these are). Review the sample from Exercise 7 to find out if you have an approximately equal gender and age representation in your sample of 40 texts.

Age \ Gender	Male speaker	Female speaker
Young speaker		
Older speaker		

10. Was random sampling a successful method? Why (not)?
11. What types of bias do we need to avoid in corpus design?
12. What type of research design (Whole corpus, Linguistic feature, Individual text/speaker) would you use in the following situations?
- (a) To find out if the zero relativizer in relative clauses (e.g. *The second thing Ø I want to say is . . .*) is more frequent in spoken or written language.
 - (b) To find out if hedges (e.g. *sort of, kind of*) are more common in male or female speech.
 - (c) To find out the frequency of the definite article *the* in written English.
 - (d) To investigate register-based variation of a large number of linguistic features including modals, discourse markers and private verbs.
13. Can you spot six errors in the following dataset based on BE06, an approximately one-million-word corpus of written British English?

Word or expression	Frequency	Frequency per million
the	5,896	5142.17
of	30,666	26745.23
and	27,909	24340.72
to	26,188	2283.98
of the	6,887	6006.47
and the	19,530	17033.01
Words total	2,293,194	—

14. Which visualization type (graph) would be appropriate in the following situations?
- Describing the frequency distribution of a linguistic variable in one corpus.
 - Comparing the distribution of a linguistic variable in two corpora.
 - Finding the relationship between two linguistic variables.
 - Estimating if the differences between the frequencies of a linguistic variable can be generalized to the population.
15. Use the Graph tool from the Lancaster Stats Tools online and the data provided there to create graphs visualizing the main patterns in those datasets.

THINGS TO REMEMBER

- Corpus linguistics is a scientific method.
- Successful application of statistical techniques in corpus linguistics depends on the use of a well-constructed unbiased corpus.
- Statistics uses mathematical expressions to help us make sense of quantitative data.
- Effective visualization summarizes patterns in data without hiding important features.
- Although most visible, p-values form only a (small) part of statistics.
- 'Statistical significance', 'practical importance' and 'linguistic meaningfulness' are three separate dimensions which shouldn't be confused.

Advanced Reading

- Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4), 243–57.
- Biber, D. & Jones, K. (2009). Quantitative methods in corpus linguistics. In A. Lüdeling & M. Kytö (eds.), *Corpus linguistics: an international handbook*, vol. 2, pp. 1287–1304. Berlin: Walter de Gruyter.
- Cumming, G. (2012). *Understanding the new statistics: effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.
- Diggle, P. J. & Chetwynd, A. G. (2011). *Statistics and scientific method: an introduction for students and researchers*. Oxford University Press.

- Gries, S. Th. (2006). Some proposals towards a more rigorous corpus linguistics. *Zeitschrift für Anglistik und Amerikanistik*, 54(2), 191–202.
- Leech, G. (1992). Corpora and theories of linguistic performance. In J. Svartvik (ed.), *Directions in corpus linguistics*, pp. 105–22. Berlin: Mouton de Gruyter.
- McEnery, T. & Hardie, A. (2011). *Corpus linguistics: method, theory and practice*. Cambridge University Press.
- Okasha, S. (2002). *Philosophy of science: a very short introduction*. Oxford University Press.
- Salsburg, D. (2001). *The lady tasting tea: how statistics revolutionized science in the twentieth century*. London: Macmillan.
- Tufte, E. (2006). *Beautiful evidence*. Cheshire, CT: Graphics Press.
- Vickers, A. (2010). *What is a p-value anyway? 34 stories to help you actually understand statistics*. Boston: Addison-Wesley.
- Yau, N. (2011) *Visualize this: the Flowing Data guide to design, visualization, and statistics*. Indianapolis: Wiley.

Companion Website: Lancaster Stats Tools Online

1. All analyses shown in this chapter can be reproduced using the statistical tools available from the companion website. The tools available for this chapter include:
 - Stats calculator
 - Randomizer
 - Graph tool
2. The website also offers additional materials for students and teachers.

2 Vocabulary

Frequency, Dispersion and Diversity

2.1 What Is This Chapter About?

Corpora consist of a large number of words – thousands, millions or even billions of words. Corpora, however, are not mere bags of words from which we can pull one word after another like a magician pulling rabbits from a hat. On the contrary, when analysing a corpus we need to think carefully about word frequencies and distributions in order to find meaningful patterns of language use. In this chapter, we'll be looking at simple statistical measures that will help us describe the occurrence of words in texts and corpora. We'll be exploring answers to five main questions:

- What is a word? (Section 2.2)
- How do we measure frequencies of words and phrases? (Section 2.3)
- How do we measure distributions of words and phrases? (Sections 2.4 and 2.5)
- How do we measure lexical diversity? (Section 2.6)
- How can these measures be used in linguistic and social research? (Section 2.7)

2.2 Tokens, Types, Lemmas and Lexemes

Think about ...

Before reading this section, think about the following questions:

1. How would you define a word?
2. How many words do the following two sentences consist of? *Took them 26 years to win the title. During that time we won it 3 times despite being in the second division for half that time.*

In corpus linguistics we deal with words and phrases. In statistical terminology, these are the **linguistic variables** that we want to investigate (see Section 1.3). To put it simply, in corpus linguistics we often count words in different contexts and then compare these counts to identify patterns of language use (there are many

examples of this type of ‘pattern spotting’ in this book). In order to count words reliably we, of course, need to know what we are counting, and therefore we first have to define a word. This might look fairly trivial (we all know what a word is), but in fact the definition of a word is quite a complex issue. This can be demonstrated with the example sentences from the ‘Think about’ task.

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
Took them 26 years to win the title. During that time we won it 3 times despite

18 19 20 21 22 23 24 25 26
being in the second division for half that time. (source: BNC, J1H)

Note: The numbers in the example above indicate individual word forms separated by space.

How many words do the two sentences contain? If you ask a corpus linguist, the answer to the question above will most likely be as follows: overall, there are 26 tokens, 23 types, 21 lemmas and 22 lexemes. Let’s try to unpack this information: a **token (running word)** is a single occurrence of a word form in the text. Each time we encounter a string of letters or numbers separated by white space (or punctuation) we count this as one token. When students ask questions such as ‘how many words do I need to write for this essay topic?’ they are asking about the token count (usually without ever having heard of the term ‘token’). When we count every single occurrence of a word form in the two sentences from the ‘Think about’ task, we get the number 26. It is important to note that the exact operationalization of what counts as a token differs in different implementations of the basic definition of ‘token’ (see above) in specific corpus analysis tools. This presents a real problem for the reliability and replicability of results, because the token counts for the same corpus such as the *British National Corpus* can differ by up to 17 per cent¹ depending on the tool used (Brezina & Timperley 2017). When using corpus tools, we should therefore be very clear about how exactly the token definition is implemented and whether the counts (especially when based on different tools) are comparable because token counts have a major effect on all statistical measures based on word/phrase frequency (e.g. relative frequencies, keyword measures, collocation measures, statistical tests using relative frequencies etc.).

A **type** is a unique word form in the corpus. When we ask about word types we are asking about how many different word forms there are in the text/corpus. In the two sentences from the ‘Think about’ task, three forms – *the* (7, 20), *that* (10, 25) and *time* (11, 26) – occur twice each (note that *times* is not counted

¹ Some tools (e.g. CQPweb, Sketch Engine) include punctuation in token counts. Others (e.g. #LancsBox) stay closer to the simple (surface) definition of the ‘token’ as presented in this book and do not count punctuation. In addition to punctuation, other sources of variation in token counting include: treatment of clitics (e.g. *'ll* in *he'll*) and hyphenated words (*well-known*), tokenization decisions based on morphological analysis (part-of-speech tagging) as well as tokenization decisions for non-segmented languages (e.g. Chinese).

together with *time* under this definition); each of these is counted as only one type. We are therefore left with 23 word types.

Both word tokens and word types are identified based on the form of a word (external appearance, if you like). To identify *lemmas* and *lexemes*, we first need to perform linguistic analysis of the text; lemmas are based on grammatical (morphological) analysis, while lexemes are based on both grammatical and semantic analysis. A **lemma** is a group of all inflectional forms related to one stem that belong to the same word class (Kučera & Francis 1967: 1).² This morphological definition sounds complicated, but the principle is very simple. We group together forms that have the same base and differ only with respect to grammar – such as the singular and the plural forms of the same noun, the present and the past tense of the same verb, the positive and the superlative form of the same adjective, etc. A **lexeme**³ is a lemma with a particular meaning attached to it, which is necessary to distinguish polysemous words (words with multiple meanings). The best way of conceptualizing a lexeme is as a subentry in a dictionary: when you open a dictionary you can see an alphabetical list of forms (headwords), many of which are split into subentries according to their meaning – these are the lexemes. Some dictionaries also list inflectional forms related to the lexeme. In the two sentences from the ‘Think about’ task *win* and *won* belong to the same lemma (because *won* is the past tense of the verb *to win*). Similarly, *time* and *times* are members of one lemma (because *times* is the plural form of *time*). We therefore get 21 as the lemma count. However, the word *time* is a polysemous word – it can mean a number of different things. In the example, it either refers to a period (11 and 26) or frequency – how often something happens (16). When we count lexemes we therefore need to distinguish between these two uses – hence there are 22 lexemes in the example.

Let us now think about how the different notions of a ‘word’ (type, lemma and lexeme) influence the kind of analysis we can carry out in corpus linguistics. Using types is the most straightforward approach, based on distinguishing different word forms regardless of their grammatical function or meaning. However, while type is a very useful category, it may obscure some meaningful differences, e.g. uses of the form *clean* as an adjective (*a clean shirt*) versus as a verb (*to clean something*). On the other hand, if we want to use lemma as the unit of analysis, we need to automatically process the corpus to assign each form its part-of-speech and lump together all inflectional forms related to the same base – this involves a certain percentage of errors. Similarly,

² Arguably, this traditional definition of the ‘lemma’ is the most useful for corpus analyses; in some operationalizations, the ‘belonging to the same word class’ part of the definition is dropped. This, however, leads to fairly counterintuitive cases where e.g. *go in it took three goes* would be lumped under the same lemma as *went in they went*.

³ The term ‘lexeme’ is sometimes used interchangeably with the term ‘lemma’. Here, I follow Biber et al.’s (1999: 54) notion of lexeme as ‘a group of related forms which share the same meaning and belong to the same word class (part of speech)’.

Table 2.1 *Type, lemma and lexeme: advantages and disadvantages*

Definition of a word	Advantages	Disadvantages
Type	Low-inference category	No distinction between forms with multiple grammatical functions and/or meanings
Lemma	Distinction between forms with different grammatical functions	POS tagging and lemmatization involved (possible sources of error)
Lexeme	Most specific category; meaning distinction taken into consideration	High-inference category (possible source of error) not yet available on a fully automatic basis

identification of lexemes, although desirable, involves semantic tagging and semantic disambiguation, which again introduces a certain percentage of errors (even higher than part-of-speech, or POS, tagging) and, moreover, cannot be done fully automatically. Table 2.1 sums up the advantages and disadvantages of different concepts of a ‘word’.

Finally, since word counts (mostly token and type counts) are a part of almost every statistical equation that is discussed in this book, it is important that you have a good grasp of these definitions. The exercises at the end of this chapter (see Section 2.8) together with answers, which are provided at the companion website, will help you test your understanding of these crucial concepts.

Reporting Statistics: Tokens, Types, Lemmas and Lexemes

1. What to Report

When talking about words in corpus linguistics, we need to specify if we mean tokens (running words), types, lemmas or lexemes. When describing corpora, we should always include the information about the exact token count. Because token counts differ from tool to tool we should also provide details about the tool and/or how the token count was arrived at.

2. How to Report: Examples

- The text consisted of 120 running words and included 69 different types.
- In our study, we used the 100-million-word *British National Corpus* (exact token count: 98,313,429; BNCweb; punctuation excluded from token count)

2.3 Words in a Frequency List

Think about ...

Before reading this section, think about the following question:

What are the most frequent words in English? Try to come up with a list of the top ten most frequent words:

- 1) _____, 2) _____, 3) _____, 4) _____, 5) _____,
- 6) _____, 7) _____, 8) _____, 9) _____, 10) _____.

Table 2.2 *Top ten words in the BNC*

Rank	Word	Absolute frequency	Relative frequency per million
1	the	6,041,234	61,448.72
2	of	3,042,376	30,945.68
3	and	2,616,708	26,615.98
4	to	2,593,729	26,382.25
5	a	2,164,238	22,013.66
6	in	1,937,819	19,710.62
7	that	1,118,985	11,381.81
8	it	1,054,279	10,723.65
9	is	990,281	10,072.69
10	was	881,473	8,965.95

Look at the wordlist in Table 2.2. It represents the ten most frequent words in the *British National Corpus* (BNC), a 100-million-word⁴ balanced sample of British English. Apart from the rank, it provides two frequency measures for each item – absolute frequency and relative frequency.

Absolute (or raw) frequency (AF) is the most straightforward statistic: it is the actual count of all occurrences of a particular word in a corpus. More precisely (and to use the terminology introduced in Section 2.2), the absolute frequency is a count of all tokens in the text or corpus that belong to a particular word type. For example, you can see that the most frequent word type in Table 2.2 is the definite article *the*. If you read through every text in the BNC and count the number of occurrences of *the* (don't worry – computers can do this for us very easily in almost no time) you'll get the number 6,041,234. Absolute frequency of words is a useful measure when we look at a single corpus. For example, we can use absolute frequency to sort a wordlist to get the

⁴ The exact word count for the BNC used for the calculations of relative frequency in Table 2.2 is 98,313,429.

most frequent items at the top (as in Table 2.2). However, when we want to compare two or more corpora⁵ we use the so-called **relative (or normalized) frequency (RF)**. Relative frequency can be calculated very easily. All we need to know is the absolute frequency of a word we are interested in, and the total number of words (tokens) in the corpus. Relative frequency is calculated as follows:

$$\text{relative frequency} = \frac{\text{absolute frequency}}{\text{number of tokens in corpus}} \times \text{basis for normalization} \quad (2.1)$$

For example, the relative frequency of the definite article *the*, as shown in Table 2.2, was calculated as follows:

$$\text{relative frequency } (\textit{the}) = \frac{6,041,234}{98,313,429} \times 1,000,000 = 61,448.72 \quad (2.2)$$

In this case, we have chosen one million as the basis for normalization, which is a common baseline in corpus linguistics. This means that, on average, there are over 61,000 instances of the definite article for every million tokens in the corpus. In fact, in all corpora of written English you can expect the definite article to be at the top of the wordlist, with an absolute frequency roughly equivalent to 6% of the overall number of tokens in the corpus. The relative frequency can therefore be considered as the **mean frequency** ('mean' is a statistical term for a type of average; see Section 1.3) – that is, the mean of the frequencies of the word in hypothetical samples of x tokens from the corpus, where x is the basis for normalization (in this case one million). This idea about the mean frequency will be useful later, when we come to discuss other statistical measures.

In smaller corpora, smaller bases for normalization than one million are more appropriate, e.g. normalization per 10,000 or even 1,000 words. The reason for this is that relative frequency is used not only to compare frequencies of a particular type in two (or more) corpora, but it is also used to present evidence about the frequency of words and phrases in a form that is easier for a reader to grasp than the absolute frequency would be. If we choose a basis for normalization that is too large relative to the actual corpus size, this can 'blow up' our numbers artificially and thus effectively misrepresent the (limited) evidence we have. For example, imagine a very small corpus that contains 11,000 tokens, including the type *homeostasis* which appears only once in the corpus. If we choose one million as the basis for normalization, we will get over 90 per million words as the relative frequency of *homeostasis*. Although the proportion is mathematically correct, this is extremely misleading for the reader, given the behaviour of rare words. When a rare word such as *homeostasis* occurs just once

⁵ It is hard to find corpora that have exactly the same number of words. Even corpora compiled according to the same sampling frame (like the Brown family corpora) differ slightly in terms of their word totals. As a general rule, when comparing the frequencies of words or phrases in two or more corpora, relative frequencies need to be used.

in a small corpus, it is highly unlikely that it would occur at the mathematically equivalent 90 times in a million words. It might just as easily occur once in the hypothetical million-word corpus, or five times, or maybe not at all: the actual smaller corpus simply does not give us enough evidence to extrapolate. In this example, a more appropriate basis for normalization would therefore be 10,000 which gives us the relative frequency of approximately 0.9. It is important to stress that relative frequencies should never be used to hide the absolute frequencies but should be reported together with absolute frequencies.

It is crucial to always remember that a corpus is a sample of language (see Section 1.4). It provides evidence about the use of words and phrases; however, this evidence can be limited even in fairly large corpora. For example, in the 100-million-word BNC, over half of the types occur only once (we call these words **hapax legomena**⁶ or **hapaxes**), and an additional 13% of the types occur only twice. In fact, less than 5% of the types occur 100 times or more in the corpus (in other words, with a relative frequency of one per million or more). This distribution of words is not specific to the BNC, but is typical of all corpora. The principle of rapidly diminishing word frequency is often referred to as **Zipf's law**.

Put informally, Zipf's law tells us that when we start with the most frequent item in the wordlist (regardless of the size of the corpus), the second most frequent item will have only half of the frequency of the first item. The third most common word will have one-third of the frequency of the first item; and so on. In other words, the amount of evidence that we can get from corpora about words diminishes rapidly. This can be seen in Figure 2.1, which shows the distribution of frequencies of words according to the rank order in the BNC on a linear scale (top) and on a logarithmic scale (bottom). Displaying frequencies on the y-axis on a logarithmic scale (log scale) means that each mark on the y-axis is the previous mark multiplied by 10. This type of display shows the distribution across the large range of frequency values (5.4 M – 1) providing a better sense of when exactly the curve starts drifting off. The linear scale (top), on the other hand, shows how dramatic the drop in word frequencies in a corpus really is.

Formally, Zipf's law can be expressed as:

$$\text{absolute frequency of a word} \times \text{its rank in a wordlist} \cong \text{constant} \quad (2.3)$$

or

$$\text{absolute frequency} \cong \frac{\text{constant}}{\text{rank in a wordlist}} \quad (2.4)$$

where the *constant* is the frequency of the first item in the wordlist.

⁶ Hapax legomenon is a Greek term and can be translated as 'once said'.

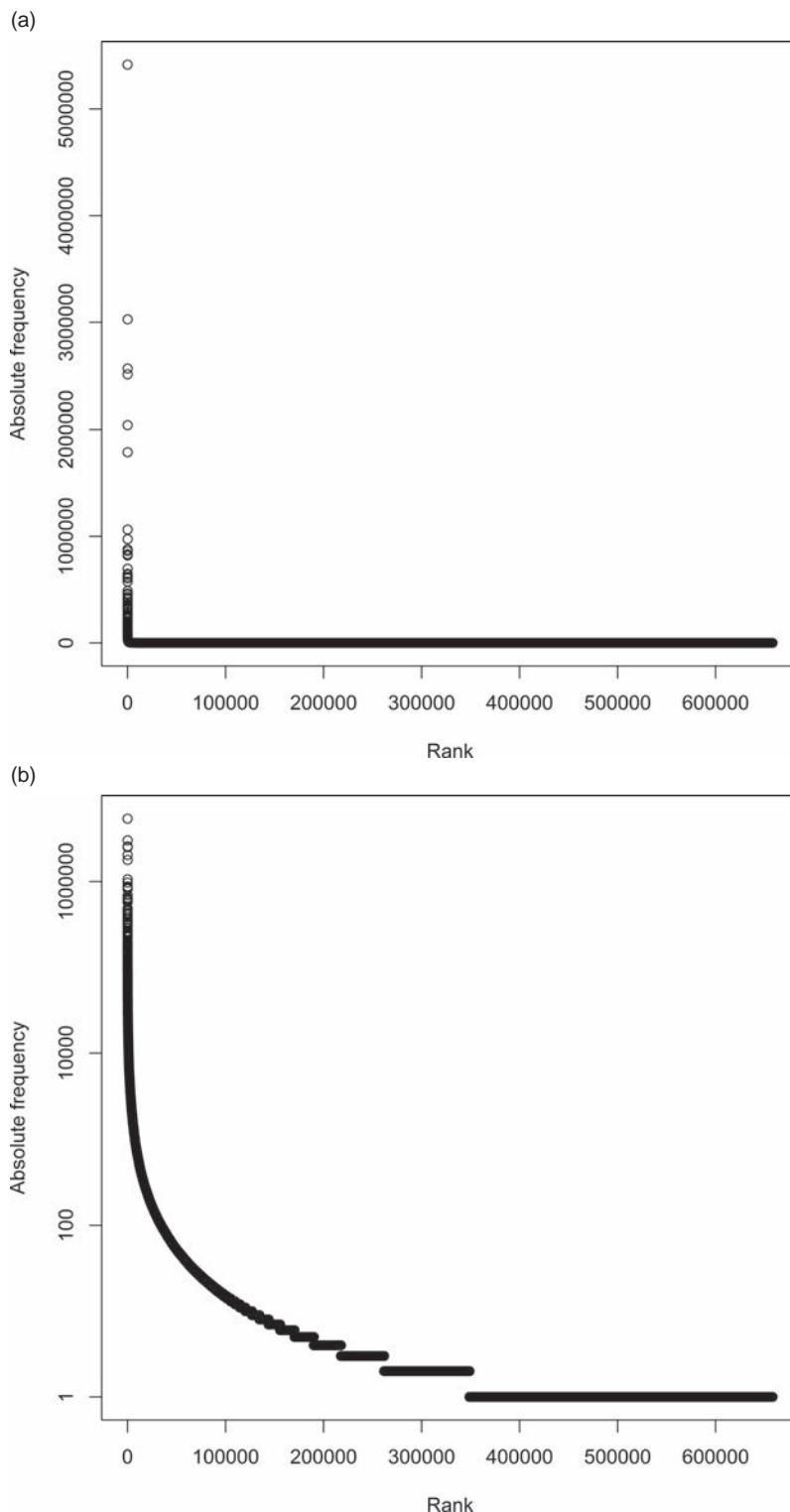


Figure 2.1 *Distribution of word frequencies in the BNC*

Note that Zipf's law represents an approximation. The actual frequencies of words in a real corpus will differ to a certain extent from those predicted by this model. The practical implications of Zipf's law for corpus linguistics are as follows: we have to learn to critically evaluate the amount of evidence we have for our claims. For hapaxes and low-frequency words, this evidence is naturally limited. To answer some research questions, we therefore need very large corpora (billions rather than millions of words), in which even fairly infrequent words occur multiple times. We also need to look for further (comparable) evidence of word behaviour from multiple sources (see Section 8.3 on meta-analysis).

Reporting Statistics: Absolute and Relative Frequencies

1. What to Report

When reporting frequency of words, we should report both the absolute (raw) and the relative frequency. The relative frequency needs to be normalized to the appropriate basis that is similar in size to the corpus or its parts (subcorpora or texts) that we are interested in.

2. How to Report: Examples

- The preposition *of* is the second most frequent item in the BNC (AF = 3,042,376, RF = 30,945.68 per million).
- The word *corpus* occurred 20 times in the text (13.3 per 1,000 words).

2.4 The Whelk Problem: Dispersion

Think about ...

Before reading this section, think about the following questions:

1. Do you know what a 'whelk' is?
2. How often do you think we use this word?

So far, we have been looking at frequencies of words. However, to fully describe a word or phrase in a corpus, we need to introduce another concept, namely dispersion. This concept is best illustrated with the so-called 'whelk problem'. This term was introduced by Kilgarriff (1997), who has pointed to unequal distribution of words in corpora. Imagine you have a corpus which contains a substantial sample

from a book on whelks (small sea creatures somewhat similar to snails). Naturally, the word *whelk* will appear many times in this book because it is about whelks. In general English, however, *whelk* is not a particularly common word because it is specific to a single genre/register – books and articles on sea life. However, here comes the problem: when we construct a frequency list based on our corpus that includes the book on whelks the word *whelk* will appear among fairly frequent items by virtue of being repeated many times in a single text. If we base our investigation on word frequency alone, our results will be extremely misleading. In addition to frequency, we therefore need to account for the dispersion – the fact that the word *whelk* occurs only in one text out of many.

Generally, **dispersion** tells us about the distribution of words or phrases throughout the corpus. For example, the definite article *the* is not only a highly frequent word, it also is fairly evenly distributed in text. This is because *the* is a grammatical word and we usually cannot put sentences together without using it. Other words which are specific to a particular context (e.g. *whelk*, *hashtag*, *corpus*) will be less evenly distributed. It is important to note that there is not a single measure of dispersion, but rather a set of different measures that can be used to investigate variation within corpora and to highlight different aspects of dispersion. At the most abstract level, dispersion directly depends on our understanding of corpora and their structure (parts) because dispersion describes distribution of words and phrases throughout the corpus or across its different parts. The decision about which dispersion measure to use is thus closely connected with the research design of our study and the design of the corpus used. Rather than engaging in a general debate about the most suitable dispersion measure in corpus research (e.g. Gries 2008, 2010; Biber et al. 2016) we need to understand the properties of the individual dispersion metrics on offer and suit these to the purposes of our research question(s) and our specific understanding of dispersion. Here, we'll consider five important dispersion measures. For an overview of a number of different dispersion measures see Gries (2008).

As an example, imagine a one-million-word corpus, which is divided into six parts of unequal size (each representing a different genre/register) as described in Table 2.3.

Table 2.3 *Example corpus: one million tokens*

	Part 1	Part 2	Part 3	Part 4	Part 5	Part 6	Whole corpus
Tokens	100,000	100,000	200,000	200,000	200,000	200,000	1,000,000
Absolute frequency of word <i>w</i>	10	4	2	0	24	10	50
Relative frequency of word <i>w</i> per 100 k	10	2	1	0	12	5	5
Includes <i>w</i> ?	YES	YES	YES	NO	YES	YES	YES

This example corpus will be used in the explanation of the different dispersion measures below.

Range₂ (R) is a very basic and fairly crude measure of dispersion.⁷ It tells us the number of corpus parts in which a word or phrase occurs, regardless of their size. The parts can be genre- or speaker-based subcorpora or even individual texts. Range is formally expressed as:

$$\text{range}_2 = \text{no. of parts with word } w \text{ (or phrase } p) \quad (2.5)$$

As we can see immediately (look at the row ‘Includes w ?’), the range for the word w in Table 2.3 is 5, because there are five parts (1, 2, 3, 5 and 6) in which the word w appears at all. This can be expressed as:

$$R(w) = 5 \quad (2.6)$$

The range₂ is sometimes also calculated as a percentage out of the total number of corpus parts:

$$\text{range}_2\% = \frac{\text{no. of parts with word } w \text{ (or phrase } p)}{\text{total no. of parts in the corpus}} \times 100 \quad (2.7)$$

Applied to the example in Table 2.3, this produces:

$$R\% (w) = \frac{5}{6} \times 100 = 83.3\% \quad (2.8)$$

We can say that the range₂ of word w in the one-million-word example corpus from Table 2.3 is over 80% because 83.3% of the corpus parts include this word.

However, range₂ is not a very good measure for quantifying the amount of dispersion across individual corpus parts because it is based on a simplistic YES/NO decision about the presence or absence of a word or phrase in each part, disregarding the actual frequencies in the different parts. Range₂ also does not take into account the size of the parts. To illustrate this, imagine a different word (w_1) which has the same absolute frequency in the example corpus as word w (i.e. 50), but a very different distribution, namely 46, 1, 1, 0, 1, and 1 in the six corpus parts. When we calculate the range₂ of w_1 , we’ll get the same number as for word w (i.e. 5 out of 6 or 83.3%), although the large majority of all occurrences of w_1 are in only one part (Part 1) whereas word w is, in comparison, more evenly spread out across the corpus. This lack of sensitivity of range₂ as a dispersion measure is its major limitation. Because of this, range₂ can be used for a first (simple) exploration of the corpus data; but for further analyses more sensitive dispersion measures are preferable.

Standard deviation is a classic measure of dispersion, which is used very often also outside corpus linguistics. It expresses how much the individual values

⁷ In statistics more generally, the term ‘range’ is also used in a different sense, to refer to the difference between the largest and the smallest values in any given dataset. This use is discussed in Section 1.3 and is indicated by the subscript 1 (range₁).

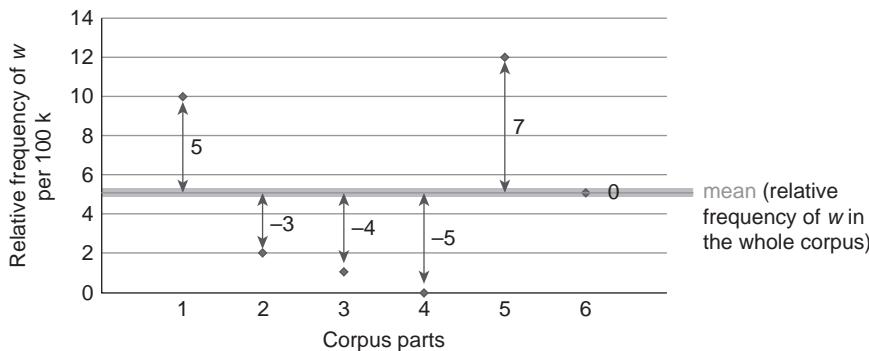


Figure 2.2 *Example corpus: calculation of SD*

in a dataset (here, the relative frequencies of *w* in the individual parts of the example corpus from Table 2.3) vary around the mean relative frequency. This becomes clear if we display the relative frequency values from Table 2.3 in graphical form.

In Figure 2.2, we can see that, with the exception of the value for Part 6, each of the values (relative frequencies) for Parts 1–5 is situated at a certain distance from the mean (the average value of *w* in the whole corpus). The individual distances from the mean are +5, −3, −4, −5, +7 and 0 (note that negative distances signify values smaller than the mean). The standard deviation is calculated based on these distances from the mean. In essence, the question which we are asking when calculating standard deviation is: how much variation around the mean can we observe in the data? Mathematically, standard deviation is expressed as:

$$\text{standard deviation}_{\text{population}} = \sqrt{\frac{\text{sum of squared distances from the mean}}{\text{total no. of corpus parts}}} \quad (2.9)$$

The superscript reminds us that we are looking at a simple form of standard deviation calculated assuming that we are dealing with the whole population. This form of standard deviation (SD_p or σ [sigma]) differs slightly from the sample standard deviation (see below). For the data in Table 2.3 SD_p is calculated as:

$$\begin{aligned} SD_p(w) &= \sqrt{\frac{(10-5)^2 + (2-5)^2 + (1-5)^2 + (0-5)^2 + (12-5)^2 + (5-5)^2}{6}} \\ &= \sqrt{\frac{25+9+16+25+49+0}{6}} = 4.55 \end{aligned} \quad (2.10)$$

A brief explanation of the procedure: squaring the distances (2) and then taking the square root ($\sqrt{}$) of the final number might seem like an unnecessarily

complicated way of calculating the variation in the data. The reason for doing this is that, as we noted above, some of the values are above the mean (and therefore the distance is a positive number), while other values are below the mean (the distance is a negative number). If we just added the distances, the positive and negative values would cancel each other out ($5 - 3 - 4 - 5 + 7 + 0 = 0$). Since squaring a negative number makes it positive, the square-and-then-square-root procedure built into the standard deviation gets us around this problem.

There is another (slightly modified) way of calculating the standard deviation which we will be using when we consider inferential statistics, rather than the descriptive statistics we are discussing here. This is so-called **sample standard deviation (SD)**. It is calculated in almost the same way as the standard deviation described in equation (2.9) above. The difference is that the sum of squared distances is divided not by the total number of corpus parts but by the total number of corpus parts minus 1 (the reason for this is connected with the notion of ‘degrees of freedom’ explained in Section 6.3).

$$\text{standard deviation}_{\text{sample}} = \sqrt{\frac{\text{sum of squared distances from the mean}}{\text{total no. of corpus parts} - 1}} \quad (2.11)$$

For the purposes of describing dispersion in a corpus the basic version of standard deviation from equation (2.9) should be used. Standard deviation is a useful measure when we want to see how homogeneous or heterogeneous the distribution of a word is. Standard deviation always needs to be considered in relation to the mean (the relative frequency of the word in the corpus overall). In our example from Figure 2.2, the *SD* (4.55) was almost as large as the mean (5). This indicates a large amount of variation, with many of the individual values appearing at a considerable distance from the mean.

Because the *SD* needs to be considered in relation to the mean, we cannot use this measure to compare the dispersions of different words (or phrases) that occur with different frequencies. In these cases, other measures of dispersion such as the coefficient of variation, Julland’s *D* or *DP* (see below) are more appropriate.

The **coefficient of variation (CV)** describes the amount of variation relative to the mean relative frequency of a word or phrase in the corpus. The more variation in the frequencies of a word/phrase in the individual parts there is, the more uneven the dispersion. The equation for *CV* is very simple:

$$\text{Coefficient of variation} = \frac{\text{standard deviation}}{\text{mean}} \quad (2.12)$$

For word *w* in the example corpus in Table 2.3 the calculation is as follows:

$$CV(w) = \frac{4.55}{5} = 0.91 \quad (2.13)$$

The coefficient of variation is a standardized measure; this means that it can be compared across different words and phrases in one corpus. The closer the

coefficient is to zero, the more even the distribution of the word or phrase is. The maximum value of the coefficient of variation depends on the number of parts in the corpus, and is equal to the square root of the number of parts minus 1 ($\sqrt{\text{no. of corpus parts} - 1}$).

Sometimes, the coefficient of variation is multiplied by 100 and presented as a percentage of variation. This is, however, problematic, because the coefficient of variation can be greater than 1 (when the SD is greater than the mean) which would give a percentage higher than 100. However, a true conversion to percentage can be achieved by considering the maximum possible variation in a given corpus. We know that the maximum value of CV depends on the number of corpus parts and can be calculated as $\sqrt{\text{no. of corpus parts} - 1}$. The following equation thus converts a CV to a percentage out of the maximum possible observed variation.

$$\text{Coefficient of variation \%} = \frac{\text{Coefficient of variation}}{\sqrt{\text{no. of corpus parts} - 1}} \times 100 \quad (2.14)$$

When we apply this to word w in the example corpus from Table 2.3, we get:

$$CV\% = \frac{0.91}{\sqrt{6 - 1}} \times 100 = 40.7\% \quad (2.15)$$

This means that the dispersion of w is less than 50% of the maximum possible variation. The maximum level of variation would be reached if the word occurred only in one part of the corpus.

Julland's D is a measure of dispersion that builds on the coefficient of variation. It is a number between 0 and 1, with 0 signifying extremely uneven distribution and 1 perfectly even distribution. Julland's D was originally developed for use in frequency dictionaries (Julland & Chang-Rodriguez 1964; Julland et al. 1970; Leech et al. 2001; Davies & Gardner 2010).

In essence, Julland's D is an inverse number to CV (and $CV\%$). While CV tells us about the amount of variation in the corpus (larger CV means more variation in the frequencies), Julland's D tells us about homogeneity of the distribution (larger Julland's D means a more even distribution and less variation). The following formula is used to calculate Julland's D :

$$\text{Julland's } D = 1 - \frac{\text{Coefficient of variation}}{\sqrt{\text{no. of corpus parts} - 1}} \quad (2.16)$$

For the example from Table 2.3, Julland's D will be calculated as follows:

$$\text{Julland's } D = 1 - \frac{0.91}{\sqrt{6 - 1}} = 0.59 \quad (2.17)$$

This value (0.59) shows an uneven distribution, which, however, is closer to 1 (perfectly even distribution) than to 0. Julland's D has been criticized in the literature (Gries 2008) for returning values that lie outside the expected range

(0–1). This criticism was, however, based on a different formula from that presented in this chapter.⁸ Another potential problem with Juilland's *D* was pointed out by Biber et al. (2016), who show that this measure is highly dependent on the number of corpus parts; with a large number of corpus parts (e.g. 1,000) the practical range of Juilland's *D* becomes very restricted because, as is implied in the formula, the actual variation in the corpus is measured against the maximum possible theoretical variation in the corpus given the number of corpus parts – the more parts the corpus has, the larger the theoretical variation is ($\sqrt{\text{no. of corpus parts} - 1}$).

As an alternative to Juilland's *D*, the last measure discussed in this section is *DP*. ***DP* (Deviation of Proportions)** is a measure proposed by Gries (2008) which compares the expected distribution of a word or phrase in different corpus parts with the actual distribution. It is a number between 0 and 1, with 0 signifying perfectly even distribution and 1 extremely uneven distribution. Note that this scale is the reverse of the Juilland's *D* scale, where 0 signifies extremely uneven distribution. *DP* is calculated as follows:

$$DP = \frac{\text{Sum of absolute values of (observed} - \text{expected proportions)}}{2} \quad (2.18)$$

The expected proportions are calculated by taking one-by-one the sizes of the corpus parts (number of tokens) and dividing them by the total number of tokens in the corpus; this is to establish their proportional contribution to the overall size of the corpus. The assumption is that if a word or phrase is evenly distributed in the corpus it should follow the proportional distribution calculated in this step, hence the expected (or baseline) distribution. The observed proportions are calculated by taking, again one-by-one, the absolute frequencies of the word or phrase of interest in the corpus parts and dividing these by the absolute frequency of the word or phrase in the whole corpus. This is done to establish how much proportionally each part of the corpus contributes to the overall frequency of the word or phrase. By comparing the observed and the expected proportions (taking absolute values of the difference) and putting the differences together we get the *DP* measure. For example, *DP* for the values in Table 2.3 is calculated as shown in Table 2.4.

$$DP = \frac{0.1 + 0.02 + 0.16 + 0.2 + 0.28 + 0}{2} = 0.38 \quad (2.19)$$

This value (0.38) indicates an uneven distribution, which, however, is closer to 0 (perfectly even distribution) than to 1. In this case, *DP* provides a similar picture to Juilland's *D*.

⁸ Gries (2008) uses the sample standard deviation (SD_{sample}) – see equation (2.11) – to calculate the coefficient of variation for Juilland's *D* formula (equation (2.16)). To avoid the problems pointed out in Gries's article, the simple descriptive population standard deviation (see equation (2.9)) should be used.

Table 2.4 Calculation of DP with the example corpus

	Part 1	Part 2	Part 3	Part 4	Part 5	Part 6	Whole corpus
Tokens	100,000	100,000	200,000	200,000	200,000	200,000	1,000,000
Absolute frequency of word w	10	4	2	0	24	10	50
Expected proportion	$\frac{100k}{1M} = 0.1$	$\frac{100k}{1M} = 0.1$	$\frac{200k}{1M} = 0.2$	$\frac{200k}{1M} = 0.2$	$\frac{200k}{1M} = 0.2$	$\frac{200k}{1M} = 0.2$	1
Observed proportion	$\frac{10}{50} = 0.2$	$\frac{4}{50} = 0.08$	$\frac{2}{50} = 0.04$	$\frac{0}{50} = 0$	$\frac{24}{50} = 0.48$	$\frac{10}{50} = 0.2$	1
Absolute differences	0.1	0.02	0.16	0.2	0.28	0	0.76

Reporting Statistics: Dispersion Measures

1. What to Report

When reporting frequencies in corpora we should also include the information about dispersion. The decision about which dispersion measure to report should be motivated by the aims of the research – what aspect of dispersion is necessary for us to be able to interpret the frequency correctly. The main options (there are others) include range (R), percentage range ($R\%$), population standard deviation (SD_p), sample standard deviation (SD), coefficient of variation (CV), percentage coefficient of variation ($CV\%$), Julland's D and DP .

2. How to Report: Examples

- The word *corpus* occurs 773 times in the BNC (6.9 per million) but only in 201 texts ($R\% = 5\%$).
- The definite article *the* is the most frequent word (type) in the BE06 corpus occurring in the texts with the mean relative frequency of 51.64 per 1,000 ($SD = 14.17$).
- Swear words are unequally distributed in the BNC64 corpus. For instance, *fuck* occurs 123 times in only 14 out of 64 speaker samples with $DP = 0.85$.

2.5 Which Words Are Important? Average Reduced Frequency

Think about ...

Before reading this section, think about the following question:

How would you find out which words in English are important to know?

Arguably, words that occur frequently across a large number of contexts are important because they are very likely to be encountered in a variety of communicative situations. For example, for learners of a language it is crucial to know which lexical items they should learn first. These are often not concrete words (as we might wrongly assume) but general abstract terms. For instance, the most widely used noun in English is *time*, which can be found frequently across different contexts in both speech and writing: we speak (and write) about time all the time. Recently, Brezina & Gablasová (2015) have carried out research into the most important words in English based on a variety of English language corpora of the total size of over 12 billion running words. The primary metric used in this study was the average reduced frequency.

Average reduced frequency (ARF) is a measure that combines frequency and dispersion (Savický & Hlaváčová 2002). The idea behind the ARF is to produce a usage coefficient evaluating the prominence of words in a corpus in terms of both their frequency and their dispersion: the more frequent and evenly distributed the word is, the more prominent it is considered to be. One of the advantages of ARF is that it does not depend on the corpus being physically divided into different parts (subcorpora). To calculate the ARF we need the following three pieces of information:

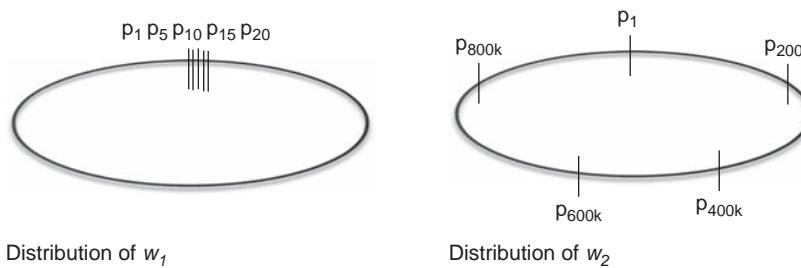
- Absolute frequency of the word
- Corpus size (total number of tokens)
- Positions of the word in the corpus

The absolute frequency and the total number of tokens were discussed in previous sections (see Sections 2.2 and 2.3). The positions of the word in the corpus are represented by numbers expressing the order in which the word appears in the corpus; these are calculated automatically by corpus analysis tools. Imagine that you begin reading through the corpus (one text after another) and assign each token a number starting from 1. Whenever you encounter the type that you are looking for, you note down the number of that token (i.e. its position in the corpus). You then continue reading until you encounter the word again, and also make a note of this position; and so on. From the individual positions of the words in the corpus, we can then calculate the distances between individual occurrences of the word in the corpus. These distances are then used in the formula for the ARF, which is as follows:

$$\text{ARF} = \frac{1}{v} \times (\min(\text{distance}_1, v) + \min(\text{distance}_2, v) + \min(\text{distance}_3, v) \dots)$$

$$\text{where } v = \frac{\text{total corpus tokens}}{\text{absolute frequency of word}}$$

and $\min(\text{distance}_n, v)$ signifies the smaller of two values: (1) value of the distance between two occurrences of the word or (2) value of v . This is done for all of the distances between two occurrences of the word in the corpus. (2.20)

Figure 2.3 *Distribution of words w_1 and w_2*

Although the equation looks complicated, the idea behind ARF is fairly simple: we notionally divide (this is not done physically but as part of the calculation) the corpus into x parts of the same size (v). The number x is the absolute frequency of the word we are interested in. This means that the number of the notional parts, and their length (v in equation (2.20)), depend on the frequency of the word in the corpus. Then we simply count the number of notional parts that include the word that we are interested in. We call this count the ‘reduced frequency’. The purpose of this exercise is to disregard occurrences of a word which are close to each other (i.e. fall within the same part) and count them only once. Then, in addition, to make this procedure robust, we repeat the process for every possible beginning-point in the corpus (think about the corpus as a circle rather than a line) and calculate the mean of all the reduced frequency values that we get via the procedure described above. This mean value is the value of the average reduced frequency.

Let’s demonstrate this with the following example: imagine a one-million-word corpus in which we search for two words w_1 and w_2 . Both words occur five times in the corpus, i.e. with absolute frequency of 5. The first word (w_1) occurs only in one text, at corpus positions 1 – 5 – 10 – 15 and 20, while the second word (w_2) is evenly distributed at positions 1 – 200,000 – 400,000 – 600,000 and 800,000. Note that ‘p’ in Figure 2.3 refers to the position of the word in the corpus.

The ARF for w_1 is calculated as follows:

First we establish the length of the notional parts, v

$$v = \frac{1,000,000}{5} = 200,000$$

Then we calculate the distances between the individual occurrences of the word (w_1) in the corpus – to do this we need to use the corpus positions. Note especially how distance₁ is calculated: it is the distance between the last and the first occurrence of w_1 in the corpus.

- distance₁ = 1st occurrence + (total corpus tokens – last occurrence) = 1 + (1,000,000 – 20) = 999,981
- distance₂ = 2nd occurrence – 1st occurrence = 5 – 1 = 4
- distance₃ = 3rd occurrence – 2nd occurrence = 10 – 5 = 5

- $\text{distance}_4 = 4\text{th occurrence} - 3\text{rd occurrence} = 15 - 10 = 5$
- $\text{distance}_5 = \text{last occurrence} - 4\text{th occurrence} = 20 - 15 = 5$

Finally, all the terms are inserted into the ARF equation:

$$\begin{aligned}\text{ARF}(w_1) &= \frac{1}{v} \times (\min(\text{distance}_1, v) + \min(\text{distance}_2, v) + \min(\text{distance}_3, v) \dots) \\ \text{ARF}(w_1) &= \frac{1}{200,000} \times (\min(999,981, 200,000) + \min(4, 200,000) \\ &\quad + \min(5, 200,000) + \min(5, 200,000) + \min(5, 200,000)) \\ &= 0.000005 \times (200,000 + 4 + 5 + 5 + 5) = 1.000095 \quad (2.21)\end{aligned}$$

As we can see from the calculations above, the ARF for word w_1 is approximately 1. This should be interpreted as follows: because the five occurrences of w_1 are all very close to each other, they should be counted as if they were only one occurrence. On the other hand, if we calculate the ARF for w_2 , we'll get a number close to 5. This means that because w_2 is evenly distributed throughout the corpus, its frequency should not be reduced – we should continue to consider each instance as a separate occurrence. Here's the mathematics:

$$\begin{aligned}\text{ARF}(w_2) &= \frac{1}{200,000} \times (\min(200,001, 200,000) + \min(199,999, 200,000) \\ &\quad + \min(200,000, 200,000) + \min(200,000, 200,000) \\ &\quad + \min(200,000, 200,000)) \\ &= 0.000005 \times (200,000 + 199,999 + 200,000 + 200,000 \\ &\quad + 200,000) = 4.999995 \quad (2.22)\end{aligned}$$

If you find it difficult to get your head around the details of this process, don't worry: the ARF is designed to be calculated automatically by a computer (see the ARF calculator in Lancaster Stats Tools online). The purpose of explaining the ARF in this section is to help you understand the general principles of this powerful measure.

Reporting Statistics: ARF

1. What to Report

ARF can be reported in addition to absolute and relative frequencies of a word. It can be used to rank-order words in a frequency list to highlight the most frequent and evenly dispersed items.

2. How to Report: An Example

- The following table shows the top four lemmas in the BNC ranked according to ARF; the absolute frequency figure is also provided.

BNC	Average reduced frequency (ARF)	Absolute frequency (AF)
the-article	3,839,770	6,050,229
be-verb	2,702,664	4,119,048
of-preposition	1,838,624	3,010,276
and-conjunction	1,671,566	2,615,148

2.6 Lexical Diversity: Type/Token Ratio (TTR), STTR and MATTR

Think about ...

Before reading this section, think about the following question:

Which of these two short texts is more lexically diverse?

Text A [BNC: KB7]	Text B [BNC: B25]
You want a cup of tea?	All sciences – physics, agriculture, medicine and even sociology – go beyond the mere solution of
Yeah I'm gonna put the kettle on. Yeah alright.	immediate problems, whether these problems are of a ‘pure’ intellectual type, or an ‘applied’ practical sort.
Yeah make us a cup of coffee Stuart.	
Well we'll probably have our dinner first then I'll probably do it.	

When looking at texts and corpora we can think about how different words (types) are used to communicate meanings. Some words (especially grammatical words) are often repeated, others are used only a few times. To measure whether overall a text or corpus uses a wide range of vocabulary or only a limited range of lexical items which get recycled, we can calculate a lexical diversity statistic (Jarvis 2013).

The simplest lexical diversity statistic is the type/token ratio (see Section 2.2 for the definition of types and tokens). **Type/token ratio (TTR)** expresses the proportion of types (different word forms) relative to the proportion of tokens (running words). The idea is that a larger number of different word forms (types) relative to the number of all words in text (tokens) points to a lexically more varied text. Type/token ratio is calculated as follows:

$$\text{type/token ratio} = \frac{\text{no. of types in text or corpus}}{\text{no. of tokens in text or corpus}} \quad (2.23)$$

For the two texts from the ‘Think about’ task the type/token ratio is 0.8 (28/35) and 0.93 (28/30) respectively. This shows that text B (academic text) is more lexically diverse than text A (informal speech). This is a valid comparison because the texts are of comparable size (have a similar number of tokens). However, we have to remember that the type/token ratio is very sensitive to the length of the text; it decreases as the text becomes longer and more words get used again (recycled). The simple type/token ratio from equation (2.23) can therefore be used only for comparison of texts of the same length. As has been shown in the literature (Covington & McFall 2010; Tweedie & Baayen 1998), no simple transformation of TTR is possible that could make TTR immune to the fluctuation due to text length, although a number of measures (e.g. Guiraud’s index, Yule’s K etc.) have been proposed that claim to do exactly that. This means that there is no magic formula which would ‘normalize’ TTR and make it comparable across texts of different lengths. Instead, different measures based on taking the average value of a set of type/token ratios from standard-length text samples have been proposed: STTR (standardized type/token ratio) and MATTR (moving average type/token ratio).

Standardized type/token ratio (STTR) is a label used by Scott (2004) to refer to a measure that is otherwise known as mean segmental type/token ratio (MSTTR) (Malvern & Richards 2002). Here, the STTR label will be used because it is known in corpus linguistics due to *WordSmith Tools* (v. 4.0), which implemented this measure under the STTR label. The calculation of STTR is very simple: we divide text into standard-size segments (e.g. 1000 words), calculate the TRR for each segment and then take the mean value of the TTRs. Because most texts do not divide exactly into standard-size segments, the last segment which is shorter than the standard size is excluded from the calculations. For shorter texts, smaller standard size of the segment has to be chosen; however, very short segments (smaller than 100 tokens) have been reported to distort the results (Malvern & Richards 2002). **Moving average type/token ratio (MATTR)** introduced by Covington & McFall (2010) is somewhat similar to STTR in that it calculates the average of TTRs in same-size segments (to avoid the problems with the dependence of TTR on the text length). However, instead of dividing the text into successive non-overlapping segments, MATTR uses an overlapping window smoothly moving through the text; for each window position the TTR of the text inside the window is calculated and then the mean value of the TTRs obtained in this way is computed. MATTR is thus a more robust measure of lexical richness than STTR because it takes into account all possible segmentations of the text.

Reporting Statistics: TTR, STTR and MATTR

1. What to Report

When reporting different versions of the type/token ratio it is important to also report the parameters influencing the outcome. For TTR the text length needs to be reported; for STTR and MATTR the standard segment size and the window size respectively need to be reported.

2. How to Report: Examples

- Simple type/token ratio (TTR) was used to compare the texts because they were of the same length (2,000 tokens). The TTR values are as follows: 0.36 (text 1), 0.33 (text 2) and 0.39 (text 3).
- MATTR (window size: 100) of Dickens's *Christmas Carol* is 0.67.

2.7 Application and Further Examples: Do the British Talk about Weather All the Time?

In this section, we'll go through an example of a research project applying the statistical procedures that were introduced in this chapter. Imagine that our task is to investigate typical topics (content words) that are frequently discussed in British English. This can give us an insight into the general public discourse practices of British society. In particular, we want to find out if *weather* is one of the prominent topics typical of British culture (as the stereotype would suggest). The corpus we want to use is BE06, a corpus of current written British English,⁹ containing samples from 15 different genres and consisting of approximately one million words. The composition of the corpus is summed up in Table 2.5.

Before looking at the data, we can already start planning our study. We know that in written English the most frequent item is almost always the definite article, which represents approximately 6% of all tokens (see Section 2.3). We can therefore fairly safely predict that there will be around 60,000 instances of the definite article in the corpus. Let's also assume that we have decided that, to make properly supported observations about British culture, we need to analyse words that occur at least 30 times in the corpus to avoid claims based on low-frequency words and hapaxes. We have chosen the minimum frequency to be 30 because there are 15 parts of the corpus and we want each word to have a chance to appear

⁹ Note that a written corpus is used here to enable us to investigate variation among a range of genres/register. This study can be supplemented by findings from a corpus of spoken language such the Spoken BNC or the Spoken BNC 2014 corpus (Love et al. 2017).

Table 2.5 *BE06*

Size (tokens)	Time period	Structure: 15 genres of written English
approx. 1 M	around 2006	(1) Press: reportage, (2) Press: editorial, (3) Press: reviews, (4) Religion, (5) Skills, trades and hobbies, (6) Popular lore, (7) Belles lettres, biography, essays, (8) Miscellaneous (government documents, foundation reports, industry reports etc.), (9) Learned and scientific writings, (10) General fiction, (11) Mystery and detective fiction, (12) Science fiction, (13) Adventure and western fiction, (14) Romance and love story, (15) Humour

at least twice in each section (see Section 1.4 on corpus sampling). We can apply Zipf's law to establish approximately how many items satisfying this criterion we can expect to find in the corpus. The question we are asking is this: given the fact that the constant (i.e. the frequency of the top item on the frequency list) is approximately 60,000 words and the cut-off frequency is 29 (i.e. we will not take into consideration words with a frequency of 29 or below), what is the approximate rank of the words in the wordlist with the cut-off frequency?

When we transform the Zipf's law equation (2.4) to express the rank, we get the following calculation:

$$\text{rank} \cong \frac{\text{constant}}{\text{absolute frequency}} = \frac{60,000}{29} \cong 2,068 \quad (2.24)$$

This means that we can expect approximately 2,000 or 3,000 items to meet the criterion for inclusion in our analysis. Fortunately, this seems to be a high enough number (that is, there will be something to analyse even if some of these items are culturally uninteresting grammatical words) and we can therefore continue to plan our research.

The next step is to decide what we actually want to analyse. We have seen that there are different definitions of a 'word' (see Section 2.2) and therefore we need to make a decision about the definition that we will be using. Our options are type, lemma or lexeme. After considering the pros and cons, let's assume we have decided to work with lemmas in our research. Lemmas are usually a good choice if we are interested in culture or discourse, since the distinctions that get submerged are the inflectional differences which are generally not relevant for this kind of research. At this stage, we can start analysing the data. The frequency list based on lemmas confirms our initial assumption (based on Zipf's law) about the number of words with the frequency of 30 and over: there are 3,196 such lemmas in the corpus. In fact, the actual number of lemmas is higher than expected, which shows that Zipf's law provides only a rough estimate of word frequency distributions.

Table 2.6 Weather-related lemmas in BE06

Weather-related lemmas	AF	Range	Juilland's <i>D</i>
cloud	40	8	0.7
cold	120	11	0.8
flood	54	10	0.2
heat	104	12	0.7
hot	108	12	0.8
ice	63	13	0.7
rain	37	12	0.8
storm	37	12	0.8
sun	71	9	0.7
temperature	66	11	0.7
warm	90	13	0.8
weather	49	10	0.7
wind	78	15	0.8

Looking now more closely at the list of lemmas, we note that among the 3,196 lemmas meeting our inclusion criterion there are 13 lemmas related to weather (this analysis had to be done manually by going through the lemma list): *cloud*, *cold*, *flood*, *heat*, *hot*, *ice*, *rain*, *storm*, *sun*, *temperature*, *warm*, *weather* and *wind*. As can be seen from Table 2.6, these lemmas appear in the corpus with absolute frequencies ranging from 37 to 120. However, as we know, frequency information always needs to be considered in combination with dispersion. Table 2.6 therefore also provides two dispersion measures for the 13 weather-related items: range and Juilland's *D*.

As we can see, most of the items are fairly evenly distributed in the corpus with Range values 8 to 15 and Juilland's *D* values 0.7 or 0.8. Nevertheless, there is one exception – the lemma *flood*. Although this lemma occurs in a majority of corpus parts (Range 10) it has a very low Juilland's *D* (0.2). This signifies a very uneven distribution in the 15 corpus genres. Indeed, a closer inspection of this word confirms that *flood* occurs mainly in one genre-based part of the corpus representing the official documents. In fact, the majority of occurrences (38 out of 54) come from a single parliamentary report on floods.

After this initial exploration, to answer the research question about the prominent topics in British public discourse we need to look at the weather terms in the context of other words (lemmas) in the corpus. For this, we need to sort the lemmas according to their prominence in written British English and consider the position of the weather terms in this ordered list. We can calculate the average reduced frequency (ARF), which combines frequency and dispersion, and order lemmas according to their ARF values. Table 2.7 shows the 13 weather-related lemmas and their ARF-based ranks in the list of 3,196 lemmas.

Table 2.7 *Ranks of weather-related lemmas in BE06*

Weather-related lemmas	ARF	ARF-based rank
hot	53.8	835
cold	48.3	940
warm	45	1,005
heat	38.2	1,195
wind	34.6	1,333
sun	30.1	1,527
weather	24.2	1,867
ice	23.1	1,941
temperature	21.7	2,048
storm	20.8	2,128
rain	20.2	2,166
cloud	14.8	2,767
flood	14.1	2,888

We can see that according to the ARF, two weather-related lemmas belong among the first thousand items in the English language; six among the second thousand and the rest (five) belong among the third thousand most frequent items. We can therefore say that weather as a topic is prominently represented in British discourse, although it is not the most widely discussed subject. To find out what the most widely discussed subjects are in British written discourse, see the companion website.

2.8 Exercises

1. Look at sentences (a)–(d) below and count the number of *tokens*, *types*, *lemmas* and *lexemes* in each.
 - (a) The City is braced for far worse figures to come in the coming months, unless the Government recovery package produces a startling turn round in optimism.
(source: BNC, CEN)
 - (b) Of 354 fifth- and sixth-formers who left Sharon's school in the summer of 1981 forty had found real jobs by 18 November, four of these having entered military service. (source: BNC, GUR)
 - (c) Erm erm erm but, yeah and people er have great areas of that taken. (source: BNC, KC3)
 - (d) Homonyms are headwords to different entries that are spelt in the same way, e.g. bow (the weapon), bow (the action), bow (the verb expressing the action). (source: BNC, EAT)
2. Use the online *Word Calculator* to compare your results from Exercise 1 to the automatically generated token, type and lemma counts. Did you get the same results? If not can you explain the differences?

3. Use the online *Word Calculator* to compare different texts from the internet. Calculate the lexical density using the three measures discussed in this chapter: simple TTR, STTR and MATTR. Compare the findings and think about which of the measures would be most appropriate to use with the text. What are your reasons for selecting the measure?
4. Calculate the relative frequencies of the following items. In each case, choose an appropriate basis for normalization.
- (a) word: *muggle*
absolute frequency: 2
corpus size: 100,000
 - (b) word: *intriguingly*
absolute frequency: 3,035
corpus size: 11,191,860,036
 - (c) word: *worse*
absolute frequency: 50
corpus size: 1,007,299

5. Look at the frequency list below. It shows ten words from the BNC, together with their ranks. Use Zipf's law to predict the absolute frequency of the items presented in the table.

Rank	Word	Absolute frequency
1	the	6,041,234
2	of	
3	and	
4	to	
5	a	
10	was	
50	so	
100	way	
1,000	limited	
10,000	conveniently	

6. Compare your results from question 5 with the actual frequencies provided in the Answers section at the companion website. How well did Zipf's law predict the frequencies?
7. Look at the absolute frequencies of four selected words in the broadly defined genre parts of the BNC (Table 2.8). An electronic version of this table is available from the companion website.

For each word, calculate:

- (a) the range
- (b) the standard deviation
- (c) the coefficient of variation
- (d) Julland's D
- (e) DP.

Table 2.8 *BNC: distribution of four selected words*

BNC part	Total no. of tokens	<i>some</i> (AF)	<i>smile</i> (AF)	<i>theory</i> (AF)	<i>chance</i> (AF)
Fiction and verse	16,143,913	24,616	5,498	347	2,645
Newspapers	9,412,174	10,520	304	266	2,589
Non-academic prose and biography	24,178,674	43,161	385	3,977	2,191
Academic prose	15,778,028	30,297	58	6,588	923
Other written material	22,390,782	37,867	488	1,268	3,323
Speech	10,409,858	20,589	112	363	1,138
Whole corpus	98,313,429	167,050	6,848	12,809	12,809

8. Use the *Dispersion calculator* to check your results from Exercise 7.
9. Calculate the ARF of the following words in the BE06 corpus (985,628 tokens):
- (a) *frigid*: AF: 2, corpus positions: 840,797 – 848,280
 - (b) *chemistry*: AF = 7, corpus positions: 160,129 – 589,607 – 594,834 – 596,351 – 611,214 – 948,612 – 950,458
 - (c) *porn*: AF = 14, corpus positions: 16,602 – 16,792 – 28,191 – 49,606 – 161,929 – 170,396 – 268,155 – 497,891 – 497,916 – 498,146 – 498,205 – 498,216 – 498,246 – 498,361
10. Use the *ARF calculator* to compare AF and ARF for different words in texts of your choice from the internet.

THINGS TO REMEMBER

- There are different concepts of a ‘word’ – token, type, lemma and lexeme.
- Zipf’s law describes the distribution of words in corpora and their rapidly diminishing frequency.
- To fully describe a word in a corpus we need to provide both the word’s frequency and its dispersion.
- Different dispersion measures (*range*, *SD*, *CV*, *CV%*, Julland’s *D*, *DP*) are appropriate in different situations.
- The average reduced frequency (ARF) is a measure that combines both frequency and dispersion; it can be used with corpora that are not divided into different parts (subcorpora).
- TTR is a measure of lexical diversity; it is sensitive to text length.
- STTR and MATTR are alternative measures of lexical diversity that can be used with texts of varying lengths.

Advanced Reading

- Baroni, M. (2009). Distributions in text. In A. Lüdeling & M. Kytö (eds.), *Corpus linguistics: an international handbook*, vol. 2, pp. 803–21. Berlin: Mouton de Gruyter.
- Covington, M. A. & McFall, J. D. (2010). Cutting the Gordian knot: the moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2), 94–100.
- Gardner, D. (2007). Validating the construct of word in applied corpus-based vocabulary research: a critical survey. *Applied Linguistics*, 28(2), 241–65.
- Gries, S. Th. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 13(4), 403–37.
- (2009). Dispersions and adjusted frequencies in corpora: further explorations. *Corpus linguistic applications: current studies, new directions*, Amsterdam: Rodopi.
- Hlaváčová, J. (2006). New approach to frequency dictionaries – Czech example. *5th edition of the International Conference on Language Resources and Evaluation*, Genoa, 22–28 May. www.lrec-conf.org/proceedings/lrec2006/pdf/11_pdf.pdf (accessed 22/6/2014).
- Savický, P. & Hlaváčová, J. (2002). Measures of word commonness. *Journal of Quantitative Linguistics*, 9(3), 215–31.
- Tweedie, F. & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, 32, 323–52.

Companion Website: Lancaster Stats Tools Online

1. All analyses shown in this chapter can be reproduced using the statistical tools available from the companion website. The tools available for this chapter include:
 - Word calculator
 - Wordlist tool
 - Dispersion calculator
 - ARF calculator
2. The website also offers additional materials for students and teachers.

3

Semantics and Discourse

Collocations, Keywords and Reliability of Manual Coding

3.1

What Is This Chapter About?

So far, we have looked at words in isolation. In this chapter, we will explore meanings of words in context, which is an area important to both linguistic and social analyses. Topics discussed here – collocations, keywords and manual coding of concordance lines – play a key role both in the study of semantics ('dictionary' meanings of words) and in discourse analysis. The chapter starts with a simple premise: word meanings can best be investigated through the analysis of repeated linguistic patterns in corpora. Techniques such as keywords help us to draw attention to words characteristic of particular texts or corpora that can be further investigated using methods such as collocation, i.e. investigating repeated co-occurrence of words, and concordancing, i.e. analysing examples of word use in context. Five questions in particular are addressed in this chapter:

- How do we identify collocations? (Section 3.2)
- What are collocation networks? (Section 3.3)
- How do we identify keywords and lockwords? (Section 3.4)
- How can the manual coding of concordance lines be made more reliable? (Section 3.5)
- How can the techniques discussed in this chapter be used in linguistic and social research? (Section 3.6)

3.2

Collocations and Association Measures

Think about ...

Before reading this section, think about the following questions:

1. What associations come into your mind when you see the word *love*?
2. Why do you think the word has these associations for you?

It is a well-known fact in corpus linguistics that words occur in combinations that we call **collocations**. More than fifty years ago, Firth (1957: 6) suggested that we should investigate 'the company words keep', which has since become an

informal definition of the collocational relationship (Gries 2013b; Brezina et al. 2015). Collocations are combinations of words that habitually co-occur in texts and corpora. Collocations can be based either on frequency alone or, as is more common, on a statistical measure called an association measure. **Association measures** (sometimes also called **collocation measures**) are statistical measures that calculate the strength of association between words based on different aspects of the co-occurrence relationship (see below). There are many different association measures, each producing a (slightly) different list of collocates (Evert 2008; Gablasova et al. 2017b). There is no one measure which would suit all purposes and research questions. We therefore need to understand how association measures operate in order to select the one that best highlights the aspects of the collocational relationship we are interested in.

The following example demonstrates how the identification of collocations works in practice:

My love is like a red, red rose that's newly sprung in June: My love is like the melody that's sweetly played in tune. As fair art thou, my bonnie lass, so deep in love am I: And I will love thee still, my dear, till a' the seas gang dry. Till a' the seas gang dry, my dear, and the rocks melt wi' the sun : And I will love thee still, my dear, while the sands o' life shall run. And fare thee weel, my only love, and fare thee weel a while! And I will come again, my love, thou'it were ten thousand mile.

(Robert Burns, ‘A Red, Red Rose’)

The example above is taken from an English version of Burns’s famous poem (originally written in Scots) ‘A Red, Red Rose’. For the purposes of illustration, all occurrences of the word *love* with their immediate context – one word to the left and one word to the right – are highlighted and the poem is displayed without line breaks to create one paragraph of run-on text.

Let’s assume that we are interested in the use of the word *love* in the poem. We will call the word *love*, our word of interest, a ‘node’. A **node** is a word that we want to search for and analyse. The words around the node are candidate words for collocates. **Collocates** are words that co-occur with the node in a specifically defined **span** around the node, which we call the **collocation window**. Metaphorically, we can imagine this as a ‘magnetic field’ around the node that attracts particular collocates like a magnet attracts metal objects. In this case, the span (collocation window) is one word to the left and one word to the right, sometimes abbreviated to 1L, 1R. In the example given above, within a 1L, 1R collocation window around each occurrence of the node, we can observe the words that co-occur with *love*, namely, in the order of appearance, *my* (three times), *is* (twice), *in* (once), *am* (once), *will* (twice), *thee* (twice), *only* (once), *and* (once) and *thou* (once). Note that in each case, the frequency of co-occurrence was provided in the brackets; we call this value the **observed frequency of collocation**. Let us consider the word which occurs most frequently with *love* in our example, *my*. At this stage, we need to ask: is *my* really a genuine collocate of

love in the poem? In other words, is *my* really strongly associated with *love*? To find out, we need to find a way to evaluate the observed frequency. We have three basic options:

1. No baseline: we compare the observed frequencies of all individual words co-occurring with the node and produce a rank-ordered list.
2. Random co-occurrence baseline ('shake the box' model): we compare the observed frequencies with frequencies expected by chance alone and evaluate the strength of collocation using a mathematical equation which puts emphasis on a particular aspect of the collocational relationship.
3. Word competition baseline: we use a different type of baseline from random co-occurrence; this baseline is incorporated in the equation, which again highlights a particular aspect of the collocational relationship.

The first (simplest) option does not involve any statistical calculation. We merely produce a rank-ordered list of words co-occurring with the node based on their frequency, such as, in our example, *my* (3), *is* (2), *thee* (2), *will* (2) ... The words towards the top of the list are the strongest collocates by the frequency count. The disadvantage of this approach is that the top collocates will typically be function words co-occurring with the node merely by the virtue of their frequency anywhere in the corpus. Frequency-based collocates are therefore fairly generic (we can expect a similar set of collocates for almost any node) and have only a limited usefulness.

The second option involves a comparison with a random co-occurrence baseline. We ask whether it is possible that the combination of words in question (e.g. *my* and *love*) occurs repeatedly only due to chance. Consider this:

1. The poem has 107 tokens (see Section 2.2 for a definition of 'token').
2. *Love* occurs 7 times in the whole poem.
3. *My* occurs 8 times in the whole poem.
4. *My* occurs 3 times in combination with *love* and 4 times in combination with other words.

Imagine also that there were no associations between words in the poem and words appeared randomly in the text. This situation is illustrated in the example below, which shows the words of Burns's poem in random order:

fare art And like red, sweetly in **love love**, And gang wi' played like dear, life
 shall rocks sprung the Till deep my my And still, weel, again, ten the the
 while! is till And As I: a' only come were sands sun: dry, and gang it a' the
 still, My thee will in my bonnie My red is a run. **[my love]** thee thou, melt the
 seas and thou' I the I lass, I melody thee a my am rose **love** dear, that's **love**
 newly **love** fare **love**, will o' so dry. fair thee will that's in while June: my seas
 tune. mile. thousand weel dear,

How many times would you expect the words *my* and *love* to co-occur by chance alone? In the random example above, *my love* occurs once; in fact, if we run the random simulation multiple times we will get an average (mean) number close to one. We call this process establishing the **random co-occurrence baseline** and the resulting value is called the **expected frequency of collocation**. The expected frequency of collocation does not have to be established empirically but can be calculated as follows:

$$\text{expected frequency of collocation} = \frac{\text{node frequency} \times \text{collocate frequency}}{\text{no. of tokens in text or corpus}} \quad (3.1)$$

When considering collocation window sizes larger than one, a correction could be applied to account for the fact that there is a greater chance of words randomly co-occurring with the node. The corrected expected frequency of collocation is calculated as follows:

$$\begin{aligned} &\text{expected frequency of collocation (corrected)} \\ &= \frac{\text{node frequency} \times \text{collocate frequency} \times \text{window size}}{\text{no. of tokens in text or corpus}} \end{aligned} \quad (3.2)$$

In our example, the expected frequency of collocation of *love* and *my* in the 1L–1R span would be calculated as follows:

$$\text{expected frequency of collocation (love, my; corrected)} = \frac{7 \times 8 \times (1+1)}{107} = 1.05 \quad (3.3)$$

We can see that the observed frequency of collocation of *love* and *my* (3) is larger than the expected frequency (1.05). To compare the difference between the two values, different association measures can be applied (see Table 3.3). All association measures that include the expected frequency in the formula (E_{11}) are based on the random co-occurrence baseline. The potential disadvantage of these measures is that they assume a particular model of language ('shake the box' model), which might be problematic (e.g. Stubbs 2001: 73–4). This model is analogous to the corpus being a box in which we have all words written on separate small cards – this box is then shaken thoroughly to obtain the baseline; however, language is much more orderly and structured and the model therefore involves a great deal of simplification.

Finally, to avoid the potentially problematic 'shake the box' model of language, some association measures operate with a different type of baseline; these measures do not include E_{11} in the equation. The baseline needs to be understood on the case-by-case basis derived from the specific formula of the association measure.

Generally, to be able to understand the equations, we need to consider the terms that enter these equations. These are best displayed in the form of **contingency tables** (showing all possible combinations – contingencies – of word co-occurrence). Table 3.1 displays observed frequencies (frequencies of co-occurrence we can 'observe' in corpora), while Table 3.2 calculates expected

Table 3.1 *Observed frequencies*

	Collocate present present (<i>affair</i>)	Collocate absent	Totals
Node present (<i>love</i>)	O ₁₁	O ₁₂	R ₁ × window size
Node absent	O ₂₁	O ₂₂	R ₂
Totals	C ₁	C ₂	N

frequencies and is therefore relevant only for the measures using the random co-occurrence baseline.¹

In the contingency tables, O stands for the observed frequency, while E stands for the expected frequency. C is a symbol for column total and R for row total. The numbers after O and E refer in turn to the relevant row and column. For example, O₁₂ thus stands for the observed frequency in the first row and the second column of the first contingency table.

The expected frequencies table is derived entirely from the observed frequencies table, using the equations shown in Table 3.2. The shaded cells in Table 3.1 represent values which we need to collect directly from the corpus (using an appropriate piece of software). These are:

- (a) Number of tokens in the whole corpus: N
- (b) Frequency of the node in the whole corpus: R₁
- (c) Frequency of the collocate in the whole corpus: C₁
- (d) Frequency of the collocation (i.e. node + collocate) in the collocation window: O₁₁
- (e) Collocation window size

To calculate a range of association measures the equations listed in Table 3.3 are used (the terms refer to the two contingency tables: Tables 3.1 and 3.2).

Finally, one question remains to be answered: which is the ‘best’ association measure? Frustratingly for some, the answer is: it depends on which aspects of the collocational relationship we want to highlight. Some collocation measures such as MI highlight rare exclusivity of the collocational relationship, favouring collocates which occur almost exclusively in the company of the node, even though this may be only once or twice in the entire corpus. Other metrics, such as Dice and log Dice, and MI2 favour collocates which occur exclusively in each other’s company but do not have to be rare. Others can take into account directionality (Delta P) or dispersion (Cohen’s d). Let us briefly consider directionality and dispersion. When exploring **directionality**, we ask the question ‘to what extent is the attraction between the node and the collocate mutual?’ In a symmetrical relationship, the attraction of the collocate to the node is almost as strong as that of the node to the collocate. *Red light* is an example of such collocational symmetry. On the other hand, in an asymmetrical relationship, the

¹ The notation in the contingency tables is based on Evert (2008).

Table 3.2 *Expected frequencies: random occurrence baseline*

	Collocate present (affair)	Collocate absent	Totals
Node present (<i>love</i>)	$E_{11} = \frac{R_1 \times C_1}{N}$	$E_{12} = \frac{R_1 \times C_2}{N}$	R ₁
Node absent	$E_{21} = \frac{R_2 \times C_1}{N}$	$E_{22} = \frac{R_2 \times C_2}{N}$	R ₂
Totals	C ₁	C ₂	N

attraction is considerably stronger in one direction. For example, in the collocation *red herring*, the attraction is much stronger from *herring* to *red* than vice versa (McEnery 2006: 18). This means that when we see the word *herring* in an English text, there is a large probability (over 20%) that the word *red* would precede. On the other hand, when we see the word *red* in a text, we can make no such strong prediction about *herring* following the word, because *red* can be followed by a very large number of different nouns and it is only in less than 0.3% cases that *red* is followed by *herring*.² In practical terms, a directional measure such as Delta P outputs two probabilities, one for each direction of the collocational relationship. Non-directional measures (all measures but Delta P from Table 3.3) output only one value so cannot be used to explore directionality. **Dispersion**, another possible aspect of the collocational relationship, is the distribution of collocates in individual corpus files or corpus parts (see Sections 1.3 and 2.4). Arguably, the more evenly dispersed a collocate is in a corpus, the more important it is for the corpus as such.

So in practical terms, how do we choose an association measure? Look at Table 3.4, which shows the typical performance of the 14 metrics presented in Table 3.3. A range of words from the highly frequent definite article *the* to low-frequency words such as *ex-teacher* or *zealand* were considered as collocates of the adjective *new* (AF = 1,233) in the one-million-word corpus BE06. The numbers in the table show how the individual association measures rank the collocates, with 1 being the most important collocate of the set of eight according to the particular statistic.

To choose a specific association measure we first need to define the type of collocations we are interested in (based on our research question). We can think of most association measures as highlighting collocations along two main dimensions: frequency and exclusivity. Frequency refers to the number of instances in which a node and collocate occur together in a corpus. Exclusivity refers to a specific aspect of the collocation relationship where words occur only or predominantly in each other's company. According to how the association

² These probabilities are based on the frequencies of *red*, *herring* and *red herring* in a 12-billion-word corpus *EnTenTen12*.

Table 3.3 *Association measures: overview*

ID	Statistic	Equation	ID	Statistic	Equation
1	Freq. of co-occurrence	O_{11}	8	T-score	$\frac{O_{11} - E_{11}}{\sqrt{O_{11}}}$
2	MU	$\frac{O_{11}}{E_{11}}$	9	Dice	$\frac{2 \times O_{11}}{R_1 + C_1}$
3	MI (mutual information)	$\log_2 \frac{O_{11}}{E_{11}}$	10	Log Dice	$14 + \log_2 \frac{2 \times O_{11}}{R_1 + C_1}$
4	MI2	$\log_2 \frac{O_{11}^2}{E_{11}}$	11	Log ratio	$\log_2 \frac{O_{11} \times R_2}{O_{21} \times R_1}$
5	MI3	$\log_2 \frac{O_{11}^3}{E_{11}}$	12	MS (minimum sensitivity)	$\min\left(\frac{O_{11}}{C_1}, \frac{O_{11}}{R_1}\right)$
6	LL (log likelihood)	$2 \times \left(O_{11} \times \log \frac{O_{11}}{E_{11}} + O_{21} \times \log \frac{O_{21}}{E_{21}} + O_{12} \times \log \frac{O_{12}}{E_{12}} + O_{22} \times \log \frac{O_{22}}{E_{22}} \right)$	13	Delta P	$\frac{O_{11} - O_{21}}{R_1} - \frac{O_{21}}{R_2}; \frac{O_{11}}{C_1} - \frac{O_{12}}{C_2}$
7	Z-score ₁	$\frac{O_{11} - E_{11}}{\sqrt{E_{11}}}$	14	Cohen's <i>d</i>	$\frac{Mean_{in\ window} - Mean_{outside\ window}}{pooled\ SD}$

Table 3.4 *Ranking of collocates of 'new' in BE06 (L3–R3)*

Collocate	C1	O11	occurrence	Freq. of co-			T-score ₁	T-score ₂	T-score (uncorr.) ^a	Dice	Log Dice	Log ratio	MS	Delta P	Cohen's <i>d</i>		
				MU	MI	MI2											
the	58,951	447	1	7	7	4	2	8	7	1	2	2	7	2	3	8	
and	27,917	203	2	8	8	5	3	7	8	2	3	3	8	3	8	7	
york	100	83	3	3	3	1	1	1	1	1	1	1	2	1	1	1	
year	708	25	4	6	6	8	6	5	6	2	4	4	4	6	4	2	5
system	285	17	5	5	5	6	7	4	5	5	5	5	5	5	5	3	3
technologies	31	14	6	4	4	3	5	3	3	4	7	5	6	4	5	3	2
zealand	14	14	6	1	1	2	4	2	2	3	6	5	6	1	5	3	4
ex-teacher	1	1	8	1	1	7	8	6	4	6	8	8	8	3	8	7	6

^a T-score shows large differences between the corrected and the uncorrected version. Other metrics such as MI are pretty stable with the corrected versions showing slightly smaller values than the uncorrected ones.

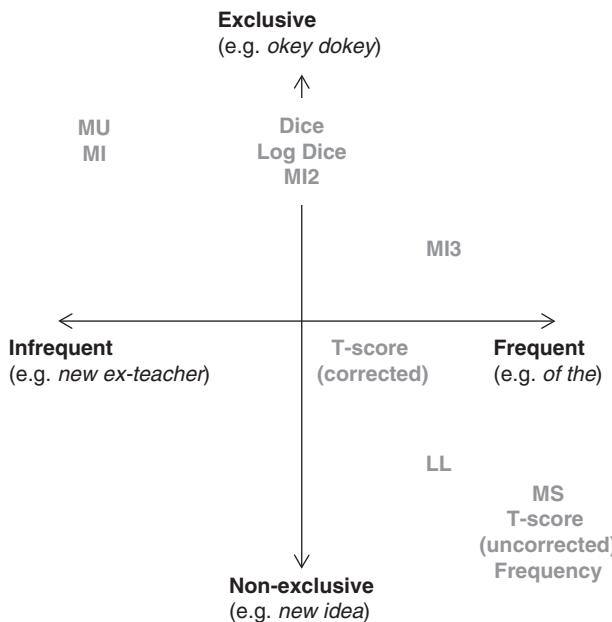


Figure 3.1 *Frequency and exclusivity scale*

measures rank the individual collocates, association measures can be placed on a two-dimensional scale indicating the extent to which these association measures highlight the frequency and/or exclusivity of the collocational relationship (see Figure 3.1). Log ratio, Delta P and Cohen's d are not included in the graph. Log ratio is a combined measure which presupposes filtering of the data (typically by LL) before the log ratio equation is applied; Delta P takes directionality into account while Cohen's d accounts for dispersion. These measures operate on separate dimensions and are therefore not displayed in Figure 3.1.

Reporting Statistics

1. What to Report

For the sake of replicability of results, all major parameters that can affect collocate identification should be reported. For this purpose, Brezina et al. (2015) introduce collocation parameters notation (CPN) that captures all important parameters for collocate identification (see Table 3.5).

Table 3.5 *Collocation parameters notation (CPN)*

Statistic ID	Statistic name	Statistic value	Statistic cut-off	L and R span	Minimum collocate freq. (C)	Minimum collocation freq. (NC)	Filter
4b	MI2	3		L5–R5	5	1	function words removed
4b–MI2(3), L5–R5, C5–NC1; function words removed							

CPN has seven different parameters. Statistic ID refers to the number in the ID column of Table 3.3. ‘a’ after the statistic ID signifies an uncorrected and ‘b’ signifies a corrected version of the same statistic – this detail refers to the correction for window sizes larger than one discussed above. This is followed by the name of the statistic and the statistic cut-off value used (in brackets), the span of the left and the right context, the minimum frequency of the collocate in the whole corpus, and the minimum frequency of the collocation (i.e. the co-occurrence of the node and the collocate). The last parameter, the filter, specifies any further procedures in the collocation extraction process, for example removal of certain words from the results (e.g. based on word-class membership), or a minimum dispersion value.

2. How to Report: An Example

- The following items were identified as top five collocates of the adjective *new* in the BE06 corpus using the MI statistic (3b–MI(5), L3–R3, C5–NC5; no filter applied): *zealand, mobilities, york, technologies* and *testament*.

3.3 Collocation Graphs and Networks: Exploring Cross-associations

Think about ...

Before reading this section, think about the following questions:

- What words come to mind when you see the word *university*? Write down at least five associations.
- Review the list and underline those words which you think are more closely associated with *university* than the rest of the words in the list.

Collocation graphs and networks build on the idea of collocation introduced in Section 3.2. A **collocation graph** is a visual representation of the collocational relationship between a node and its collocates. Instead of a list of collocates

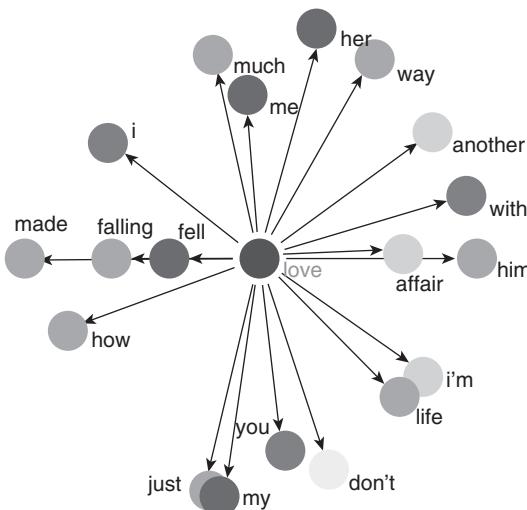


Figure 3.2 *Collocation graph: 'love' in BE06 (10a – log Dice (7), L3–R3, C5–NC5)*

displayed in a table (the usual format), the graph shows the relationship between the node and the collocates by displaying collocates closer to or further apart from the node. Figure 3.2 shows an example of a collocation graph based on a one-million-word corpus BE06 using log Dice as the association measure. The centre of the graph is occupied by the node (*love*), around which are placed the collocates. The graph displays three dimensions of the collocation relationship: (i) strength of the association, (ii) frequency of the collocate and (iii) position of the collocate in text. The strength of association as measured by the association measure (here log Dice; see Section 3.2) is expressed as the length of the link between the node and the collocate: the closer the collocate is to the node, the stronger the relationship (think of a magnet). The frequency is displayed as the shade of the colour of the collocate: the more intense the colour the more frequent the collocate is. Finally, the position of the collocate in text (whether it occurs predominantly before or after the node) is shown by the position of the collocate in the graph (left, middle or right). For instance, *fall*, *falling* and *fell* occur always before the node *love*, while *you* occurs both before and after *love* with approximately the same frequency.

Collocation networks are extended collocation graphs, which show larger association patterns than those seen from the immediate collocates discussed so far (Phillips 1985; Williams 1998; Brezina et al. 2015). A **collocation network**, as the term suggests, is a network of linked collocations (see Figure 3.3) that starts with an initial node (N1) around which a set of first-order collocates is identified (C1–C5). Any of these collocates can in turn be considered as a new node (N2) for which another set of collocates,

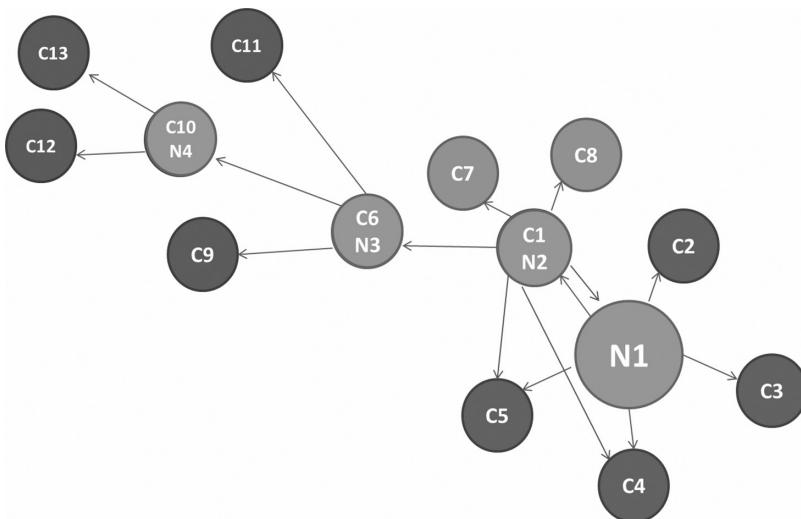


Figure 3.3 *Collocation networks: concept demonstration*

second-order collocates (C6–C8), is identified. This process can be repeated multiple times (four times in Figure 3.3) to build an extensive network which shows how individual words are connected through word associations and cross-associations. Collocation networks thus work with a simple assumption: even distant collocates can participate in the meaning of a word of interest by sharing a similar conceptual space with this word through the link of associations with other words that directly collocate with the word of interest.

It is best to illustrate this idea with an example of a relatively simple collocation network based on the one-million-word LOB corpus (see Figure 3.4 – note the CPN in the caption). The network shows the relationship between the word *time* and the word *money*. Although *money* does not directly collocate with *time* (or vice versa), we can see that these two words are connected via their mutual associations with words such as *spend*, *spent*, *saved*, *waste*, *lose* and *(a) lot*. This observation can provide evidence for the existence of the well-known conceptual metaphor TIME IS MONEY proposed by Lakoff & Johnson (1980). Based on this evidence we can claim, together with Lakoff & Johnson, that in English we use the word *time* in a similar way to the word *money* and that we often understand time through financial metaphors.

Another example that demonstrates the concept of collocation networks is the word *university* and its cross-assOCIATIONS that can be seen from Figure 3.5. You can compare this collocation network with your answers to questions 1 and 2 from the ‘Think about’ task at the beginning of this section. We can see that although the word *university* itself shows only seven immediate collocates that satisfy the identification criteria recorded by means of CPN, the conceptual space

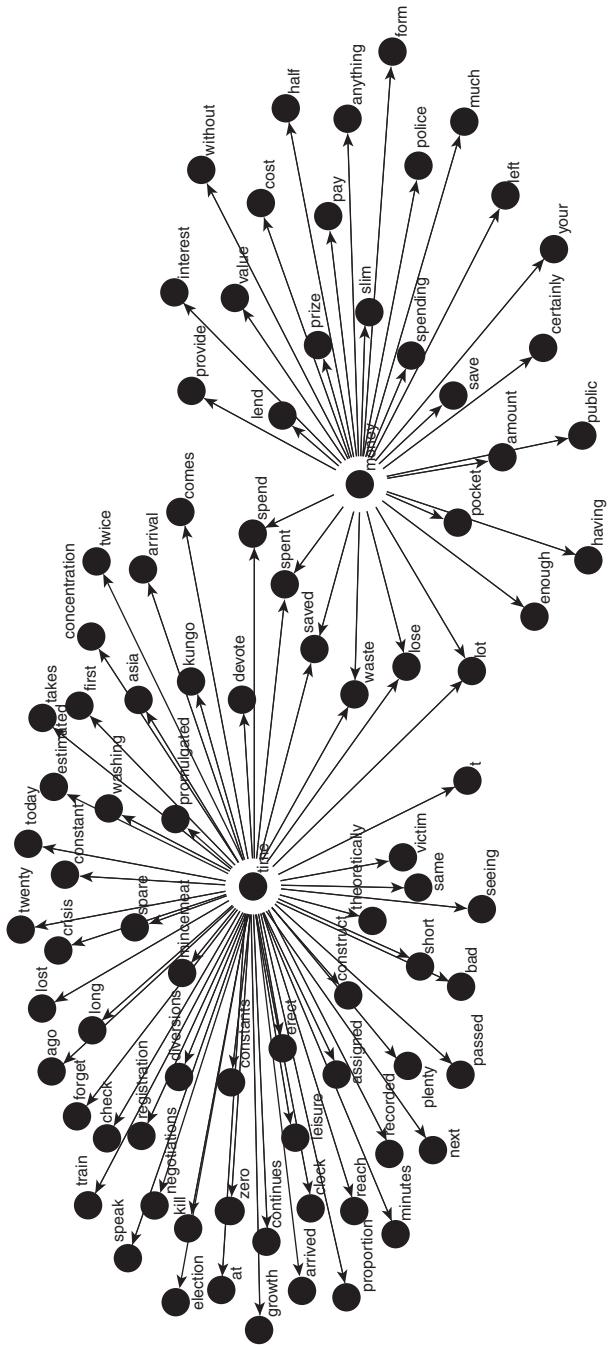


Figure 3.4 Third-order collocates of time in LOB (3a–MI(5), R5–L5, C4–NC4, no filter applied)

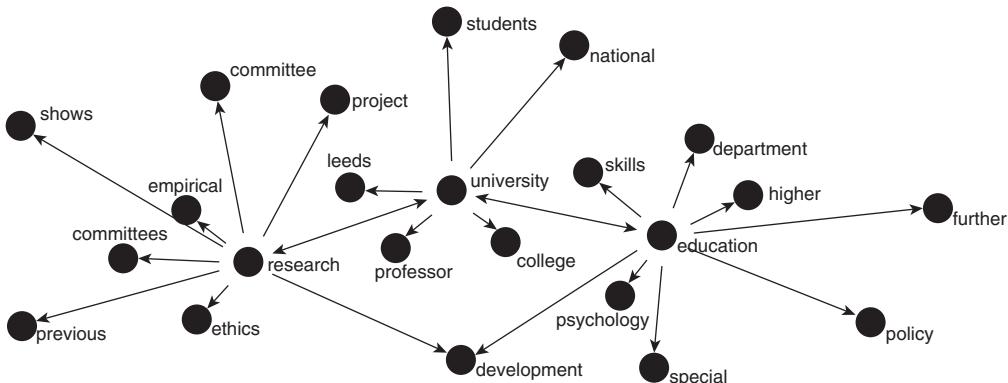


Figure 3.5 Collocation network of '*university*' based on BE06 (3b–MI(3), L5–R5, C8–NC8)

that *university* occupies is much larger, as can be seen from the more distant collocates shown in Figure 3.5.

Collocation graphs and networks are useful summaries of complex meanings of words in texts and corpora. These networks can provide useful information about key topics in texts and discourses as well as their connection. For effective building of collocation networks specialized software is necessary that is able to run multiple comparisons of word co-occurrence and display the data in a visual form. Such a tool, #LancsBox (Brezina et al. 2015), is now available for free download. A link to the tool is provided at the companion website. #LancsBox enables the user to upload their own corpora and easily identify collocation networks with appropriate CPN parameters. The tool implements all association measures from Table 3.3 and also allows the user to modify existing association measures or to add new ones.

3.4 Keywords and Lockwords

Think about ...

Before reading this section, think about the following question: Which of the lists in Table 3.6 best captures words frequently occurring in American but not in British English? Provide reasons for your choice.

Identifying keywords is one of the crucial techniques in corpus linguistics (Scott 1997), yet it is also a procedure that is often misunderstood. **Keywords** are words that are considerably more frequent in one corpus than in another corpus; we can therefore say that keywords are words that are typical of the corpus of interest when

Table 3.6 *AmE06: American English keywords*

Keyword list 1	Keyword list 2	Keyword list 3	Keyword list 4
U.S.	U.S.	LABOR	TOWARD
PERCENT	PERCENT	NEIGHBORHOOD	NEIGHBORHOOD
AMERICAN	PROGRAM	DEFENSE	RECOGNIZE
PROGRAM	TOWARD	CONGRESSIONAL	NEIGHBORS
TOWARD	AMERICAN	ATLANTA	COLORED
STATES	BUSH	PGF2A	MANHATTAN
FEDERAL	FEDERAL	MACDOWELL	FAVORITE
BUSH	STATES	MRNA	RECOGNIZED
PRESIDENT	CENTER	NEIGHBORS	CENTER
CENTER	MR.	ABBY	REALIZE
MR.	PRESIDENT	GENOME	RECOGNIZING
PROGRAMS	PROGRAMS	FLORIDA	TRAVELED
UNITED	UNITED	9–11	SIGNALED
STATE	WASHINGTON	DOE	COLOR
CONGRESS	CONGRESS	POE	CALIFORNIA
WASHINGTON	AMERICANS	ROUSSEAU	GOTTEN
AMERICANS	STATE	NS1	LABOR
DEFENSE	CALIFORNIA	REZKO	FAVOR
CALIFORNIA	AMERICA	MITCH	FINALLY
WAR	DEFENSE	ADDITIVES	CENTERS

compared to another corpus. However, it is important to remember that ‘keywords’ is a relative term depending on the differences in lexical frequencies in the two corpora in question. Keywords are important when identifying key concepts in discourses, typical vocabulary in a genre/language variety, lexical development over time, etc. Complementary to keywords are lockwords, a term introduced by Baker (2011). **Lockwords** are words that occur with similar frequencies in two corpora that we compare. Keywords (and lockwords) rely on our ability to compare two corpora in a meaningful way. This, however, is more complicated than it sounds and leads to some controversy about how to best perform the comparison.

So, let us unpack the keyword procedure: a **corpus of interest (C)**, sometimes referred to as a ‘focus corpus’ (Kilgarriff 2012) or ‘node corpus’ (Scott 1997), is compared with a baseline **reference corpus (R)** using a statistical measure to identify words that are used either more often or less often in C when compared to R. If a word is used more often in C than in R then we call it a positive keyword (+); if, on the other hand, a word is used less in C than in R we use the term negative keyword (−). Finally, if a word is used in C with a comparable frequency to its frequency in R we talk about a lockword (0). In practice, this is done automatically by corpus software, which creates two wordlists, one based on C and the other based on R, and compares the items on these wordlists one-by-one to decide which category the compared words belong to. The keyword procedure is schematically summarized in Table 3.7.

Table 3.7 *Decisions about keywords: BASIC options*

Corpus of interest C	Reference corpus R	Decision
Frequent	Infrequent	+ (positive keyword)
Infrequent	Frequent	- (negative keyword)
Comparable freq.	Comparable freq.	0 (lockword)

So far, the principle of comparison seems straightforward. However, there are a number of specific decisions we have to make when operationalizing this general principle. The following is an exploration of the different options we have when we identify keywords and lockwords.

1. How to Choose a Reference Corpus?

Typically, a reference corpus is larger than or similar in size to the corpus of interest so as to provide a large enough amount of evidence about word frequencies (see question 2 below). Generally speaking, the larger and the more similar the reference corpus is to the corpus of interest the more reliable and focused the comparison is. This is because every two corpora are different in a number of aspects which get highlighted to different degrees in the keyword procedure. To illustrate the point of the relative heterogeneity of any two corpora, let's compare two excerpts taken from a corpus of American (AmE06) and a corpus of British (BE06) English respectively, each excerpt consisting of exactly 100 words.

Text A	Text B
Democrats call those shifts too little, too late. "Changing direction in Iraq starts with changing the people we send to Washington," Shays' challenger, Diane Farrell, said Saturday in the Democratic response to Bush's radio address.	Something behind him went 'gloink'. It was a small, subtle and yet curiously intrusive sound, and it accompanied the appearance, on a shelf above Rincewind's desk, of a beer bottle where no beer bottle had hitherto been. He took it down and stared at it. It had recently contained a pint of Winkle's Old Peculiar. There was absolutely nothing ethereal about it, except that it was blue. The label was the wrong colour and full of spelling mistakes but it was mostly there, right down to the warning in tiny, tiny print: May Contain Nuts. Now it contained a note.
Democrats, who since the Vietnam War have battled voter perceptions that they are soft on defense, are finding a more receptive audience for the argument that they could do a better job of protecting America and conducting its foreign policy. In the poll, 52% say the Iraq war has made the USA less safe from terrorism.	
Nonetheless, Republicans continue to view the issue of terrorism as ...	

We can start enumerating the differences between the texts by pointing out that text A is about American politics, while text B is a part of a story about a person called Rincewind. Text A comes from a newspaper, text B from a novel. Text A uses American spelling (*defense*), whereas text B sticks with the British spelling conventions (*colour*). We may also begin to observe individual lexical differences reflecting the topics of the two texts. Some readers might also notice the unusual onomatopoeic word *gloink* in text B and words related to the war in Iraq (*Bush*, *Iraq*, *terrorism*, *war* etc.) in text A. These are only a few differences that illustrate the aspects in which two texts can be different from each other – the same is true, on a large scale, about two corpora.

The crucial question to ask in this context therefore is: what kind of language do the corpus of interest (C) and the reference corpus (R) represent and how is the composition of each corpus reflected in the comparison – the keyword procedure? (See Section 1.4 for the discussion of corpus representativeness.) In practice, every keyword (or lockword) identified should be related back to what we know about the composition of the two corpora (C and R) and the multiple sources of difference between them. We should also consider which words would get highlighted as keywords had we chosen a different reference corpus. In fact, we can select multiple reference corpora and carry out the keyword procedure with each of them and compare the results.

2. How to Deal with Absent Words?

When comparing two corpora, it often happens that a word that occurs frequently in one corpus is absent from the other corpus. The question is how to deal with these words in the keyword procedure. We know that unless a corpus represents the population (all language use), absence of evidence is not evidence of absence (see Section 1.4). The answer to the question thus largely depends on the relative sizes of the two corpora we compare and our understanding of their representativeness and sampling. As a general rule, we should always carefully evaluate the amount of evidence we have in the two corpora in question for the claim that something is a keyword (positive or negative). The practical questions to ask here are:

- Is X a positive keyword or is the reference corpus not large enough?
- Is Y a negative keyword or is the corpus of interest not large enough?

In practice, some corpus tools let us set the minimum cut-off limits for the frequencies of words in C and R before considering them in the keyword procedure. This, however, needs to be done very thoughtfully.

3. What Statistical Measure to Use for Comparing Corpora?

To answer this question, let us take as an example the two corpora from which excerpts A and B (see above) were taken, AmE06 and BE06, each consisting of approximately one million words. Let us consider five lexical items that stand out as different in texts A and B – *war*, *defense*, *pint*, *Rincewind* and *gloink* – as well as two more general (grammatical) words, *the* and *he*. This time, however, the comparison will be made between the whole corpora that include texts A and B respectively. Let's also assume that AmE06 is our corpus of interest (C) and BE06 is the reference corpus (R). We can see in Table 3.8 that the words *war*, *defense* (American spelling) do indeed appear more frequently in the American English than the British English corpus, while *pint* and *Rincewind* are more prominent in the British English corpus. *Gloink* is a so-called hapax legomenon, a word that appears only once in the corpus, with its one occurrence being in text B. *The* and *he*, which are common grammatical words, occur with relatively comparable frequencies. If we translate these observations into decisions about keywords, we can hypothesize to have two positive keywords (+), one negative keyword (–) and two lockwords (0). For *Rincewind* and *gloink* there is not enough evidence in the American corpus to make a decision about their keyness.

To help make these decisions, traditionally the log-likelihood (LL) statistic has been used to establish whether the differences between C and R are likely to be due

Table 3.8 Comparison of selected lexical items in BE06 and AmE06

Words	C: AmE06 AF (RF ^a)	R: BE06 AF (RF ^a)	RF ratio AmE06/ BE06 ^b	Hypothesis
war	620 (609.11)	267 (265.00)	2.30	+
defense	120 (117.89)	1 (0.99)	118.78	+
pint	1 (0.98)	16 (15.88)	0.06	–
Rincewind	0 (0)	10 (9.93)	0.00	NA
gloink	0 (0)	1 (0.99)	0.00	NA
the	59,901 (58,848.84)	58,960 (58,519.23)	1.01	0
he	7,310 (7,181.60)	6,827 (6,775.96)	1.06	0

^a relative frequency per million words

^b relative frequency ratio between AmE06 and BE06

to chance or are statistically significant. As with collocations (see Section 3.2) two contingency tables, one with observed and one with expected frequencies, are used and the values are entered into the log-likelihood equation below.³

$$\text{log likelihood}_{\text{short}} = 2 \times \left(O_{11} \times \log \frac{O_{11}}{E_{11}} + O_{21} \times \log \frac{O_{21}}{E_{21}} \right) \quad (3.4)$$

O_{11} stands for the frequency of the word of interest (w) in C, while O_{21} is the frequency of the same word in R. E_{11} and E_{21} are the frequencies that we would expect by chance in C and R respectively if there were no difference in the frequency of w in the two corpora. The expected frequencies are calculated as follows:

$$E_{11} = \frac{\text{tokens in C} \times (\text{freq. of } w \text{ in C} + \text{freq. of } w \text{ in R})}{\text{total no. of tokens in C and R}}$$

$$E_{21} = \frac{\text{tokens in R} \times (\text{freq. of } w \text{ in C} + \text{freq. of } w \text{ in R})}{\text{total no. of tokens in C and R}}$$

For example, to calculate the log-likelihood statistic for the word *war* from Table 3.8, we would use the absolute frequencies (AF) and the following corpus sizes: AmE06: 1,017,879 tokens; BE06: 1,007,532 tokens.

$$E_{11} = \frac{1,017,879 \times (620 + 267)}{2,025,411} = 445.77$$

$$E_{21} = \frac{1,007,532 \times (620 + 267)}{2,025,411} = 441.23$$

$$\begin{aligned} \text{log likelihood}_{\text{short}}(\text{war}) &= 2 \times \left(620 \times \log \frac{620}{445.77} + 267 \times \log \frac{267}{441.23} \right) \quad (3.5) \\ &= 140.87 \end{aligned}$$

Based on the value of the LL statistic that is larger than 3.84, which is the cut-off point for significance at $p<0.05$ level (to be found in statistical tables or outputted directly by a computer program), we could conclude that the difference between the use of *war* in AmE06 and BE06 is statistically significant; this means that we have enough evidence in the data to reject the null hypothesis, which says that there is no difference between the frequencies of *war* AmE06 and BE06 (see Section 1.3). In fact, we can report the p-value as being smaller than 0.0001 ($p<0.0001$) because the log likelihood value is larger than 15.13. We can thus conclude that *war* is a positive keyword for AmE06 (American English). Although log likelihood can help us decide whether we have enough evidence in the corpus to say that the frequencies of the same word differ between C and R, there is growing evidence (Brezina

³ In fact, there are two forms of the equation, the short and the long LL equation, which differ only slightly in the final output. The short LL equation can be seen as a simplification of the long LL equation for the keyword analysis. The long form of the LL equation was listed in Table 3.3 under association measures.

& Meyerhoff 2014; Bestgen 2014) that log likelihood used for corpus comparison is prone to identifying far too many keywords (false hits). This is because, among other things, relatively small frequency differences between C and R can reach statistical significance in large enough corpora. To address the issue of a large number of identified keywords, keywords were traditionally sorted according to the LL statistic (sometimes labelled somewhat unhelpfully as ‘keyness’) and only those words with the largest LL values (e.g. top 10, 50 or 100 keywords) were considered for further analysis. This, however, raises the question of whether the LL statistic is the best measure to show the size of the difference in the frequencies of words in C and R and how to meaningfully interpret results such as 140.87.

Instead of using the values of the LL statistic, which are relatively difficult to interpret, for the identification and sorting of keywords, Kilgarriff (2009) suggests looking at the ratio between the relative frequencies of words in C and R. Because the ratio can be calculated only if the values in R are greater than zero (division by zero is not defined in mathematics), Kilgarriff suggests adding a constant k to both relative frequencies before calculating the ratio. The constant can be any positive number of our choice, but typically it is 1, 10, 100 or 1,000. The resulting measure is called the simple maths parameter (SMP) and is calculated as follows:

$$\text{simple maths parameter} = \frac{\text{relative frequency of } w \text{ in C} + k}{\text{relative frequency of } w \text{ in R} + k} \quad (3.6)$$

The constant k simultaneously serves as a filter that allows focusing on words above certain relative frequencies in the corpus. For example, if we use 1 as the constant, we highlight low-frequency unique words, while 100 would filter out words that occur with the relative frequency smaller than 100 per million words if the relative frequency per million words is used. The simple maths parameter for the word *war* (with $k = 100$) is calculated as follows.

$$\text{simple maths parameter } (\textit{war}) = \frac{620 + 100}{267 + 100} = 1.96 \quad (3.7)$$

The interpretation of the value of the simple maths parameter is more straightforward than that of the LL statistic: given that the relative frequencies of the word are larger than 100 (which we specified when choosing k), we can say that *war* occurs approximately twice as much in C as in R.

Currently, the question of which statistic best suits the identification of keywords is an open one. Other suggestions for sorting principles are %DIFF (Gabrielatos & Marchi 2012), log ratio (Hardie 2014) and Cohen’s d (Brezina 2014); the last of these takes dispersion into account. The keyword procedure can be also applied to categories at a higher level of abstraction such as lemmas or key semantic domains (Rayson 2008). Through this we can identify larger concepts or semantic areas typical of one type of discourse compared with another, i.e. that in the reference corpus.

Table 3.9 *American English keywords: different keyword identification procedures*

Log likelihood	SMP (with 100 as the constant)	Log ratio	Cohen's <i>d</i>
U.S.	U.S.	LABOR	TOWARD
PERCENT	PERCENT	NEIGHBORHOOD	NEIGHBORHOOD
AMERICAN	PROGRAM	DEFENSE	RECOGNIZE
PROGRAM	TOWARD	CONGRESSIONAL	NEIGHBORS
TOWARD	AMERICAN	ATLANTA	COLORED
STATES	BUSH	PGF2A	MANHATTAN
FEDERAL	FEDERAL	MACDOWELL	FAVORITE
BUSH	STATES	MRNA	RECOGNIZED
PRESIDENT	CENTER	NEIGHBORS	CENTER
CENTER	MR.	ABBY	REALIZE
MR.	PRESIDENT	GENOME	RECOGNIZING
PROGRAMS	PROGRAMS	FLORIDA	TRAVELED
UNITED	UNITED	9–11	SIGNALED
STATE	WASHINGTON	DOE	COLOR
CONGRESS	CONGRESS	POE	CALIFORNIA
WASHINGTON	AMERICANS	ROUSSEAU	GOTTEN
AMERICANS	STATE	NS1	LABOR
DEFENSE	CALIFORNIA	REZKO	FAVOR
CALIFORNIA	AMERICA	MITCH	FINALLY
WAR	DEFENSE	ADDITIVES	CENTERS

In sum, the term ‘keywords’ might be slightly misleading because it suggests that there is a single set of words that characterize a particular corpus. However, as we have seen, the keyword list is a result of multiple decisions in the process starting with the selection of the reference corpus and finishing with the choice of the particular statistic. To demonstrate this point, let’s return to the question from the ‘Think about’ task at the beginning of the chapter. If you have chosen your favourite keyword list for American English, you might be interested in knowing which procedure was used to identify these keywords. Table 3.9 also demonstrates clearly that there is no one single answer to the question of what the keywords are in a particular corpus.

Reporting Statistics

1. What to Report

The outcomes of the keyword procedure are influenced by three crucial parameters: (i) the selection of the reference corpus, (ii) implementation of minimum frequency cut-off points and (iii) the choice of the statistical measure. It is also customary to report whether all identified keywords or only the top 10, 50, 100 etc. were used. The parameters listed above are usually reported in the Procedure subsection of the Method section. This, however, needs to be complemented with a careful description of the corpus of interest C in the Data subsection of the Method section.

2. How to Report: Examples

Data

- AmE06, which represents written American English sampled in 2006, was used. AmE06 consists of 15 genre-based subdivisions which can be grouped into four broader genre divisions: newspapers, general prose, academic prose and fiction ...

Procedure

- Words typical of American English were identified by comparing AmE06 and BE06 using the keyword procedure. BE06 was used as a reference corpus because it was created using the same sampling frame (Brown family sampling frame) and therefore represents the same genres of written English sampled around the same point in time (2006) as AmE06. For the identification of keywords, Kilgarriff's (2009) SMP was used with the constant 100; no frequency cut-off points were applied. The top 20 positive keywords were further analysed.

3.5 Inter-rater Agreement Measures

Think about ...

Before reading this section, think about the following question:

Which of the concordance lines in Table 3.10 show the use of the word 'religion' in a positive context (i.e. where the writer evaluates religion positively)? Note down your answers.

Finally, let us focus on an area that has not yet received sufficient attention in corpus-based discourse studies, the inter-rater agreement. **Inter-rater agreement**, which is an estimate of how reliable and consistent a coding is, should be reported in studies working with a **judgement variable**. This is a variable that involves categorization or evaluation of cases (e.g. concordance lines) by the analyst that might bring an element of subjectivity into the study. The larger the element of subjectivity, the larger the need for double coding and reporting of inter-rater agreement. For example, if we want to group the occurrences of the word *time* into ten different semantic categories (think of dictionary definitions), this would involve a certain amount of subjectivity because semantics is notoriously fuzzy. Should *time* in a particular context be categorized as X or should it be subsumed under Y? Another example is the categorization from the 'Think about' task where you were asked to decide which concordance lines show the use of

Table 3.10 BE06: selected concordances for 'religion'

	Left context	Node	Right context	P/N ^a
1	use it to pursue their own needs, don't blame the	religion.	A lot of my friends have faced racism from white	
2	it's no wonder the confused flock to fundamentalist	religion,	which brooks no deviation from its rigid truth. And	
3	the war on terror – as by internal conflicts of class,	religion	and ethnicity. Closely examined, Muslim societies	
4	dissolve our complacent, parochial notions about	religion,	democracy, secularism and capitalism. They	
5	who are also authority figures within an organized	religion,	have the right to speak freely in the public square	
6	discrimination as per the Employment Equality	(religion	or Belief) regulations. Iqbal Sacranie, Secretary-	
7	crime, in 50% of cases, the actual or perceived	religion	of the victim was Islam. The criticisms of those who	
8	and I want him to stay that way". A	religion,	old or new, that stressed the magnificence of the Universe	
9	"He's positive science is incompatible with	religion,	but he waxes ecstatic about nature and the universe	
10	two different attitudes towards worship.	religion	is that piety or reverence that emerges from the	
	20 True			

^a P = Positive; N = Negative

the word *religion* in a positive context and which in a negative context. Clearly, your understanding of what religion is as well as your evaluation of religion may have an impact on judging the examples as positive or negative (see the discussion below). On the other hand, situations such as categorizing sentences according to the grammatical categories of the verbs into active and passive constructions involves relatively little subjective judgement and doesn't require double coding. Such situations also usually lend themselves to automatic methods; in a part-of-speech-tagged corpus, active and passive constructions can be searched for and counted automatically.

So how is inter-rater agreement approached? Before calculating inter-rater agreement, we need to find a second rater, another researcher on the project or a colleague who is willing to help. The first step is to carefully explain the coding system to them – usually a written coding scheme with concrete examples which both raters can refer to is a good idea. We then ask the second rater to independently code the same dataset or (especially if the dataset is large) a random sample taken from the dataset. Finally, using an appropriate statistical procedure, we can estimate how reliable our coding is or, in other words, how much subjectivity is involved in the coding.

Table 3.11 Double coding: concordances from the 'Think about' task

No.	Religious person	Atheist	(Dis)agreement
1	1	1	AGREE
2	0	0	AGREE
3	1	0	DISAGREE
4	0	0	AGREE
5	1	0	DISAGREE
6	1	1	AGREE
7	1	0	DISAGREE
8	1	1	AGREE
9	0	0	AGREE
10	1	0	DISAGREE
Positive	7	3	—
Negative	3	7	—
Total	10	10	

Before discussing the details of the statistical procedure, let's have a look at some data. Let's assume that we have asked two raters, a religious person and an atheist, to code the concordance lines from the 'Think about' task. The results can be seen in Table 3.11. '1' symbolizes a positive evaluation, '0' means a negative rating of the example.

We can see that the religious person and the atheist agree in six cases and disagree in four cases. These four cases of disagreement represent situations in which the context does not provide enough evidence about the evaluation of the example and the raters therefore rely on their own understanding of the situations. The first statistic that we can calculate easily is the raw agreement. **Raw agreement** is a metric, often expressed as a percentage, which provides the proportion of agreement cases in all cases. Raw agreement is calculated as follows:

$$\text{raw agreement} = \frac{\text{cases of agreement}}{\text{total no. of cases}} \quad (3.8)$$

For the example of the two raters, the raw agreement is:

$$\text{raw agreement} = \frac{6}{10} = 0.6 \quad (3.9)$$

We can thus say that in the example above, the raw agreement is 0.6 or 60%, which is a relatively low number. Ideally, we would be aiming at agreement of 80% and above.⁴ The judgement variable as operationalized in the example is therefore highly problematic because it involves a large amount of subjectivity,

⁴ However, we need to bear in mind that there is no magic number beyond which (raw) agreement becomes universally acceptable. The analyst always needs to evaluate the nature of disagreement and the robustness of coding in the context of a particular study.

which can be seen from the fact that the two raters following the same instructions came to different conclusions in 40% of cases.

Raw agreement is a useful first approximation of the reliability of the rating. However, we need to consider the fact that some agreement between the raters would be achieved even if both raters coded the cases randomly. More sophisticated inter-rater agreement measures such as Cohen's Kappa (κ) or Gwet's AC₁ therefore take this agreement by chance into account and subtract it from the raw agreement. Cohen's κ has been traditionally used for nominal variables; recent studies (e.g. Gwet 2002), however, recommend Gwet's AC₁ statistic. Both Cohen's κ and Gwet's AC₁ are based on the same equation (3.10). They differ, however, in the estimation of the agreement by chance.

$$\text{Cohen's } \kappa/\text{AC}_1 = \frac{\text{raw agreement} - \text{agreement by chance}}{1 - \text{agreement by chance}}, \quad (3.10)$$

where agreement by chance is calculated for the two measures as:

κ	agreement by chance = chance of being categorized as X by both raters + chance of being categorized as Y by both raters
AC_1	agreement by chance = $2 \times$ chance of being categorized as X \times chance of not being categorized as X

In the example above, Cohen's κ is calculated as:

$\text{agreement by chance} = 0.7 \times 0.3$ $+ 0.3 \times 0.7 = 0.42$	Note: 0.7×0.3 is the mathematical expression of the chance of positive rat- ing by the first rater (7 out of 10) and by the second rater (3 out of 10). 0.3×0.7 is the mathematical expression of the
--	---

$\kappa = \frac{0.6 - 0.42}{1 - 0.42} = 0.31$	(3.12) chance of negative rating by both raters.
---	--

In the example above, AC₁ is calculated as:

$\text{agreement by chance} = 2 \times \frac{10}{2 \times 10} \times \left(1 - \frac{10}{2 \times 10}\right)$ $= 0.5$	Note: In $\frac{10}{2 \times 10}$, 10 is the number of positive ratings by both raters (7 + 3) and 2×10 is the number of ratings by both raters. $\left(1 - \frac{10}{2 \times 10}\right)$ is the complementary probability, i.e. the probability of not being categorized as positive by chance.
--	---

$\text{AC}_1 = \frac{0.6 - 0.5}{1 - 0.5} = 0.2$	(3.14)
---	--------

Table 3.12 Overview of inter-rater agreement measures

Type of judgement variable	No. of values	No. of raters	Statistic(s) to use
Nominal (categories)	2 and more	2	Gwet's AC ₁ and Cohen's κ
	2 and more	3 and more	Gwet's AC ₁ and Fleiss's κ
Ordinal (ranks)	2 and more	2 and more	Gwet's AC ₂
Interval/Ratio (scale)	range	2 and more	Interclass correlation (ICC)

Both measures, κ and AC₁, produced low numbers: 0.31 and 0.2 respectively. These are closer to 0 (which is the baseline for random agreement) than to 1 (which shows absolute/perfect agreement).

The scale below can help interpret the results of κ ⁵ and AC₁. 0.67 and 0.8 as the cut-off points for agreement and very good agreement respectively are based on Krippendorff's (2012 [1980]) recommendations for content analysis of texts. Krippendorff talks about a tentative conclusion for values between 0.67 and 0.8 and a definite conclusion for values above 0.8. These, as any effect size cut-off points need to be, are interpreted as recommendations based on experience with rater behaviour in a particular field, not God's truth.



In addition, κ and AC₁ as effect size measures can be complemented with a p-value which will take into account the sample size (number of cases rated) and will tell us whether there is a statistically significant agreement between the raters. The null hypothesis we are testing is: there is agreement by chance alone. If the p-value is smaller than (0.05), conventionally we would reject the null hypothesis and say that the agreement is not due to chance.

So far, we've looked at a simple situation where we have two raters who have been coding a dataset using two categories (positive or negative). However, if we have more raters or a different number of categories or the rating is done using ranks or numeric values that can be placed on a scale, we need to select different measures of rater agreement. Table 3.12 shows the statistics that can be used in such situations. All these statistics are implemented in the Agreement tool in Lancaster Stats Tools online.

⁵ Note that κ does not have a straightforward minimum value. This, however, does not matter because it is used for evaluating rater agreement, not disagreement, and we thus look at the values larger than 0.

Reporting Statistics

1. What to Report

The following information should be provided for the reader to evaluate the reliability of coding of a judgement variable: (i) number of raters, (ii) amount of data coded (the whole dataset or random sample), (iii) inter-rater agreement measure, (iv) p-value and (v) interpretation of the result. The information described above should be reported in the Method section of the research report.

2. How to Report: An Example

- Following the coding scheme described above, two independent raters coded a random sample of 100 concordance lines from a total of 1,053 containing the word *religion* in the corpus. Gwet's AC₁ measure showed agreement between the raters (AC₁ = 0.7, p < 0.001). A review of the differences between raters found no systematic pattern of disagreement. Given the nature of the judgement variable, the amount of agreement was deemed sufficient.

3.6 Application and Further Examples: What Do Readers of British Newspapers Think about Immigration?

This section illustrates the statistical procedures introduced in this chapter in the context of a short discourse analysis study. The study focuses on the perception of ‘East European immigrants’ by readers of two British newspapers, the *Guardian* (a ‘heavyweight’ newspaper, politically leaning to the left) and the *Daily Mail* (a right-wing mass-market newspaper), using two corpora based on readers’ comments below articles on immigration on the newspaper websites. It is assumed that because the two respective newspapers attract a different readership, the analysis of the comments will reveal different perspectives on immigration.

At this stage, it might be useful to provide some historical context for the study. In January 2014, Britain opened its job market to citizens from Romania and Bulgaria. In the run-up to this event, the British press frequently debated the possible impact of this decision on the British economy and quality of life in Britain. In the media, comparisons were also made with a previous event ten years earlier (2004) when the job market opened to citizens of Poland, Hungary, the Czech Republic and Slovakia.

Table 3.13 *Keywords*

<i>Guardian</i> corpus	<i>Daily Mail</i> corpus
Guardian	UKIP
Balls	THE
Russia	DM
Duffy	homeless
economic	benefits
argument	police
debate	NOW
white	TO
Russian	squatters
post	NO

The data covers the period from 2010 to 2013. All articles in the *Guardian* and the *Daily Mail* containing the search term ‘east europeans’ or ‘eastern europeans’ were identified and the reader comments on these articles were extracted. ‘East (ern) Europeans’ is a collective term frequently used by the British press to refer to people from new European Union countries (e.g. Romania, Bulgaria, the Czech Republic or Poland). Overall, 942,232 tokens were extracted from the *Guardian* (GU corpus) and 2,149,493 from the *Daily Mail* (DM corpus).

First, to show an overall difference between the two corpora, top ten positive keywords were identified for both the GU and DM corpora. When extracting keywords for one of the newspapers, the comments of the readers from the other newspaper acted as a reference corpus in order to highlight words specific to the *Guardian* or *Daily Mail* readership. For the identification of keywords, Kilgarriff’s (2009) SMP was used with the constant 100; no frequency cut-off points were applied. The keywords are displayed in Table 3.13.

Apart from the two ‘obvious’ keywords *Guardian* and *DM* referencing the two newspapers, we can see an interesting pattern of differences between the two corpora. While the keywords in the GU corpus appear more neutral and related to the theoretical aspects of the immigration debate (*economic*, *argument*, *debate*), the keywords found in the DM corpus predominately point to negative aspects of immigration (*homeless*, *benefits*, *police*, *squatters*). The emotional intensity in the discourse of *Daily Mail* readers can also be seen from the frequent use of capitalization⁶ (THE, NOW, TO, NO) as illustrated in the examples below:

- (1) NO NO NO NO WAY. ENOUGH. (DM, 25/04/2011)
- (2) TOTAL DESTRUCTION OF THE UK (DM, 03/03/2011)

⁶ Note that the keywords were extracted as case sensitive. In standard texts, the case can usually be ignored because it does not convey any linguistic or social meaning. Keywords are therefore usually extracted as case insensitive. However, in online discussion forums capitalization is one of the means of emphasis.

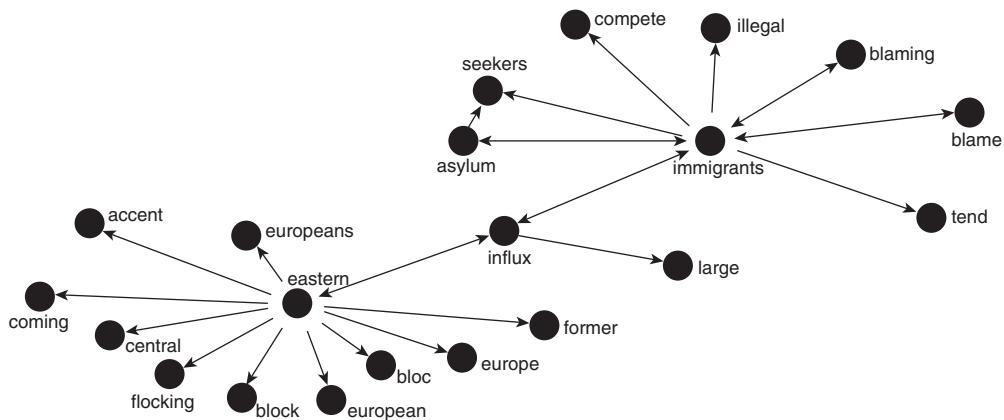


Figure 3.6 Collocation networks around 'immigrants' in the *Guardian* (3a–MI(6), R5–L5, C10–NC10; no filter applied)

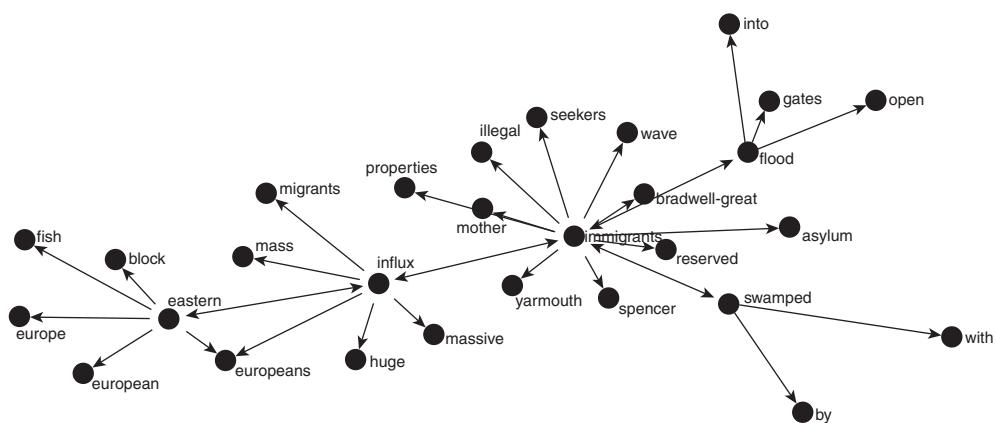


Figure 3.7 Collocation networks around 'immigrants' in the *Daily Mail* (3a–MI(6), R5–L5, C20–NC20; no filter applied)

As the next step, a collocation network with the initial node 'immigrants' was built for each corpus. The results can be seen in Figures 3.6 and 3.7.

Figure 3.6 shows the construction of immigrants in the discourse of the *Guardian* readers. The immediate collocates point to the coordination of the word *immigrants* with *illegal* and *asylum seekers*. Although word co-occurrence typically implies shared meaning, we can see from the examples below that *Guardian* readers often make a clear distinction between

legal immigrants and asylum seekers on the one hand and illegal immigrants on the other hand. The coordination thus in this context presents these categories as mutually exclusive.

- (3) Rarely is the distinction made between asylum seekers, immigrants and illegal immigrants. Personally, I have no time for people who easily take a swipe at hard working low-paid legal migrants who often take jobs that unemployed UK citizens sometimes find unpalatable. (GU, 29/04/2010)
- (4) There you go again. Is “immigrants and asylum seekers” some kind of single entity to you? (GU, 29/03/2010)

In addition, the collocates *blame* and *blaming* also stand out in the *Guardian* debate. Although *blame* and *blaming* are words with negative semantic prosody, here they critically reflect the tendency to use immigrants as scapegoats for different social issues which many *Guardian* readers perceive as unjustified – see the example below.

- (5) Sure there are issues, but blaming immigrants for everything isn’t going to address the real issues is it? (GU, 06/06/2010)

When we compare Figures 3.6 and 3.7, the most distinct difference can be seen in the labels used to denote the quantity of the immigrants coming to Britain. While the *Guardian* readers largely use the term *influx*, the *Daily Mail* readers also use *flood*, *wave* and *swamped*.

To investigate if the word *immigrant(s)* was used predominantly in a positive or negative context in the two reader discourses, a random sample of 100 concordance lines was extracted from each corpus and manually coded by two raters on a 5-point Likert scale from very positive (1) to very negative (5). Inter-rater agreement (DM: AC₂ = 0.93; GU: AC₂ = 0.8) was deemed high. The summary of the reader evaluations is shown in Table 3.14.

We can see that *Guardian* readers use the term *immigrant(s)* in more positive than negative contexts, while the *Daily Mail* readers use very few positive evaluations of *immigrant(s)*; the DM corpus is thus dominated by negative and very negative evaluations (together over 50% of comments).

Table 3.14 *Evaluations of ‘immigrant(s)’ in the GU and DM corpora*

	1 (very positive)	2 (positive)	3 (neutral)	4 (negative)	5 (very negative)
GU	4%	31%	39%	26%	0%
DM	2%	6%	40%	45%	7%

3.7 Exercises

Collocations

- Which association measures would you use in the following research scenarios? Note that more than one answer is possible in each case – think of your rationale for the answer you choose.
 - You need to identify technical terms connected with the word *process* in a corpus of research articles on organic chemistry, e.g. *petrochemical process*. Note that technical terms are exclusive and relatively rare combinations of words with a specific meaning.
 - You want to study the associations that the word *enemy* (node) has in the newspaper discourse. You are interested to see content words around the node rather than frequent grammatical words.
 - You want to write a dictionary of collocations for learners of English that would include a broad range of fixed expressions such as *find out*, *take responsibility*, *dire consequences* etc. The collocations you include need to be recognizable as specific meaningful units and they need to occur as frequent combinations.
- Look at the information in Table 3.15 about the co-occurrence of the word *issue* in an L3–R3 collocation window in BE06, a one-million-word corpus of written English. Use the online *Collocation Calculator* to calculate four association measures: MI, LL, Delta P and log Dice.
 - Number of tokens in the whole corpus (N): 1,001,514
 - Frequency of the node in the whole corpus (R_1): 164
 - Collocation window size: 6 (3L, 3 R)

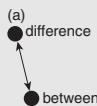
Table 3.15 *Collocates of issue in BE06*

Collocate	C_1	O_{11}	MI value	LL value	Delta P values	log-Dice value
the	58,591	101				
this	4,815	38				
important	322	7				
address	88	6				
bbc	98	5				
HUPO-PSI	1	1				

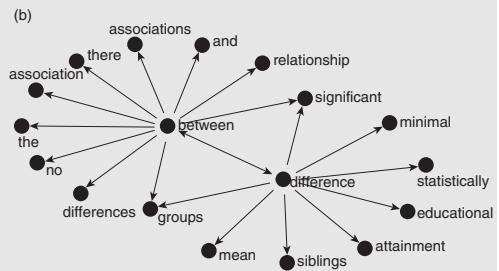
- Discuss how the association measures from Exercise 2 rank the six collocates. Which association measure would you choose?

Collocation Networks

- Compare the pairs of collocation networks in Figure 3.8 based on (a) the BE06 non-academic subcorpus, an 840,000-word sample of written British English ranging from newspapers and general prose to fiction, and (b) the BE06 academic subcorpus, which consists of over 160,000 words of academic English. Note that the BE06 non-academic subcorpus is more than five times larger than its academic English counterpart. Pay attention to the frequencies of the initial node and the CPN parameters, especially the cut-off points and their effect on the collocates that are shown in the graphs.

BE06 – non-academic (840k)

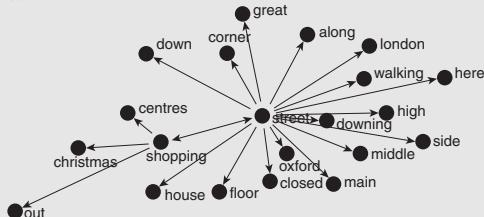
AF (node 'between'): 641

BE06 – academic (160k)

AF (node 'between'): 482

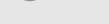
10a-log Dice(10), R5-L5, C5-NC5; no filter applied

(a)



AF (node 'shopping'): 72

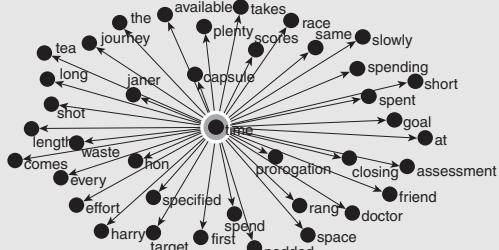
(b)



AF (node 'shopping'): 1

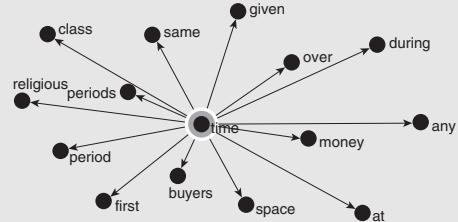
3a-log MI(5), R5-L5, C5-NC5; no filter applied

(a)



AF (node 'time'): 1,444

(b)



AF (node 'time'): 210

3a-log MI(5), R5-L5, C5-NC5; no filter applied

Figure 3.8 Selected collocation networks

5. Use #LancsBox, which is downloadable from <http://corpora.lancs.ac.uk/lancsbox>, to build collocation networks based on the LOB corpus (available via #LancsBox). LOB is a one-million-word corpus representing written British English of the 1960s.

Nodes to search for:

- *university*
- *time*

Compare the collocation networks of *time* and *university* based on LOB with the collocation networks built using BE06, which represents British English around 2006, shown in Section 3.3. Is there any difference/indication of language development?

Keywords

6. Review the following situations and decide upon an appropriate type of the reference corpus (e.g. general language corpus, specialized corpus representing . . .). Justify your answer.
- (a) In a literary stylistic study, we compiled a corpus of all works by a certain author; we want to identify keywords typical of this author of interest.
 - (b) We are interested in keywords typical of the genre of academic writing. We have compiled a corpus of research articles and books in multiple disciplines representing all major academic fields.
 - (c) We are interested in keywords typical of spoken language. Our corpus of interest is the spoken part of the *British National Corpus*.
7. Calculate the SMP statistic for the words in Table 3.16. Decide which of the words belongs to (i) positive keywords (+), (ii) negative keywords (-) and (iii) lockwords (0).

Table 3.16 *Keywords*

Word	C (tokens: 1,007,532)	R (tokens: 1,017,879)	SMP (simple maths parameter)	Decision (+/-/0)
BBC	106	3		
before	970	854		
London	471	119		
nation	51	195		
she	4,162	4,494		
slowly	83	94		
today	270	278		
tomorrow	47	48		
Washington	27	222		
which	2,680	2,056		

Inter-rater Agreement

8. The following ratings were obtained in three situations involving a judgement variable. Calculate the inter-rater agreement in each situation.

- i Situation 1: In a discourse analysis study, a judgement variable with three possible values (1, 2 and 3) was coded by three independent raters. The variable of interest was a nominal variable capturing a discourse category.

Rater A: 2, 1, 1, 2, 1, 1, 3, 3, 2, 2, 3, 1

Rater B: 2, 1, 2, 2, 1, 1, 2, 2, 2, 2, 3, 1

Rater C: 2, 1, 1, 2, 1, 2, 2, 2, 2, 3, 1

- ii Situation 2: In an applied linguistic study, texts from second language speakers were used. Based on the texts, the proficiency of the second language speakers was coded using hierarchically ordered categories (ordinal variable) ranging from 1 (lowest proficiency) to 6 (highest proficiency). A random sample of 20 per cent of the texts was double coded to assess the robustness of the coding.

Rater A: 4, 4, 4, 3, 4, 4, 3, 3, 4, 4, 3, 3, 2, 4, 4, 4, 3, 4, 4, 4

Rater B: 4, 4, 4, 3, 4, 3, 3, 3, 3, 4, 4, 5, 2, 5, 5, 4, 4, 4, 4, 5

- iii Situation 3: Two transcribers were given the same recording to transcribe. It contains a spoken interaction between six different speakers. Because speaker attribution in a dialogue between multiple speakers is notoriously difficult, the reliability of the speaker codes (1 to 6) at the beginning of each turn was checked by an inter-rater agreement measure.

Transcriber A: 1, 4, 5, 4, 3, 4, 2, 4, 1, 2, 6, 1, 4, 2, 1, 6, 1, 6, 4, 1

Transcriber B: 1, 4, 5, 4, 3, 4, 2, 4, 1, 2, 6, 2, 4, 6, 2, 4, 2, 4, 6, 2

9. Look at the examples in Table 3.17 taken from the *Trinity Lancaster Corpus*. They show how speakers of English as a foreign language express disagreement. Decide how polite (or impolite) these speakers are when they express disagreement. Use the following rating on a 5-point Likert scale:

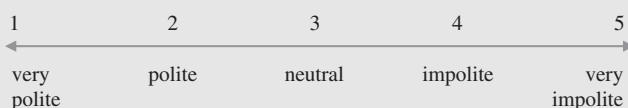


Table 3.17 Examples for rating

Example	Rating
(a) I completely disagree with this because er I I repeat as I said . . .	
(b) I agree with this point but don't you think maybe the ti= fact that times are changing is a good thing?	
(c) but I personally would disagree that that money would necessarily be spent on that	
(d) erm no no it's not so	
(e) well I'm not totally convinced but er you know I live in a really traditional family	
(f) mm I can understand your opinion erm but I was still wondering . . .	
(g) I can't agree with you	
(h) er er I I think erm I I think they I I think they are wrong	
(i) I think they're completely wrong	
(j) no way	
(k) I think he's stupid	
(l) I I I can understand what you're saying but I'm not I don't agree with that	

After the rating, answer the following questions:

- How confident are you about the ratings you have provided?
 - Would you consider politeness a robust judgement variable?
 - How important do you think it is to have another rater for this judgement variable?
10. Compare your coding in Exercise 9 with the coding of the same dataset by a different rater (e.g. ask a friend to help you with this exercise). Using the Agreement calculator, calculate the appropriate agreement measure.

Measure calculated: _____, Value: _____

- If available, keep adding more raters and calculating the inter-rater agreement.
11. Imagine you need to produce a research report based on the dataset discussed in Exercises 9 and 10. Report the results of the inter-rater agreement measure from Exercise 10. Refer back to the ‘Reporting statistics’ box.

THINGS TO REMEMBER

- There are many association measures each highlighting different aspects of the collocational relationship (e.g. frequency or exclusivity). There is no one best association measure.
 - Collocations can be presented in a tabular (table) or visual form (graph).
 - Collocation networks show complex cross-associations in texts and discourses.
 - The keyword procedure is in essence a comparison which depends on a number of parameters. There is no such thing as one set of keywords.
 - For judgement variables an inter-rater agreement statistic should be reported. Gwet’s AC₁ and AC₂, Cohen’s and Fleiss’s as well as interclass correlation can be used depending on the situation.
-

Advanced Reading

- Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context. *International Journal of Corpus Linguistics*, 20(2), 139–73.
- Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (eds.), *Corpus linguistics: an international handbook*, vol. 2, pp. 223–33. Berlin: Walter de Gruyter.
- Evert, S. & Krenn, B. (2001). Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, pp. 188–95.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378.
- Gablasova, D., Brezina, V. & McEnery, A. M. (2017). Exploring learner language through corpora: comparing and interpreting corpus frequency information. *Language Learning*, 67(S1), 130–54.
- Gries, S. Th. (2013). 50-something years of work on collocations: what is or should be next . . . *International Journal of Corpus Linguistics*, 18(1), 137–66.
- Gwet, K. L. (2014). *Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among raters*. Gaithersburg, MD: Advanced Analytics.
- Kilgarriff, A. (2012). Getting to know your corpus. In *Text, Speech and Dialogue*, pp. 3–15. Berlin: Springer.
- Scott, M. (1997). PC analysis of key words – and key key words. *System*, 25(2), 233–45.
- Sinclair, J., Jones, S. & Daley, R. (2004). *English collocation studies: the OSTI report*. London: Continuum.

Companion Website: Lancaster Stats Tools Online

1. All analyses shown in this chapter can be reproduced using the statistical tools available from the companion website. In particular, the available tools include:
 - Collocation calculator
 - #LancesBox
 - Keywords calculator
 - Agreement calculator
2. The website also offers additional materials for students and teachers.

4 Lexico-grammar

From Simple Counts to Complex Models

4.1 What Is This Chapter About?

This chapter focuses on the statistical analysis of lexico-grammatical features in language (such as articles, passive constructions or modal expressions). We start with a discussion of two types of approaches to lexico-grammar¹ in corpora. The first approach uses the ‘Whole corpus’ research design and compares the frequencies of a linguistic variable (and its variants) in broadly defined subcorpora. The second approach employs the ‘Linguistic feature’ research design and carefully defines the contexts in which a particular variable can occur (i.e. its lexico-grammatical frame) and analyses factors which contribute to the occurrence of one variant of the variable as opposed to another. Following the second approach, the chapter shows how lexico-grammatical variation can be summarized using cross-tabulation and what statistical measures can be computed based on cross-tabulation summary tables. These measures range from simple percentages to the chi-squared test and logistic regression. Since logistic regression represents an advanced statistical procedure, large parts of the chapter are devoted to explaining this method and the interpretation of its output. In this chapter, we’ll be exploring the answers to four questions:

- How can lexico-grammatical variation best be described? What types of research design can be used? (Section 4.2)
- How can lexico-grammatical variation be summarized and what simple statistical measures can be computed? (Section 4.3)
- How can we build complex models that account for multiple variables that predict lexico-grammatical variation? (Section 4.4)
- How can the statistical techniques discussed in this chapter be used in the analysis of lexico-grammar? (Section 4.5)

¹ In this chapter, the term ‘lexico-grammar’ is used for regularities in language that involve a large number of features along the cline between lexis and grammar explored using corpus techniques. Specifically, the features investigated in this chapter are identifiable within a particular lexico-grammatical frame (see the discussion in Section 4.2).

4.2 Analysing a Lexico-grammatical Feature

Think about ...

Before reading this section, think about the following situation:

A friend who is learning English shows you a sentence in a newspaper article entitled ‘Google unveils new logo at turning point in company’s history’.² The sentence reads: *The logo has undergone many, mainly small, changes in its history.* Your friend asks you: ‘Why is the definite article used in this sentence?’ Think about how you would answer this question. What grammatical rules explain the use of the definite article in this case?

When looking at lexico-grammar from a broad perspective, we can see that there is a large amount of variation related to the situations in which language is used. For example, even purely grammatical words, such as articles in English, which we might expect to be stable in language, show considerable variation in their distribution in different registers of spoken and written English (speech, fiction, newspapers, general writing and academic writing). The stacked bar chart in Figure 4.1 displays frequencies of *the* and *a/an* in different registers of the BNC. A **stacked bar chart** is a graphical representation of (relative) frequencies of a linguistic variable with multiple variants (*the* and *a/an* in Figure 4.1) in different parts of a corpus (subcorpora). This type of visualization is useful for comparisons of the distribution of a linguistic feature. Such visualizations are often used, for example, in Biber et al.’s (1999) grammar, which discusses a number of lexico-grammatical patterns in different registers.

From Figure 4.1 we can see that the definite article is much more frequent than the indefinite article in all the BNC subcorpora. In addition, while the variation in the use of the indefinite article in the subcorpora is not striking, there are clear differences in the use of the definite article. Speech and fiction have considerably fewer definite articles than general prose and academic writing. Newspapers are somewhere in the middle between these two groups, with the definite article forming on average 6 per cent of the newspaper subcorpus.

We can now return to the ‘Think about’ task and reflect on the question using the evidence presented in Figure 4.1. Based on this evidence, we can comment on the distribution of the articles in different registers and even suggest that when in doubt it is a better bet to use the definite rather than the indefinite article regardless of the text type or register. But would these observations help us really answer the question asked in the ‘Think about’ task? Let’s consider the following: although we have explored the overall distribution of the articles in the BNC, we have so far paid little attention to the

² www.theguardian.com/technology/2015/sep/01/google-logo-history-new-doodle-redesign

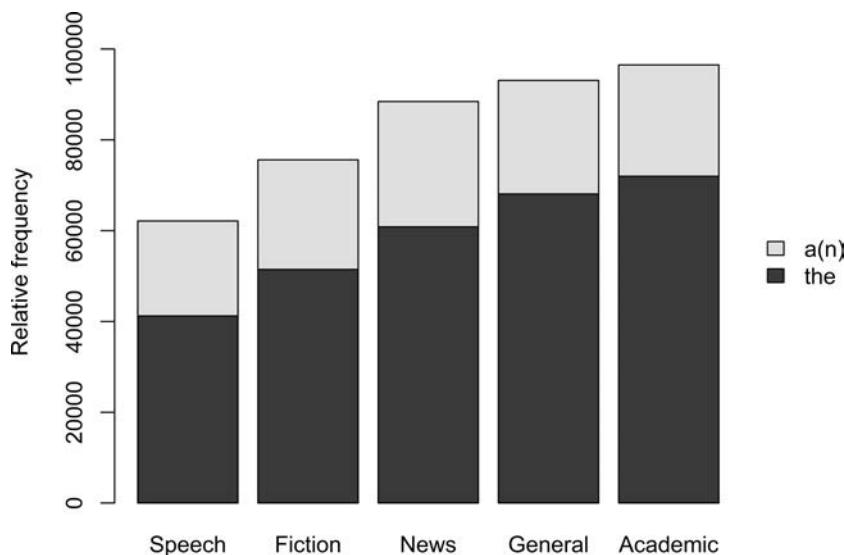


Figure 4.1 *The definite and indefinite articles in BNC subcorpora*

immediate linguistic contexts (co-texts³) in which articles in English occur. However, these contexts are crucial for a deeper understanding of lexico-grammatical features and the language-internal principles of their use. Such understanding will, in turn, help us answer the question of why a particular lexico-grammatical feature is used in a specific context, the question asked in the ‘Think about’ task about the definite article. If we relied only on the information presented in Figure 4.1, we might have concluded that the rules governing the use of the definite article are different in different registers. But why should the definite article be so preferred in academic writing, for instance? As it turns out, the high incidence of articles is directly related to frequent use of nouns: the more nouns there are in a text or register the more likely it is that the articles will appear in front of them. This is a slight simplification because nouns can be used as parts of compounds or modifiers of other nouns, in which case they don’t take an article. The articles in this case are indicators of broader functional variation among registers (see Chapter 5 for exploration of register variation in depth). So how can we explore, in a systematic way, the linguistic contexts in which the articles occur?

The first step is to use a different research design (see Biber & Jones 2009). So far, we have explored the BNC using the **Whole corpus design**. In order to bring

³ Literally, co-text refers to the actual words that surround the linguistic feature of interest. This term is used to refer to the immediate language context of the linguistic feature of interest, which can help us determine different properties of the linguistic feature such as its syntactic position or function; co-texts can be observed in a concordance.

File	Article_type	Context_type	Noun_typ	Left	KWIC	Right	
A0T	definite	non-determined	proper	y Grey Walter , who called it	<<< the >>>	Contingent Negative	
A0Y	definite	determined	count_pl	pported by a pillow , and all	<<< the >>>	natural outlets of th	
A1D	definite	determined	uncount	th his obsessive recording of	<<< the >>>	weather . Matteo Fa	
A6L	definite	determined	count_pl	he first bank that comes into	<<< the >>>	ratings is NatWest o	
A7D	definite	determined	uncount	ence , she was in England for	<<< the >>>	publication of her n	
A7L	definite	determined	count_sg	British filmmaking , nor was	<<< the >>>	film industry the on	
ABD	indefinite	non-determined	count_sg	debate in a journal , Nature ,	<<< a >>>	week before the co	
AJX	definite	determined	proper	lens promised by Labour and	<<< the >>>	Liberal Democrats .	
ALB	indefinite	non-determined	count_sg	n mind when they state : -- [<<< A >>>	framework for estab	
ALU	definite	determined	count_sg	of a beautiful red colour and	<<< the >>>	other half of a deep	
ASA	definite	determined	count_sg	could n't believe my luck . In	<<< the >>>	end none of us won	
ASL	definite	determined	count_sg	t where it starts will become	<<< the >>>	anus . The embryo h	

Figure 4.2 The vs a(n) dataset: linguistic feature design (an excerpt)

attention to the linguistic contexts in which the articles occur, the **Linguistic feature design** is required (see Section 1.4 for discussion of research designs). In this case, we will search for all occurrences of the definite and the indefinite article in the BNC and code each example for contextual features such as context type and noun type. The dataset produced is shown in Figure 4.2. Because there are too many examples in the BNC to be coded manually⁴ (over 8.5 million), we can take a random sample of, for example, 100, 500 or 1,000 instances and base our analyses on it.

The dataset in Figure 4.2 is an excerpt from a random sample of 100 concordance lines⁵ downloaded from the BNC. In the dataset, three nominal variables, that is variables which express unordered categories (see Section 1.3), were manually coded. These are: article type, context type and noun type. In addition, the dataset includes the file name and the concordance line itself, which was used in manual coding. Out of the three variables, the article type is our **linguistic variable** of interest. Sometimes it is also called an **outcome variable** because its value depends on the other two variables (context type and noun type), which are called **explanatory variables or predictors**. The term ‘predictor’ suggests that we can use these variables to predict the outcome, i.e. value of the outcome variable (definite vs indefinite article, in our case). How this predicting works will

⁴ Corpus tools can assist with coding of the examples in various ways. While some variables such as the type of article in Figure 4.2 can be automatically assigned to each observation, the more complex functional categories such as the type of context in which articles occur usually require a human coder who can evaluate each example holistically. Categories such as the type of noun (singular, plural, proper) that occurs after the article can be automatically pre-processed using the appropriate part-of-speech (POS) tag of the noun. However, manual checking of this automatic process is required because POS tagging is never 100% accurate. Also, the information about countability of nouns needs to be added manually.

⁵ This number of concordance lines was chosen for simple demonstration. In a real study, a larger dataset would be desirable.

become clear in Section 4.4, where a technique called logistic regression is introduced.

Let us look at some examples of the variation in the dataset in Figure 4.2. With regard to the type of context in which articles occur, the basic split is between contextually determined and contextually non-determined uses as coded in the dataset. The former represents cases where the article marks a person, object or abstract entity previously mentioned or implied (example 1) or specified in the context immediately following (example 2). On the other hand, contextually non-determined uses include reference to a person, object or entity mentioned for the first time and otherwise not specified (examples 3 and 4).

- (1) Lands were granted to a group of men known as feoffees, who became the legal owners of the land, while the grantor enjoyed the use of the lands – in other words, all the rights and profits arising from them. But because the feoffees were the legal owners, the lands could not be taken into wardship if the grantor died leaving an heir under age . . . (BNC, file: E9V)
- (2) In September, a month after the RSPCA conference, she was in England for the publication of her new book. (BNC, file: A7D)
- (3) The kit includes a fine brass pendulum and chain along with a detailed book to point you in the right direction. (BNC, file: CBC)
- (4) This effect, which is strongest over the frontal lobes, was first observed in 1964 by Grey Walter, who called it the Contingent Negative Variation. (BNC, file: A0T)

Before looking at the types of analyses we can carry out using a Linguistic feature design (see Sections 4.3 and 4.4), it is important to note some important features of lexico-grammatical variables and the space within which they operate. Lexico-grammatical variables as discussed in this chapter are variables which can be expressed as choices between two or more variants, such as the definite and the indefinite article. The realization of one variant as opposed to the other(s) depends on the linguistic context: one context favours one linguistic variant, another context another variant. To use a linguistic feature design, we need to be able to identify all contexts in which the linguistic (outcome) variable operates – this is called a **lexico-grammatical frame, envelope of variation or variable context** (Tagliamonte 2006: 70ff; Grieve-Smith 2007). For example, if we are interested in the choice between the definite and the indefinite article we need to identify all contexts in which these articles appear by searching for all articles in the corpus. These contexts form the grammatical frame of interest. However, if we were interested in a more general use of determiners in front of English nouns, in other words where to use not only articles but also other determiners such as *this, that, my, your* or no determiners at all, the grammatical frame would be different. In this case, we would have to search for all nouns and investigate whether each is preceded by a determiner and if so, what determiner it is.

Table 4.1 provides examples of lexico-grammatical features that can be investigated using the Linguistic feature research design. It shows the type of research

Table 4.1 Examples of lexico-grammatical variables with a grammatical frame

Research question	Outcome variable options	Lexico-grammatical frame
When do we use the passive construction?	ACTIVE, PASSIVE	All verb forms that can be used in passive, i.e. transitive verbs.
In what contexts do we use <i>which</i> and in what contexts <i>that</i> in relative clauses?	<i>which, that</i>	All relative clauses.
When do speakers use <i>that</i> deletion? E.g. <i>I think Ø this is good.</i>	<i>that, Ø [no relativizer]</i>	All clauses where <i>that</i> occurs or is deleted.
What is the difference between various modal expressions of strong obligation?	<i>must, have to, need to</i>	All contexts in which strong deontic modals occur.

questions that can be asked as well as the outcome variable options (choices that the speakers have) and the lexico-grammatical frame.

A word of caution: we need to realize that not all linguistic variables lend themselves to the type of research design described above. **Ambient linguistic variables** such as discourse markers, hesitations or swearwords that can appear in any possible context and do not have a clearly defined lexico-grammatical frame cannot be treated in the same way as the variables from Table 4.1. For these variables, Individual text/speaker research design is appropriate (see Section 6.3). The difference between a linguistic variable with a lexico-grammatical frame and an ambient linguistic variable is illustrated below. Whereas the passive in example (5) has a clearly defined lexico-grammatical frame (all verb forms with a transitive verb), the discourse markers, as is apparent from example (6), can appear at almost any place in an utterance. The variable with a clear lexico-grammatical frame has also a clear number of variants: all verb forms with a transitive verb can be either active or passive; there are, however, a large number of discourse markers not in a direct competition – they can appear repeatedly, some of them competing for the same syntactic slot, as example (6) illustrates.

(5) It's about time that was done. (BNC, file: KBB)

(6) Well, you know, it you see, time were, I don't know I suppose, I don't know but I never seemed to be afraid ... (BNC, file: HDK)

In sum, we have seen two approaches to the analysis of lexico-grammatical variation. One approach looked at general distributions of lexico-grammatical features in broadly defined subcorpora, while the other focused on individual linguistic contexts, in which lexico-grammatical features are used. The approach that would arguably be the most fruitful for answering the question in the ‘Think about’ task is the exploration of the linguistic contexts in which articles in English

are used (following the Linguistic feature research design). The statistical techniques employed in such an exploration are the topic of the remainder of this chapter.

4.3

Cross-tabulation, Percentages and Chi-squared Test

Think about ...

Before reading this section, think about the following questions.

1. Which of the following expressions do you say most often?

- *I must go.*
- *I have to go.*
- *I need to go.*

2. Can you think of contexts in which you would use each of them?

A good starting point for any data exploration is a simple summary table. In Chapter 3 (see Sections 3.2 and 3.4), we have seen contingency tables that were constructed for collocation and keyword analyses to show all possibilities (contingencies) of word (co)occurrence. **Cross-tabulation (cross-tab)**, which is explored in this section, is a similar technique (Hill et al. 2006: 32–8); it examines the relationship between **categorical variables** (nominal and ordinal variables), i.e. variables that are used to categorize observations (see Section 1.1). **Cross-tabulation tables**⁶ are created by the cross-plotting (hence cross-tabulation) of one linguistic variable and one or more explanatory variables. The simplest form of cross-tabulation is a 2×2 table (two main rows and two main columns) with one linguistic variable and one explanatory variable,⁷ each with two categories (sometimes called levels) as shown in Table 4.2.

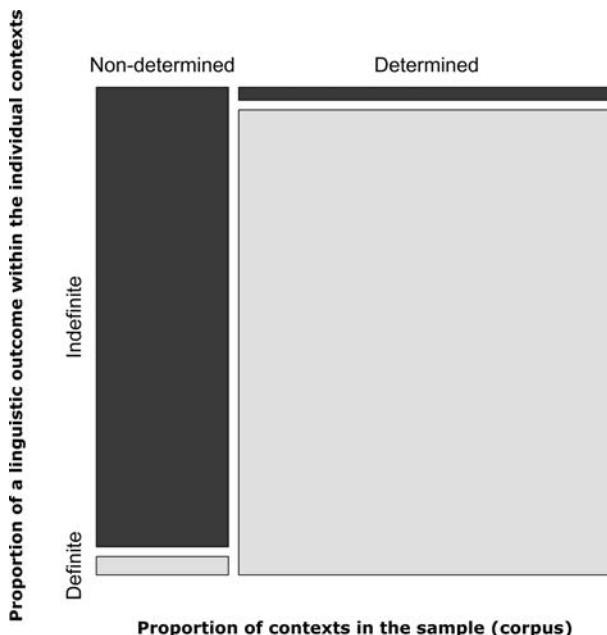
Table 4.2 displays article use in different types of context (see Section 4.2). The linguistic variable, ‘Article type’, can take on two possible values that are encoded as two categories (levels) of the variable: ‘Indefinite’ and ‘Definite’; the explanatory variable labelled as ‘Context type’ again has two categories (levels): ‘Non-determined’ and ‘Determined’. Conventionally, categories of the linguistic variable are listed as columns and the categories of explanatory variables are listed as rows. In addition, the table includes the **column** and the **row totals**

⁶ There are different terms used for referring to these tables: ‘cross-tabulation table’, ‘contingency table’, or ‘pivot table’. The last is a term used in spreadsheet programs such as Excel or Calc.

⁷ In statistical jargon, this would be called a two-way table because it includes two variables. We can also produce three-way, four-way (see Table 4.4) etc. tables with one linguistic and two, three etc. explanatory variables; however, the more complex the tables are, the more difficult it is to interpret them.

Table 4.2 *Cross-tabulation: article type by contextual determination*

Context type \ Article type	Indefinite	Definite	Total
Non-determined	25	1	26
Determined	2	72	74
Total	27	73	100

Figure 4.3 *A mosaic plot: article type by contextual determination*

(sometimes called the ‘marginal frequencies’) (light shaded in Table 4.2) as well as the **grand total** (dark shaded in Table 4.2). The cross-tabulation table includes all possible combinations of the variable categories (levels) and the frequency count in each subcategory (cell in the table). Table 4.2 lists the following combinations: (i) Contextually non-determined indefinite article, (ii) Contextually non-determined definite article, (iii) Contextually determined indefinite article and (iv) Contextually determined definite article.

A small aside: the information included in a simple cross-tab table can be visualized in a mosaic plot (Figure 4.3). A **mosaic plot** turns the frequency information into the size of the areas in the plot. In addition, it also displays the proportions of the categories of the predictor variable in the corpus and the proportions of the values of the linguistic variable (outcomes) within each predictor category. In Figure 4.3, we can clearly see that the largest area

(light-shaded large rectangle) represents the combination of contextually determined definite article with 72 observations. The smallest area (light-shaded small rectangle), on the other hand, represents contextually non-determined definite article with one observation. The widths of the rectangles show the proportion of each context type in the sample or corpus (Non-determined: Determined = 26: 74) and their heights show the proportion of outcomes (definite and indefinite articles) within each context type (predictor) category. Note that the presentation of the data in a mosaic plot is the reverse of the cross-tab table: in the mosaic plot, the categories of the explanatory variable are listed horizontally, while the categories of the linguistic variable are displayed vertically. For more information see Theus and Urbanek (2008: 50ff) and Friendly (2002).

Returning to the cross-tab table: often, in addition to frequency counts, cross-tabulation tables include percentages for easier comparison across categories. The percentages can be computed from row totals, column totals or the grand total according to the following equation:

$$\text{percentages in a cross-tabulation table} = \frac{\text{cell value}}{\text{relevant total}} \times 100 \quad (4.1)$$

It is important to note that each of the three percentage options (percentages out of row total, column total or grand total) has a completely different interpretation and is useful for a different type of comparison. For example, the percentage out of row total of the first cell in Table 4.2 (contextually non-determined indefinite article) is calculated as follows:

$$\% \text{ non-determined contexts with an indefinite article} = \frac{25}{26} \times 100 = 96.2\% \quad (4.2)$$

We can say that out of all non-determined contexts a large majority (96.2%) prefer the indefinite to the definite article. In other words, in this context, there is a 96.2% (or 0.962) probability of occurrence of the indefinite article. More generally, we can interpret the percentage based on the row total as the probability (preference where over 50% or dispreference where less than 50%) of a given variant of the linguistic variable in a given context.

Alternatively, we can calculate the percentage based on column totals. Using the same example from Table 4.2 we'll get the following:

$$\% \text{ indefinite articles occurring in non-determined context} = \frac{25}{27} \times 100 = 92.6\% \quad (4.3)$$

We can say that out of all indefinite articles, 92.6% occur in non-determined contexts. Although this may sound similar to the previous option, the logic of this statement is different. Here, we don't compare the two variants of the linguistic variable but two contexts (determined and non-determined).

Table 4.3 *Percentage options in cross-tabulation*

Relevant total	Interpretation	Use
Row	Probability of one variant of the linguistic variable in a given context	Comparison of the use of different linguistic variants in a particular context
Column	Probability of one context with a given type of a linguistic variant	Comparison of different contexts
Grand	Representation of different sub-categories in the corpus	Corpus description

Finally, we can calculate the percentages from the grand total as follows:

$$\% \text{ contextually non-determined indefinite articles in the corpus} = \frac{25}{100} \times 100 = 25\% \quad (4.4)$$

Note that because the grand total in Table 4.2 is 100, the percentages out of the grand total will be the same as the actual values in the cells. These percentages reflect the representation of different subcategories (e.g. contextually non-determined indefinite article) in the corpus. Such percentages are typically indicative of the sampling of the corpus rather than any inherent principles of language or grammar. Table 4.3 provides a summary of the options for calculating percentages in cross-tabulation, their interpretation and use. Clearly, for the purposes of comparison of lexico-grammatical features, percentages based on row totals are appropriate.

So far, we have seen only a very simple form of cross-tabulation (2×2 table). With a larger number of explanatory variables (and their categories) we can create more complex cross-tabulation tables. In the ‘Think about’ task, you were asked to think of the contexts for the use of *must*, *have to* and *need to*. All of these are modal expressions with a similar meaning indicating strong obligation, in other words, indicating that something is necessary or should be done. Table 4.4 displays different contexts in terms of the variety of English, genre and the subject (first-person pronoun, second-person pronoun or other) of the modal expression. Every row of the table indicates the use of *must*, *have to* and *need to* in a specific context such as American academic writing with ‘I’ as the subject (first row). The percentages in Table 4.4 were calculated out of row totals and can be interpreted as probabilities of each of the three modal expressions in a particular context determined by the predictor combinations (specific variety, genre and subject).

The last point to discuss in this section is the appropriate test for statistical significance that can be used with cross-tabulation; a statistical significance test evaluates the amount of evidence against the null hypothesis (see Section 1.3).

Table 4.4 *Strong (semi)modals in different genres of British and American English: cross-tabulation*

Variety	Genre	Subject	Modal			Total
			must	have to	need to	
AM	Academic	I	0 (0.0%)	1 (100.0%)	0 (0.0%)	1
		you	3 (33.3%)	1 (11.1%)	5 (55.6%)	9
		other	63 (64.3%)	18 (18.4%)	17 (17.3%)	98
	Fiction	I	8 (12.7%)	34 (54.0%)	21 (33.3%)	63
		you	16 (24.6%)	35 (53.8%)	14 (21.5%)	65
		other	51 (28.5%)	87 (48.6%)	41 (22.9%)	179
	General	I	8 (19.5%)	27 (65.9%)	6 (14.6%)	41
		you	4 (10.0%)	24 (60.0%)	12 (30.0%)	40
		other	152 (56.5%)	72 (26.8%)	45 (16.7%)	269
	Press	I	1 (50.0%)	1 (50.0%)	0 (0.0%)	2
		you	2 (20.0%)	7 (70.0%)	1 (10.0%)	10
		other	44 (33.6%)	48 (36.6%)	39 (29.8%)	131
BR	Academic	I	0 (0.0%)	0 (0.0%)	1 (100.0%)	1
		you	1 (100.0%)	0 (0.0%)	0 (0.0%)	1
		other	37 (39.8%)	23 (24.7%)	33 (35.5%)	93
	Fiction	I	19 (28.8%)	34 (51.5%)	13 (19.7%)	66
		you	14 (27.5%)	29 (56.9%)	8 (15.7%)	51
		other	88 (41.3%)	107 (50.2%)	18 (8.5%)	213
	General	I	9 (23.1%)	27 (69.2%)	3 (7.7%)	39
		you	21 (32.8%)	21 (32.8%)	22 (34.4%)	64
		other	193 (53.0%)	105 (28.8%)	66 (18.1%)	364
	Press	I	4 (25.0%)	12 (75.0%)	0 (0.0%)	16
		you	2 (18.2%)	7 (63.6%)	2 (18.2%)	11
		other	60 (48.4%)	40 (32.3%)	24 (19.4%)	124
Total			800	760	391	1951

Because cross-tabulation works with categorical data, a test called **chi-squared** (often written as χ^2) can be used (Balakrishnan et al. 2013; Azen & Walker 2011: 58–9; Sheskin 2007: 493–561). Chi-squared is appropriate for simple tables with one linguistic and one explanatory variable such as Table 4.2. Before using the test, we need to check the assumptions and prerequisites of appropriate use. Chi-squared has two assumptions:

1. **Independence of observations.** We assume that every observation such as the use of an article in English (see Table 4.2) is independent of another observation. However, with corpora we need to relax this assumption (there is, of course, a price to pay for this): when we look at texts in corpora, we see that linguistic features in the same text are connected with one another. A corpus, no matter how well sampled, is not a (random) sample of linguistic

features (or words) but a sample of texts which combine a number of interconnected linguistic features (Kilgarriff 2005). This means that by definition, the assumption of independence of observations is violated to some extent. This violation may lead to an increased number of falsely significant results, i.e. ‘false hits’.

2. **Expected frequencies greater than 5** (in contingency tables larger than 2×2 at least 80% of expected frequencies greater than 5). Expected frequencies are baseline frequencies that we calculate as part of the chi-squared procedure (see the next paragraph). With corpus datasets, which are typically large, this assumption is usually not difficult to meet. However, when this assumption is violated, the **log-likelihood test** (also known as **likelihood ratio test** or **G test**) or the **Fisher exact** test is more appropriate for the estimation of statistical significance (p-values) (see Upton 1992; Rayson et al. 2004; Sprent 2011).

Chi-squared is calculated according to the following equation:⁸

$$\text{Chi-squared} = \text{Sum for all cells of } \frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}} \quad (4.5)$$

We know from Chapter 3 (Section 3.2) that **expected frequencies** are frequencies that we would expect to see if there was no relationship between the variables in the data, i.e. if the null hypothesis were true. Expected frequencies function as a baseline that we use to establish if there is a real relationship between variables, in this case an effect of the explanatory variable on the linguistic variable. Expected frequencies are calculated as follows:

$$\text{Expected frequency} = \frac{\text{row total} \times \text{column total}}{\text{grand total}} \quad (4.6)$$

Let’s take as an example the use of articles in English discussed above. The data, so-called **observed frequencies**, are provided in Table 4.2. Using equation (4.6) and the data from Table 4.2 we calculate the expected frequencies. These are displayed in Table 4.5.

The chi-squared test value (statistic) is then calculated as follows:

$$\begin{aligned} \text{Chi-squared} &= \frac{(25 - 7.02)^2}{7.02} + \frac{(1 - 18.98)^2}{18.98} + \frac{(2 - 19.98)^2}{19.98} + \frac{(72 - 54.02)^2}{54.02} \\ &= 85.25 \end{aligned} \quad (4.7)$$

⁸ A technical detail: sometimes you may get a slightly different chi-squared value than that produced by equation (4.5). This is due to the fact that some packages use chi-squared with the so-called Yates’s correction. However, for corpus comparison where we deal with large amounts of data the difference is negligible in practical terms.

Table 4.5 *Expected frequencies: article type by contextual determination*

Context type \ Article type	Indefinite	Definite	Total
Non-determined	$\frac{26 \times 27}{100} = 7.02$	$\frac{26 \times 73}{100} = 18.98$	26
Determined	$\frac{74 \times 27}{100} = 19.98$	$\frac{74 \times 73}{100} = 54.02$	74
Total	27	73	100

For a 2×2 table, the 0.05 significance critical value (cut-off point) is 3.84 and the 0.01 significance critical value is 6.63;⁹ this means that any value of the chi-squared test for a 2×2 table greater than 3.84 and 6.63 is significant at the 0.05 level and the 0.01 level respectively. In our case, the p-value associated with the test value 85.25 is very small $p < 0.0000000000000001$, which is usually reported as < 0.0001 . Remember, statistical testing is not a competition for the smallest p-value (see Section 1.4). In the statistical testing procedure, we evaluate the amount of evidence which we have in the data for the rejection of the null hypothesis that states that there is no relationship between the variables in the data (the use of articles and the type of context in Table 4.2). In addition to statistical significance, we also express in standardized terms the size of the difference between the categories in the cross-tab table. This is done by reporting the effect size measure. For the chi-squared test, we have several options. The statistic that is calculated for the overall global effect in the contingency table is Cramer's V . Cramer's V ¹⁰ is a standardized chi-squared value that adjusts for the total number of observations (grand total in the cross-tab table) because the values of the chi-squared statistic grow with sample size. This growth is an intended effect of the null-hypothesis testing because the chi-squared test evaluates the amount of evidence against the null hypothesis. The effect size (Cramer's V), however, needs to be comparable across different sample sizes (corpora) and therefore uses the following standardization:

$$\text{Cramer's } V = \sqrt{\frac{\text{chi-squared}}{\text{total observations} \times (\text{no. of rows or columns, whichever is smaller} - 1)}} \quad (4.8)$$

⁹ In statistical terms, we talk about chi-squared distributions with particular degrees of freedom (df). The critical points for significance can be found in statistical tables under the appropriate df. A 2×2 table corresponds to $df = 1$, a 2×3 table to $df = 2$, a 3×3 table to $df = 4$ etc. More generally, $df = (\text{number of rows} - 1) \times (\text{number of columns} - 1)$. However, don't worry if this footnote is too technical; the online statistical tool provided for this book calculates p-values automatically so you'll never have to search for df and their corresponding p-values.

¹⁰ A similar statistic to Cramer's V is **Phi** (sometimes written as ϕ). Phi is used only for 2×2 tables and has the same values as Cramer's V . The difference in labels is historical.

Table 4.6 Interpretation of Cramer's V

Degrees of freedom (table type)	Effect size		
	Small	Medium	Large
1 (2×2)	.10	.30	.50
2 (2×3 or 3×2)	.07	.21	.35
3 (2×4 or 4×2)	.06	.17	.29

Applied to our example we get:

$$V = \sqrt{\frac{85.25}{100 \times (2 - 1)}} = 0.923 \quad (4.9)$$

The resulting value of 0.923 can be interpreted as a very large effect size. Cramer's V ranges from 0 to 1 and Table 4.6 shows Cohen's (1988) recommended interpretation of the values.

Another option, which is in many cases preferable to the general effect size such as Cramer's V , is the **probability ratio (PR, also known as relative risk or risk ratio)**. As the term suggests, it is a ratio of two probabilities from the cross-tab table comparing the probability of a particular linguistic outcome (e.g. the definite article) occurring in one context type relative to the same outcome occurring in the other context type. Because this effect size is a ratio of two values, it is suitable only for simple 2×2 cross-tab tables where there are only two categories (levels) of each variable. It is calculated as follows:

$$\text{probability ratio} = \frac{\text{probability of outcome of interest in context 1}}{\text{probability of outcome of interest in context 2}} \quad (4.10)$$

The probabilities of the definite and indefinite article in two context types are calculated in Table 4.7.¹¹ As can be seen, these are the values in the cells divided by the row totals.

Unlike in controlled experimental research such as in medical studies,¹² where the researchers are interested in one type of effect (particular outcome in a treatment group as opposed to the control group), in corpus linguistics we can calculate more than one probability ratio depending on the focus of our investigation. There are four options of probability ratios based on Table 4.7:

- PR of indefinite article in non-determined vs determined context = $\frac{0.962}{0.027} = 36$
- PR of indefinite article in determined vs non-determined context = $\frac{0.027}{0.962} = 0.03$

¹¹ Conventionally, probabilities are reported on a scale from 0 to 1. These are the same numbers as percentages based on row totals except for the multiplication by 100.

¹² A small terminological remark: the terms 'relative risk' or 'risk ratio', which are alternative labels for 'probability ratio', come from medical and epidemiological studies where the risk of a disease is calculated.

Table 4.7 *Probabilities: article type by contextual determination*

Context type \ Article type	Indefinite	Definite	Total
Non-determined	25/26 = 0.962	1/26 = 0.038	26
Determined	2/74 = 0.027	72/74 = 0.973	74
Total	27	73	100

- PR of definite article in non-determined vs determined context = $\frac{0.038}{0.973} = 0.04$
- PR of definite article in determined vs non-determined context = $\frac{0.973}{0.038} = 25.3$

For instance, we can see that the indefinite article is 36 times more likely to occur in a non-determined context than in a determined context. Conversely, the indefinite article's probability to occur in a determined context is 0.03 times its probability to occur in a non-determined context; that is a very small probability in comparison. The scale on which the probability ratio operates is 0 to infinity and the interpretation is as follows:

- A probability ratio of 1 means there is no difference between the two contexts.
- A probability ratio smaller than 1 means the linguistic outcome of interest is less likely to occur in context 1 than in context 2.
- A probability ratio larger than 1 means the linguistic outcome of interest is more likely to occur in context 1 than in context 2.

Sometimes, an alternative measure called **odds ratio** is reported instead of probability ratio. Odds ratio uses odds instead of probabilities. However, its interpretation appeals less to common sense than probability ratio because we tend to understand probabilities better than odds (Cohen 2000; Davies et al. 1998). To transform probability ratio to odds ratio we can use the following equation:

$$\text{odds ratio} = \text{probability ratio} \times \frac{\text{probability of outcome of interest in context 2}}{\text{probability of outcome of interest in context 1}} \quad (4.11)$$

For more discussion of odds ratio see Section 4.4. Finally, it should be noted that in addition to the effect size measure we should also compute the confidence intervals (95% CIs) for effect size to be able to estimate the range within which the effect is likely to occur in the population (language use in general).

Reporting Statistics: Cross-tabulation and Chi-squared

1. What to Report

In the case of simple situations with one linguistic and one explanatory variable, we report a cross-tab table with percentages as well as the chi-squared test results. For the chi-squared test the following should be reported: (i) degrees of freedom (see note 9), (ii) test value, (iii) p-value, (iv) effect size (probability ratio or Cramer's V or both) and (v) 95% confidence interval for the effect size.

In more complex situations (with more explanatory variables), the cross-tab tables alone are sufficient with detailed description of the important/interesting contrasts. If we want to report inferential statistics with complex tables, logistic regression needs to be performed (see Section 4.4).

2. How to Report: An Example

- There was a significant association between the context type and article type ($\chi^2 (1) = 85.25, p < .001$). The overall effect is large: Cramer's $V = .923$, 95% CI [.727, 1]. The definite article is 25.3 (95% CI [3.7, 172.95]) times more likely to appear in the determined context than in the non-determined context.

Tables 4.2 and 4.4 provide examples of simple and complex cross-tab tables respectively.

4.4 Logistic Regression

Think about ...

Before reading this section, review the meaning of the following terms used earlier in this chapter. Can you define them all?

CATEGORY OF A VARIABLE, EXPLANATORY VARIABLE, LEVEL OF A VARIABLE, LINGUISTIC VARIABLE, OUTCOME, OUTCOME VARIABLE, PREDICTOR, PREDICTOR VARIABLE, VALUE OF A CATEGORICAL VARIABLE, VARIANT (OF A LINGUISTIC VARIABLE)

In this section, we will discuss a powerful statistical technique called **logistic regression**. So far, we have explored the nature of lexico-grammatical variables (Section 4.2) as well as some relatively simple techniques for dealing with lexico-grammatical variation (Section 4.3). In the following discussion, we'll look at the analysis of lexico-grammatical variation with logistic regression, a technique that uses explanatory (sometimes also called predictor) variables, which can be both categorical and scale, to estimate their effect on the linguistic (outcome) variable,

which has to be a categorical variable.¹³ This means that we are looking for any contextual features (predictors) that have an effect on the occurrence of different variants of a linguistic variable such as the definite or indefinite article. When analysing the data, we build a mathematical model (an equation) that represents complex relationships between the variables in the data. This model is then used to explain the effect of the predictor variables (and their combinations) on the outcome variable. Because logistic regression is a fairly complex statistical technique, this section introduces a number of new terms and technical details. However, don't worry if you don't grasp every detail of this procedure. What is important to know are the basic principles of this technique and the interpretation of the output; the computation of logistic regression is carried out automatically and can be done using the Logistic regression tool from Lancaster Stats Tools online.

For those interested in the details, the following is one possible form of the logistic regression equation:

$$\text{probability of outcome of interest} = \frac{e^{(\text{intercept} + b_1X_1 + b_2X_2\dots + \text{unaccounted variation})}}{1 + e^{(\text{intercept} + b_1X_1 + b_2X_2\dots + \text{unaccounted variation})}} \quad (4.12)$$

Explaining this equation introduces some new terminology: e is a mathematical constant that is approximately equal to 2.71828,¹⁴ the intercept is a baseline value, b_1 , b_2 etc. are so-called estimates or coefficients and X_1 , X_2 etc. are predictors used in the model. All of these terms will be explained in this section. For now, in very simple terms, we can say that logistic regression estimates the probability of a particular outcome of interest based on individual predictors, which contribute to this outcome to a different degree (some are more relevant than others). In essence, it is a complex classification exercise in which we combine relevant predictors (pointing to patterns, regularities or ‘rules’ of lexico-grammar) that can predict to which category a particular instance of a linguistic feature belongs. The question we are asking is this: does this context favour the use of variant A (e.g. definite article) or variant B (e.g. indefinite article)? This idea is visually captured in Figure 4.1.

Before proceeding to the explanation of the details of the logistic regression technique, let us review the basic terminology. In the ‘Think about’ task, you reviewed terms that are used synonymously in the context of logistic regression applied to the study of lexico-grammar; these terms can be divided into four groups:

¹³ If the outcome variable is a scale variable, ordinary least squares (OLS) regression needs to be used. See Gries (2013a: 261–82) for a detailed explanation of this procedure.

¹⁴ e is an irrational number (that's why it cannot be written precisely as fraction or a finite series of numbers after the decimal point). You may also know e as the base of the natural logarithm (\ln).

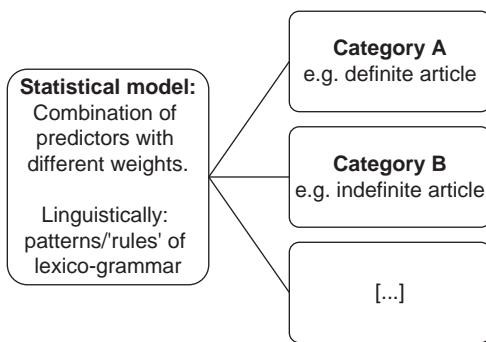


Figure 4.4 *Logistic regression: a basic schema*

- Features of lexico-grammar that are the focus of the research: ‘linguistic variable’ = ‘outcome variable’.
- Terms related specifically to the structure of the outcome variables: ‘variant (of a linguistic variable)’ = ‘outcome’.
- Contextual variables that help us explain the use of lexico-grammatical features: ‘explanatory variable’ = ‘predictor variable’ = ‘predictor’.
- Terms related to the structure of categorical variables that are used both as outcome variables and predictor variables:¹⁵ ‘category of a variable’ = ‘level of a variable’ = ‘value of a categorical variable’.

There are several stages of the logistic regression procedure: (i) Data checking (prerequisites and assumptions), (ii) Building a model and (iii) Interpreting the model. To illustrate the logistic regression technique, we’ll use the example of the definite article discussed previously (see e.g. Figure 4.2).

STAGE 1: Data checking (prerequisites and assumptions)

Before analysing the data (which we call building a model), we need to check that the dataset is suitable for this type of analysis and that the assumptions of the statistical test are met. First, we need to make sure that the dataset is organised according to the principles of the **Linguistic feature research design** (see Section 1.4). This means that each occurrence of the linguistic feature of interest such as the definite and the indefinite article is on a separate line and is properly annotated for explanatory variables as shown in Figure 4.2. We also need to check that the linguistic feature lends itself to this type of research and that the **lexico-grammatical frame** has been properly defined (see Section 4.2).

¹⁵ NB: Predictors can also be scale variables.

Second, as in any type of quantitative analysis, the variables should be measured and coded accurately and consistently. Multivariate analyses (such as logistic regression) are especially sensitive to measurement errors, which can have a multiplicative effect due to multiple variables used in model building. Because many, especially functional, aspects of lexico-grammatical variables (e.g. syntactic/semantic function of a linguistic feature) are coded manually, the consistency and accuracy of coding should not be underestimated. If a **judgement variable** is involved, double coding of a certain portion of the data (e.g. randomly selected 20%) is recommended (see Section 3.5). For more discussion on data preparation and management see Osborne (2012: 195ff).

Third, we need to have enough data. As a general principle, the more explanatory variables we use, the more data (cases or lines in the dataset) we need to have. Hosmer et al. (2013: 407–8) discuss a ‘rule of 10’, which states that the maximum number of coefficients that can be included in a model (b_1 , b_2 etc. in equation (4.12)) is estimated by taking the frequency of the least frequent outcome and dividing it by 10. The best way to count the number of coefficients is to look at the logistic regression output for coefficients (see the discussion below) and count the number of rows in the table leaving out the intercept. For example, using the dataset presented in Table 4.2, with the least frequent outcome being the indefinite article (27 cases), we could include a maximum of two coefficients (27/10) in the model. Hosmer et al. (2013: 408), however, consider also the limitations of this rule of thumb, pointing out that there are many different factors that play a role in the issue of data sufficiency in logistic regression. In corpora, we usually have enough data for this type of investigation, although the coding of cases typically involves a certain amount of manual labour and can potentially be time-consuming. It is also advisable to select predictor variables wisely (ideally based on previous literature and linguistic theory) rather than throw in as many explanatory variables as possible.

Fourth, we need to check whether it makes sense to build a model. If one predictor perfectly explains (classifies) all cases, then building a model that estimates the effect of different predictors is redundant. There are also mathematical reasons for why such a model cannot be constructed (Hosmer et al. 2013: 147–8). In statistical terminology, this issue is called **(quasi) complete separation**. For example, the type of noun (singular countable, plural countable, uncountable, proper) predicts very accurately the occurrence of the indefinite article: *A/an* can appear only in front of a countable noun in singular. Table 4.8 demonstrates this situation with our example dataset.

Fifth, we need to check the assumptions of the test itself (see Osborne 2015: 85–130). There are three main assumptions: (i) independence of observations, (ii) no (multi)collinearity and (iii) linearity. **(i) Independence**

Table 4.8 *Probabilities: article type by noun type*

Noun type \ Article type	Indefinite	Definite
Noun type		
Countable sg.	27	34
Countable pl.	0	19
Uncountable	0	10
Proper	0	10

of observations. As with the chi-squared test discussed in Section 4.3, we assume that every observation such as the use of an article is independent of another observation. In corpora, this assumption is usually violated to some extent due to the nature of language, where linguistic features are interconnected, and also due to corpus sampling that is done at the level of texts, not individual linguistic features. Because when studying lexico-grammatical variation, we usually presuppose a certain uniformity among texts of the same type, we don't have to worry about this problem too much apart from making sure that in our sample, linguistic features come from a range of texts. This can be done by taking a random subsample of linguistic features from the corpus. For studies which need to control for individual texts or speakers a method called **mixed-effect modelling** is used (see Section 6.5 for a detailed discussion of this). **(ii) No (multi)collinearity between predictors.** Collinearity is characterised by high correlation ($r = 0.8$ and above) between predictor variables (see Section 5.2 for an explanation of correlation). This needs to be checked for both scale and categorical variables.¹⁶ Collinearity actually means that the correlated variables measure a similar thing (construct, to use a technical term) and therefore we don't need both/all of them. If we encounter (multi)collinearity, we can either exclude or combine variables to avoid this problem. **(iii) Linearity.** Logistic regression belongs to a group of statistical methods that use **linear modelling**. This means that the relationship between variables is captured by a (regression) line (see Section 1.2). However, not all relationships between variables are linear. For example, a language change over time is usually **curvilinear** – it can be expressed as a curve rather than a line (see Section 7.5 for more discussion of non-linear models). In logistic regression, we assume that there is a linear relationship between scale predictors (if we have any) and the log odds of the outcome variable (see below for the explanation of log odds). For more details about testing this assumption see Hosmer et al. (2013: 94ff) and Field et al. (2012: 344–5).

¹⁶ To run a correlation test with categorical variables, replace the categories' labels with numbers, e.g. non-determined = 0; determined = 1.

STAGE 2: Building a model

Having checked the data and the assumptions, we can focus on data analysis (building the model). The first step here is to decide the baseline values of categorical variables including the outcome variable; the baseline value for scale variables is always 0, so we do not need to specify this. The **baseline values**, sometimes called the **reference levels**, are values against which the model compares the effects of the predictor(s) on the outcome. Simply put, these are the values that are not the focus of the investigation. For example, if our research question is ‘in what contexts is the definite article used in English?’ the definite article will be our **outcome of interest**. We therefore set the indefinite article as the baseline value. Similarly, if we assume (based on the literature and linguistic theory) that determined contexts favour definite articles, we’ll select the non-determined contexts as the baseline value of the predictor variable.

Practical advice: it is useful to distinguish between baseline values and other values of categorical variables using a prefix letter (A_, B_, C_). A baseline value is coded with prefix A_ (e.g. A_indefinite), while other values are coded with other prefix letters (e.g. B_determined). This is done because software packages (including the Logistic regression tool in Lancaster Stats Tools online) enter the data in the model alphabetically by default. An example of a dataset that is ready to be analysed using the logistic regression technique is displayed in Figure 4.5. In this dataset, we have three predictors ('Context_type', 'Noun_type', 'NP_Length') and one outcome variable ('Article_type'). With the exception of 'NP_Length', which measures the length of the noun phrase (number of characters) at the scale level, all other variables are categorical (nominal) and therefore have a prefix letter.

ID	Context_type	Noun_type	NP_Length	Article_type
1	A_nondetermined	D_proper	29	B_definite
2	B_determined	B_count_pl	15	B_definite
3	B_determined	C_uncount	7	B_definite
4	B_determined	B_count_pl	7	B_definite
5	B_determined	C_uncount	27	B_definite
6	B_determined	A_count_sg	14	B_definite
7	A_nondetermined	A_count_sg	4	A_indefinite
8	B_determined	D_proper	17	B_definite
9	A_nondetermined	A_count_sg	51	A_indefinite
10	B_determined	A_count_sg	26	B_definite
11	B_determined	A_count_sg	3	B_definite
12	B_determined	A_count_sg	4	B_definite
13	B_determined	A_count_sg	6	B_definite

Figure 4.5 Article use in English: a dataset (an excerpt)

The next step is to decide how and what variables to enter in the model. The goal is to find a model that explains as much variation in the data as possible with as few variables as possible. This is sometimes called a **parsimonious model**. We therefore want to have only the most relevant variables in the model. This can be achieved in two different ways. We can decide a priori based on previous literature and linguistic theory which variables to enter, run the analysis and then leave in only variables that have a statistically significant effect. This is so-called **block entry**. Another option is a **stepwise entry** where we let the statistical software add or remove variables step-by-step until the information criterion called AIC (see below) of the model is no further improved. Within the stepwise procedure, we can decide between a **forward**, a **backward** and a **hybrid** procedure. The forward procedure starts with no predictor variables in the model and keeps adding available variables one-by-one. The backward procedure starts with a model which includes all available variables and deletes them one-by-one. The hybrid procedure combines the forward and the backward technique adding and deleting predictors depending on reassessment of the model at each stage. The advantages of the block entry include the researcher's full control over the process with the important decision about which variables to include being based on theoretical grounds (variables relevant from the perspective of linguistic theory) rather than an automated, theory-blind procedure carried out by the statistical software. The stepwise procedure is justified in exploratory analyses where no clear theoretical grounds for variable inclusion are available (Hosmer et al. 2013: 93–4; Osborne 2015: 251–3). In lexico-grammatical corpus research, block entry is usually preferable.

With the dataset in Figure 4.5 (use of the definite/indefinite article), we have several options for building a model. The most interesting ones are listed in Table 4.9. The first model (md0) is a so-called **baseline (or null) model**, which includes no predictors and is used only as a reference point for more complex models; this model, which doesn't include any information about the contexts in which the linguistic variables occur, can be improved by adding relevant predictors. Md1 includes only one predictor (Context_type), while the other two models (md2 and md3) incorporate two predictors (Context_type plus one other predictor). So which model to choose?

By default, statistical packages provide a comparison of any model we build with the baseline model. If our model doesn't perform better than the baseline model, it is pretty useless because the selected predictors have little effect and we can therefore discard the model. In Table 4.9, we can see that both md1 and md3 are significantly better than the baseline model and we can therefore consider them further. Statistical significance is established by means of the **log-likelihood test** (also known as **likelihood ratio test**), a well-known measure in corpus linguistics (see Sections 4.3 and 3.4). Md2, on the other

Table 4.9 *Models: an overview*

Model label	Outcome	Predictors included	Result based on output of statistical software (see ‘STAGE 3: Interpreting the model’ below)
md0	Article_type	[none]	Baseline model with intercept only.
md1	Article_type	Context_type	Statistically significant, i.e. significantly better than a baseline model, AIC = 30.86
md2	Article_type	Context_type, Noun_type	Large standard errors → something went wrong; Noun_type causes a problem with complete separation.
md3	Article_type	Context_type, NP_Length	Statistically significant, i.e. significantly better than a baseline model; however, not significantly better than md1; AIC = 31.91 (larger than AIC for md1)

Table 4.10 *A part of the logistic regression output: large standard errors*

	Estimate (log odds)	Standard error
(Intercept)	-21.056	4530.376
Context_typeB_determined	23.889	4530.376
Noun_typeB_count_pl	18.733	6706.381
Noun_typeC_uncount	18.733	9244.108
Noun_typeD_proper	39.961	8958.692

hand, shows a typical warning sign of a regression model – extremely large **standard errors**. Table 4.10 shows a part of the logistic regression output with standard errors highlighted. As we can see, the standard errors are many times larger than the estimates; this is a red flag that tells us that something went wrong with the model. The meaning of both the standard errors and estimates (coefficients) will be explained below when we discuss the interpretation of the model. As it turns out, with md2 the problem is the complete separation issue connected to the Noun_type predictor (see above); md2 can therefore be discarded.

With md2 out of the way, let’s focus on the comparison of md1 and md3. At first it might seem that a model with more predictors is a better one. This, however, is not true. In the same way as we compared our models with the baseline model, we can compare any two models (md1 and md3 in our case) to see if one is significantly better than the other. In addition to statistical significance, which is measured by the **log-likelihood test**, we also use **AIC (Akaike information criterion)** to establish which model is the most efficient by reaching significance with as few variables as possible. AIC is calculated as follows:

$$\text{AIC} = \text{unexplained variation by the model} + 2 \times (\text{number of outcome categories} - 1 + \text{number of predictors}) \quad (4.13)$$

AIC has the following interpretation: when comparing two models based on the same dataset, the smaller the AIC, the better the model. The logic of the measure is also fairly straightforward: we take the variation in the data that the model doesn't explain and in addition penalize the model for using many predictors (the more predictors the more we penalize the model). As mentioned earlier, a successful model is a parsimonious model that explains as much variation as possible with as few predictors as possible. Note that AIC is not a standardized effect size measure and therefore cannot be used to compare models based on different datasets. As we can see from Table 4.9, md3 is not significantly better than md1 and, moreover, md1 has a smaller AIC value (30.86) than md3 (31.91); we'll therefore opt for the model with only one variable (md1).

Finally, at this stage we need to decide if we want to include only the selected predictors' so-called **main effects** or also **interactions** between the predictors. Interactions are specific predictor combinations (combinations of contexts in which the linguistic variable occurs) which can, in some cases, have a significant effect on the outcome, despite the fact that the individual predictors in isolation do not. If interactions are not tested, this effect may not be discovered. However, as with variable selection, the inclusion of interactions should be based on previous literature and linguistic theory.

STAGE 3: Interpreting the model

Logistic regression produces an output that consists of two main parts: (i) Model summary and (ii) Coefficients (estimates). As an example, a basic output for md1, our preferred model, is reproduced below. Note that this particular form of output comes from the logistic regression tool from Lancaster Stats Tools online and may differ to some extent from outputs of other statistical packages.

Model md1

Model summary: Likelihood ratio test (LL): 89.79 ($p < .0001$) \rightarrow SIGNIFICANT; C-index: 0.96 \rightarrow OUTSTANDING; Nagelkerke R^2 : 0.86; AIC: 30.87

Coefficients:

	Estimate (log odds)	Standard error	Z value (Wald)	p-value	Estimate (odds)	95% CI lower	95% CI upper
(Intercept)	-3.219	1.020	-3.156	0.002	0.04	0.002	0.189
Context_type	6.802	1.247	5.457	0.000	900	116.878	21421.229
B_determined							

Looking at the model summary, we can see that overall the model is statistically significant with a likelihood ratio test value of 89.79¹⁷ and a very low p-value ($p < .0001$); this means that the model md1 with one predictor (Context_type) is significantly better than a baseline model with no predictors. Besides, many additional measures such as *C*-index and different versions of pseudo- R^2 can be used to further evaluate the model's performance. The *C*-index (concordance index¹⁸) is a measure of the classification success of the model, in other words, how well the model can predict the outcome (*the* or *a(n)* in our example). We are looking for a *C*-index value above 0.7¹⁹ (Hosmer et al. 2013: 177).²⁰ Sometimes, pseudo- R^2 values such as Nagelkerke R^2 are also reported. These measures try to capture the amount of variation in the data explained by the model. The range of pseudo- R^2 is from 0 (no variation explained) to 1 (all variation explained). Unfortunately, pseudo- R^2 measures are generally not very reliable (Osborne 2015: 51).

Let us now focus on the effect of individual predictors by looking at the second part of the output called 'Coefficients'. The coefficients (estimates) are displayed in a table. The first row always displays a so-called **intercept** (or **constant**), a baseline value in the model estimating the situation where all predictors are at their baseline values. Remember that for categorical predictors such as the Context_type these are the values that we set as baseline; for scale predictors this is always 0 (see 'STAGE 2: Building a model' above). In md1, the intercept estimates the odds of the definite article occurring in the baseline (that is non-determined) context. The odds are very small: 0.04 (with 95% CI 0.002, 0.189).

Let us pause for a moment and explain the units which are used to measure the effect of the predictors. The units are **odds** and **log odds (logits)**, the latter being computed by taking the natural logarithm of odds. Internally, the logistic regression, as the name suggests, operates with log odds. Because these are relatively difficult to interpret, we usually convert log odds to simple odds. Odds are defined as follows:

$$\begin{aligned} \text{odds} &= \frac{\text{probability of outcome } o}{\text{probability of not outcome } o} \\ &= \frac{\text{probability of outcome of interest in baseline context}}{\text{probability of baseline outcome in baseline context}} \end{aligned} \quad (4.14)$$

¹⁷ Note that the same value of 89.79 would be obtained by performing the log-likelihood test based on Table 4.2 because the simplest case of logistic regression with one predictor is identical to performing the log-likelihood test. Note also that a very similar value of 85.25 was obtained in Section 4.3 when performing the chi-squared test on Table 4.2.

¹⁸ Other labels of the same measure include index of accuracy, Gini-index and AUC.

¹⁹ If the index is above 0.8, the model has an excellent classification success; above 0.9 (which is rare), the classification success is outstanding.

²⁰ Although the *C*-index has been widely used, it was recently criticized and another measure (H-measure) was suggested instead (Hand 2010). The debate is fairly technical; I still recommend reporting the *C*-index with the caveat that no single measure can capture the model's predictive success in all circumstances.

In practice, odds are often used in sports betting. There we ask questions such as what are the odds of our team winning the game? If the odds are 2 to 1 the probability of our team winning is twice as large (i.e. 66.7%) as the probability of our team losing (33.3%). In our example, the odds of the baseline value (intercept), i.e. a definite article occurring in the non-determined context, are calculated as follows (relevant probability values are available in Table 4.7):

$$\text{odds (intercept)} = \frac{\text{probability of } \textit{the} \text{ in non-determined context}}{\text{probability of } a(n) \text{ in non-determined context}} = \frac{0.038}{0.962} = 0.04 \quad (4.15)$$

Note that this value corresponds with the odds value of the intercept in the output table, column 6 ‘Estimate (odds)’. Returning to the table of coefficients in the output, the remaining rows, which we usually focus on because they answer the research question about the effect of different predictors, list scale predictor variables and so-called **dummy variables** (combinations of categorical predictors and their non-baseline values). Thus, the row ‘Context_typeB_determined’ shows the effect of the determined context on the use of the definite article. This effect is measured in **log odds ratios** (useful for internal operation of logistic regression) and **odds ratios** (useful for the interpretation of the effect) by making a comparison with the baseline value. The odds ratio is calculated as follows:

$$\begin{aligned} \text{odds ratio} &= \frac{\text{probability of outcome of interest in context of interest}}{\text{probability of baseline outcome in context of interest}} \\ &\quad \times \frac{\text{probability of outcome of interest in baseline context}}{\text{probability of baseline outcome in baseline context}} \\ &= \frac{\text{odds (predictor value of interest)}}{\text{odds (intercept)}} \end{aligned} \quad (4.16)$$

An odds ratio larger than 1 indicates that the odds of the outcome of interest occurring in the context of interest are larger than those of the same outcome occurring in the baseline context; an odds ratio smaller than 1, on the other hand, shows the dispreference of the context of interest for the outcome of interest. If the odds ratio is 0, there is no effect.

In our example, the odds ratio that shows the effect of the determined context type on the use of the definite article is calculated as follows (relevant probability values are available in Table 4.7):

$$\begin{aligned} \text{odds ratio (Context typeB}_{\text{determined}}\text{)} &= \frac{\text{probability of } \textit{the} \text{ in determined context}}{\text{probability of } a(n) \text{ in determined context}} \\ &\quad \times \frac{\text{probability of } \textit{the} \text{ in non-determined context}}{\text{probability of } a(n) \text{ in non-determined context}} \\ &= \frac{\frac{0.973}{0.027}}{\frac{0.038}{0.962}} = \frac{36}{0.04} = 900 \end{aligned} \quad (4.17)$$

We can see that this value is provided in the output table row 2 ‘Context_typeB_determined’, column 6 ‘Estimate (odds)’. It can thus be said that the odds of the definite article occurring in the determined context are 900 times (with 95% CI [117, 21,421]) larger than the odds of the definite article occurring in the non-determined context.

For each row, the output table also lists **standard errors**, the significance statistic **Wald’s z** and the corresponding **p-value**, which tells us if the specific estimate is statistically significant. Standard errors express how accurately the estimates reflect the value in the population; the smaller standard errors are the better. The **Wald statistic (z)** is computed by dividing the estimate with its standard error. The effect size for each parameter is the odds ratio discussed above, which is supplemented with 95% confidence intervals, showing us where the odds ratio is likely to lie in the population.

In more complex models (those with multiple predictors), we evaluate the estimates one by one. The following table shows the addition to the coefficients output when we include the scale NP_length predictor in the model measuring the length of the noun phrase.

	Estimate (log odds) error	Standard (Wald)	Z value p-value	Estimate (odds)	95% CI_l	95% CI_u
NP_Length	0.037	0.039	0.939 0.348	1.037	0.966	1.138

We can see that NP_length is not a significant predictor ($p > 0.05$) and the odds ratio is very small. 1.037 is close to 1, which means no effect. This observation is confirmed by the confidence interval which actually includes 1; this is a sign of a statistically non-significant result, because in the population the effect can well be null (odds ratio 1). The reason for showing this output is to discuss the meaning of the odds ratio in the case of a scale predictor. The odds ratio of a scale predictor indicates how many times larger (if > 1) or smaller (if < 1) the odds of the outcome of interest are compared to the odds of the baseline outcome with one unit change of the scale predictor. In our example, the unit of noun phrase length was one character. The estimate (1.037) therefore indicates that the odds of the definite article *the* increase 1.037 times with every character added to the noun phrase. This means the longer the noun phrase is the more likely we are to see the definite article. However, as noted, this effect is not statistically significant and we would normally exclude it from the model because there is not enough evidence for it in the data.

In sum, logistic regression is a powerful method that gives us a detailed insight into the effects of different contexts on the linguistic output. It needs to be noted that in this section we have explored a type of logistic regression called **binomial logistic regression**, that is a logistic regression with an outcome variable with two categories like *the* and *a/an*. This represents a typical case in lexico-grammatical research that investigates a competition between two

linguistic features within a given lexico-grammatical frame. If, however, the outcome variable has more than two options (categories), a similar technique called **multinomial logistic regression** is used. Multinomial regression follows the same principles as binomial regression, but the comparisons are somewhat more complex (see Field et al. 2012: 8.9.1–8.9.2 and Arppe 2008 for an example of such a procedure).

Reporting Statistics: Logistic Regression

1. What to Report

Because logistic regression is a complex procedure, its successful use depends on a number of steps, as discussed in this chapter. These steps should be briefly outlined in our research report to allow replicability. First, we should indicate how the lexis-grammatical frame was defined and which variables were used and why. Second, we should let the readers know how the data was obtained (e.g. a random subsample of all occurrences of the linguistic features from a corpus) and coded as well as whether any part of the data was double coded (if so, inter-rater agreement statistic needs to be reported). Third, we need to provide details about the overall statistics of the model (LL, p-value, C-index) as well as a table of individual coefficients, including statistical significance information, the odds ratios and 95% confidence interval for the odds ratios.

2. How to Report: An Example

- Because the focus of the research was the variation between the definite and the indefinite article, all occurrences of these linguistic features were found in the corpus. One hundred cases were then randomly selected and coded for the presence or absence of the definite article (outcome variable). In addition, two contextual variables (predictors) reported in the literature as having an effect on the use of articles in English were also measured. These were the context type and length of the noun phrase.
- The context type was a significant predictor of the type of article used. Entry of determined context into the model significantly improved model fit ($LL = 89.79$, $p < .0001$). The model also has outstanding classification properties ($C\text{-index} = .95$). The length of the noun phrase, on the other hand, didn't have a significant effect. As can be seen from the coefficients table below, the use of the definite article is much more likely in the determined context ($OR = 900$, 95% CI [117, 21,421]) than in the non-determined context.

	Estimate (log odds)	Standard error	Z value (Wald)	p-value	Estimate (odds)	95% CI lower	95% CI upper
(Intercept)	-3.219	1.020	-3.156	0.002	0.04	0.002	0.189
Context_type	6.802	1.247	5.457	0.000	900	116.878	21,421.229
B_determined							

4.5 Application: *That* or *Which*?

Lexico-grammatical variables as discussed in this chapter are *variables* which can be expressed as choices between two or more variants, such as the definite and the indefinite article.

Figure 4.6 *A sentence from this book corrected for ‘grammar’*

While writing this chapter, I encountered the following situation: my word processor underlined with a wiggly line a phrase that included a relative pronoun *which*, signalling a potential grammatical error or inaccuracy (see Figure 4.6). The correction offered was to either add a comma in front of *which* or use the relative pronoun *that* instead, with the reasoning being as follows: ‘If these words are not essential to the meaning of your sentence, use “which” and separate the words with a comma’ (Microsoft 2010).

Having overcome the initial resentment at the fact that a computer was correcting my grammar, I started thinking about how to test whether the rule the grammar checker was applying actually reflects how language is used. I devised a study using BE06 and AmE06, two one-million-word corpora of current written English. The study is reported below.

This study is based on BE06, a balanced corpus of contemporary British English, and AmE06, a balanced corpus of contemporary American English. Each corpus consists of four major written genres (general prose, fiction, newspapers and academic writing).

In BE06 and AmE06, there are 4,736 instances of *which* and 22,749 instances of *that*. However, while *which* is used predominantly as a relativizer in contexts such as the one in Figure 4.6, *that* has many other functions as well (determiner, demonstrative, emphaser etc.). In this study, the relevant lexico-grammatical uses of *which* and *that* (and their possible alternation) were defined using the appropriate part-of-speech tags (DDQ for *which* and CST for *that* from the CLAWS7 tagset²¹) in combination with the syntactic position after a noun. Only instances where *which* and *that* are in principle interchangeable were considered.²² This established the lexico-grammatical frame for this study.

The suggestion from the word processor (‘If these words are not essential to the meaning of your sentence, use “which” and separate the words with a comma’) has two aspects: a formal and a functional aspect. The formal aspect requires *which* to be separated by a comma, while the functional aspect requires *which* to be used in situations where the clause introduced by *which* is not

²¹ Part-of-speech tags are grammatical labels automatically attached to words using a part-of-speech tagger; for more information see <http://ucrel.lancs.ac.uk/claws7tags.html>

²² The cases were excluded where *that* follows a noun denoting a person or group of people (*For a guy that did well he dressed down* (BE06_K16)) and therefore alternates with *who* rather than *which*.

Table 4.11 Cross-tabulation: separator use with which and that relativizers

Relativizer \ Presence of separator	separator (, or –)	no separator	Total
which	1,396 (63%)	804 (37%)	2,200
that	191 (3%)	7,281 (97%)	7,472
Total	1,587	8,085	9,672

essential to the overall meaning of the sentence (can be left out), something that is in grammatical terminology called a ‘non-restrictive’ (or ‘non-essential’) clause. By implication, *that*, the other relativizer, is to be used in complementary distribution to *which*, that is in ‘restrictive’ (‘essential’) modifying clauses, not separated by a comma. To test both the formal and the functional aspect of the suggestion, two research questions were formulated:

- RQ1 (formal): Is *which* preceded by a comma or a dash (–) while *that* appears without a comma or a dash (–)?
- RQ2 (functional): What are the factors that affect the use of *which* and *that*?

The results of the search of BE06 and AmE06 that answer the first question are displayed in Table 4.11. Note that the table displays the results of automatic corpus searches without manual checking (see RQ2).

Overall, *that* (7,472) is used much more often than *which* (2,200) in clauses modifying a noun. Almost two-thirds of the clauses with *which* are separated by a comma or dash, while *that* is preceded by a comma or dash only in 3 per cent of cases. When looking closely at this 3 per cent, we notice that a large majority of these cases include a parenthetical phrase before the *that* clause, e.g. *The communications circuit proposed by Robert Darnton rightly identifies many factors, besides authorship, that govern any reading experience* (BE06_J02). A parenthetical phrase is a phrase that interrupts the main flow of the sentence and is typically separated by commas from both sides. The separator that we observe thus belongs to the parenthetical phrase, not the relative clause itself.

The chi-squared test confirmed that there is a significant association between the relativizer (*which* or *that*) and the presence of a comma or dash ($\chi^2 (1) = 4,595.47$, $p < .001$). The overall effect is large: Cramer’s $V = 0.689$, 95% CI [.669, .709]. A comma or dash is 24.8 (95% CI [21.5, 28.7]) times more likely to appear in front of *which* than in front of *that*. Conversely, no separator is 2.7 (95% CI [2.5, 2.8]) times more likely to appear in front of *that* than in front of *which*. This complex relationship is visualized in Figure 4.7.

In sum, we can see a clear preference for no separator occurring before *that*; a separator may or may not appear before *which*, although there is a preference for a comma or dash to appear. However, it cannot be stated as a categorical rule that *which* should be always separated with a comma because as we can see from Table 4.11, there are numerous (804) counter-examples.

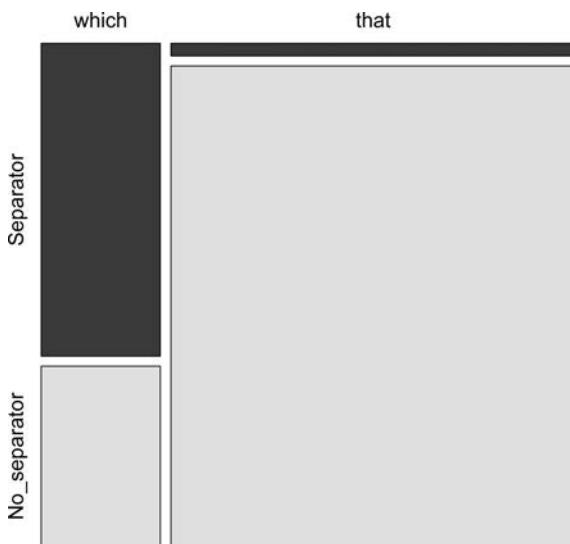


Figure 4.7 *Visualization of the relationship between which and that and a separator*

To answer the second research question, a subset of 360 random concordance lines taken from 9,672 instances of *which* and *that* listed in Table 4.11 was manually coded for a number of contextual variables such as the variety of English, presence of a separator, clause type and syntactic type. In addition, the length of the relative sentence that followed *which* or *that* was measured and recorded as the ‘Length’ variable. For example, the following sentence would be coded as Variety: British, Separator: NO, Clause type: Restrictive, Syntax: Subject,²³ Length: 5.

Apple will just veto and refuse to distribute any application which does not meet its terms. (BE06_E33)

Table 4.12 lists all categorical outcome variables included and the frequencies of *which* and *that* in the contexts defined by these outcome variables. The ‘Length’ variable, which is measured as a scale level, cannot be cross-tabulated.

From Table 4.12, we can see that the picture is quite complex and it is not easy to make sense of all the factors by simply looking at the table. For this reason, logistic regression analysis was used to identify the relevant factors that play a role in the selection of *which* or *that*.

Overall, the model that includes all the predictor variables (‘Variety’, ‘Separator’, ‘Clause type’, ‘Syntax’ and ‘Length’) is significant (LL: 222.31;

²³ *Which* is here the subject of the relative clause. Cf. *Still, it was his only answer, which he repeatedly struck* (AmE06_K15), where *which* is the object.

Table 4.12 *Which and that in different contextual situations: cross-tabulation*

Variety	Separator (, or –)	Clause	Syntax	Relativizer		
				<i>that</i>	<i>which</i>	Total
American	NO	Non-restrictive	Object	3	1	4
			Subject	11	2	13
		Restrictive	Object	18	2	20
			Subject	126	5	131
	YES	Non-restrictive	Object	0	6	6
			Subject	0	20	20
		Restrictive	Object	0	0	0
			Subject	1	0	1
British	NO	Non-restrictive	Object	3	2	5
			Subject	3	8	11
		Restrictive	Object	14	8	22
			Subject	76	15	91
	YES	Non-restrictive	Object	0	4	4
			Subject	0	31	31
		Restrictive	Object	1	0	1
			Subject	0	0	0
Total				256	104	360

$p < .0001$) and has outstanding classification properties (C -index: 0.91). Table 4.13 displays individual coefficients in the model. We can see that with the exception of ‘SyntaxB_Subject’, all estimates are significant ($p < 0.05$). Thus British variety, presence of a separator, occurrence in a non-restrictive clause and greater clause length favour the use of *which*.²⁴ For example, the odds of *which* being used in the British English are 5.3 (95% CI [2.5, 12.1]) times the odds of the same relativizer being used in American English. Similarly, the odds of *which* used after a separator are 53.8 times (95% CI [12.9, 376.4]) the odds of *which* used without a separator. Also, the longer the sentence the more likely it is to use *which*. The odds increase 1.1 times (95% CI [1.02, 1.15]) with each additional word.

So much for the study. The burning question, however, remains: was the computer right after all? If the suggestion by the computer were to be taken as a categorical rule, the answer is certainly ‘no’. The study demonstrated that there is a combination of multiple factors that favour or disfavour the use of *which* (and *that*) and these factors have to be interpreted as probabilities (or odds, to be precise), not certainty.

²⁴ Note that the relativizer *that*, American variety, no separator, restrictive clause and syntactic role as an object were set as the baseline values of the respective variables.

Table 4.13 Which *or* that: logistic regression estimates

	Estimate (log odds)	Standard error	Z value (Wald)	p-value	Estimate (odds)	95% CI lower	95% CI upper
(Intercept)	-3.354	0.563	-5.958	0.000	0.035	0.011	0.099
VarietyB_BR	1.667	0.397	4.195	0.000	5.296	2.511	12.080
SeparatorB_YES	3.985	0.825	4.832	0.000	53.795	12.876	376.448
ClauseB_Non_restr	2.046	0.446	4.588	0.000	7.733	3.235	18.812
SyntaxB_Subject	-0.614	0.421	-1.460	0.144	0.541	0.240	1.260
Length	0.079	0.029	2.739	0.006	1.083	1.023	1.147

4.6 Exercises

1. Look at the topics and research questions in Table 4.14. Decide if the linguistic feature research design is appropriate (YES/NO). If yes, define the appropriate lexico-grammatical frame. The first row was completed for you as an example.

Table 4.14 Appropriate research design

TOPIC: Research question	Linguistic features (examples)	Linguistic feature research design?	Lexico-gramma- tical frame
DATIVE ALTERNATION: What factors have an effect on dative alternation in English?	<i>She handed the student the book.</i> <i>She handed the book to the student.</i>	YES	All dative constructions.
A/AN ALTERNATION: When is a non-standard version of the indefinite article (<i>a</i> before a word beginning with a vowel) used in spoken English?	<i>an apple, a apple</i>		
SWEARWORDS: Do speakers use more strong or weak swearwords?	<i>fuck, cunt, motherfucker etc.</i> vs <i>damn, crap, hell etc.</i>		
GENITIVE ALTERNATION: What factors influence the choice between <i>s-</i> and <i>of-</i> genitive?	<i>president's speech, the speech of the president</i>		
EPISTEMIC MARKERS: Does corpus data support the hypothesis that 'we only say we are certain when we are not' (Halliday)?	<i>This is certainly the case.</i> <i>This is maybe the case.</i>		
ATTENDED/UNATTENDED THIS: What factors influence the presence of a noun after <i>this</i> ?	<i>This is an example.</i> <i>This sentence is an example.</i>		

2. Analyse the following cross-tab table: add row and column totals as well as the grand total. Is there a difference in the use of *must*, *have to* and *need to* between British and

American English? Add the percentages that will help you answer this question and calculate the chi-squared test (with raw frequencies).

Variety	Modal			Total
	<i>must</i>	<i>have to</i>	<i>need to</i>	
American	352	355	201	
British	448	405	190	
Total				

3. Interpret the following mosaic plot in Figure 4.8. It displays the use of three modal expressions of strong obligation (*must*, *have to* and *need to*) in BE06, a one-million-word corpus of written British English. The genres displayed are academic writing ('Acad'), fiction ('Fiction'), general prose ('General') and newspapers ('Press').

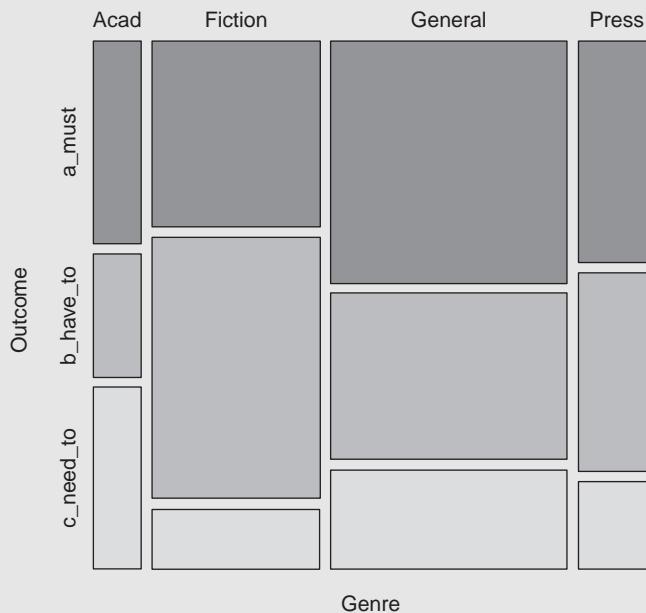


Figure 4.8 Must, have to and need to in British English (BE06)

4. Interpret the following models, outcomes of logistic regression, that come from a study of modal expressions of strong obligation. The study was guided by the following research question: in what contexts do speakers use *must* as opposed to semi-modals of strong obligation (*have to* and *need to*)?

The following variables were coded in the dataset:

- Outcome variable: *must* vs *have to* and *need to* combined (baseline).
- Predictor 1 (Variety): British vs American (baseline).
- Predictor 2 (Genre): fiction vs general prose vs press vs academic writing (baseline).
- Predictor 3 (Subject): *I* vs *you* vs other subject (baseline).

Model1: main effect 'Variety'

Overall model statistics: Likelihood ratio test (LL): 3.52 ($p = 0.061$) → NOT SIGNIFICANT; C-index: 0.52 → NOT ACCEPTABLE; Nagelkerke R^2 : 0; AIC: 2641.65

	Estimate (log odds)	Standard error	Z value (Wald)	p-value	Estimate (odds)	95% CI lower	95% CI upper
(Intercept)	-0.457	0.068	-6.711	0.000	0.633	0.554	0.723
VarietyB_BR	0.173	0.092	1.875	0.061	1.189	0.992	1.426

Model2: main effects 'Variety' and 'Genre'

Overall model statistics: Likelihood ratio test (LL): 54.49 ($p < 0.001$) → SIGNIFICANT; C-index: 0.6 → NOT ACCEPTABLE; Nagelkerke R^2 : 0.04; AIC: 2596.67

	Estimate (log odds)	Standard error	Z value (Wald)	p-value	Estimate (odds)	95% CI lower	95% CI upper
(Intercept)	-0.027	0.147	-0.184	0.854	0.973	0.729	1.300
VarietyB_BR	0.163	0.094	1.738	0.082	1.177	0.980	1.416
GenreB_Fiction	-0.870	0.165	-5.277	0.000	0.419	0.303	0.579
GenreC_General	-0.172	0.157	-1.092	0.275	0.842	0.618	1.146
GenreD_Press	-0.529	0.185	-2.860	0.004	0.589	0.410	0.846

Model3: main effects 'Variety', 'Genre' plus interactions

Overall model statistics: Likelihood ratio test (LL): 75.54 ($p < 0.001$) → SIGNIFICANT; C-index: 0.61 → NOT ACCEPTABLE; Nagelkerke R^2 : 0.05; AIC: 2581.63

	Estimate (log odds)	Standard error	Z value (Wald)	p-value	Estimate (odds)	95% CI lower	95% CI upper
(Intercept)	-0.027	0.147	-0.184	0.854	0.973	0.729	1.300
VarietyB_BR	-0.857	0.288	-2.979	0.003	0.424	0.240	0.742
GenreB_Fiction	-1.581	0.238	-6.646	0.000	0.206	0.128	0.326
GenreC_General	-0.578	0.225	-2.573	0.010	0.561	0.359	0.868
GenreD_Press	-1.166	0.266	-4.387	0.000	0.312	0.184	0.522
VarietyB_BR: GenreB_Fiction	1.440	0.337	4.275	0.000	4.221	2.190	8.214
VarietyB_BR: GenreC_General	0.893	0.321	2.785	0.005	2.443	1.308	4.605
VarietyB_BR: GenreD_Press	1.319	0.376	3.506	0.000	3.738	1.796	7.852

Model4: main effects 'Variety', 'Genre', 'Subject' plus 'Variety', 'Genre' interactions

Overall model statistics: Likelihood ratio test (LL): 145.56 ($p < 0.0001$) → SIGNIFICANT; C-index: 0.66 → NOT ACCEPTABLE; Nagelkerke R^2 : 0.1; AIC: 2515.61

	Estimate (log odds)	Standard error	Z value (Wald)	p-value	Estimate (odds)	95% CI lower	95% CI upper
(Intercept)	0.542	0.200	2.714	0.007	1.720	1.168	2.561
VarietyB_BR	-0.930	0.290	-3.210	0.001	0.395	0.222	0.693
GenreB_Fiction	-1.318	0.242	-5.441	0.000	0.268	0.165	0.428
GenreC_General	-0.450	0.228	-1.974	0.048	0.638	0.406	0.993
GenreD_Press	-1.191	0.268	-4.445	0.000	0.304	0.179	0.511
SubjectB_I	-1.084	0.174	-6.232	0.000	0.338	0.239	0.472
SubjectC_you	-0.917	0.158	-5.794	0.000	0.400	0.291	0.542
VarietyB_BR: GenreB_Fiction	1.482	0.340	4.353	0.000	4.400	2.267	8.620
VarietyB_BR: GenreC_General	0.952	0.324	2.941	0.003	2.592	1.379	4.915
VarietyB_BR: GenreD_Press	1.490	0.379	3.927	0.000	4.438	2.118	9.384

THINGS TO REMEMBER

- When analysing lexico-grammatical variation we need to pay attention to individual linguistic contexts and define a lexico-grammatical frame.
- Cross-tabulation can be used for a simple analysis of categorical variables. In addition to frequencies, cross-tab tables can also include percentages based on row totals (most useful for investigation of lexico-grammar), column totals and the grand total.
- The data in cross-tab tables can be effectively visualized using mosaic plots.
- We can test the statistical significance of the relationship between variables in a two-way cross-tab table (i.e. a table with one linguistic and one explanatory variable) using the chi-squared test. The effect sizes reported are Cramer's V (overall effect) and probability or odds ratios (individual effects).
- Logistic regression is a sophisticated multivariable method for analysing the effect of different predictors (both categorical and scale) on a categorical (typically binary) outcome variable.
- In logistic regression, we look at both the general performance of a model as well as individual coefficients showing the effect of the predictor variables on the outcome of interest.

Advanced Reading

- Balakrishnan, N., Voinov, V. & Nikulin, M. S. (2013). *Chi-squared goodness of fit tests with applications*. Waltham, MA: Academic Press.
- Friendly, M. (2002). A brief history of the mosaic display. *Journal of Computational and Graphical Statistics*, 11(1), 89–107.
- Geisler, C. (2008). Statistical reanalysis of corpus data. *ICAME Journal*, 32, 35–46.
- Gries, S. Th. (2013). *Statistics for linguistics with R: a practical introduction*. Berlin: De Gruyter Mouton, pp. 247–336.
- Hosmer, D. W., Lemeshow, S. & Sturdivant, R. X. (2013). *Applied logistic regression*, 3rd edn. Hoboken, NJ: John Wiley & Sons.
- Osborne, J. W. (2015). *Best practices in logistic regression*. Thousand Oaks, CA: Sage.

Companion Website: Lancaster Stats Tools Online

1. All analyses shown in this chapter can be reproduced using the statistical tools available from the companion website. The tools available for this chapter include:
 - Cross-tab
 - Categories comparison
 - Logistic regression calculator
2. The website also offers additional materials for students and teachers.

5 Register Variation

Correlation, Clusters and Factors

5.1 What Is This Chapter About?

This chapter discusses a group of methods that can be used for the simultaneous analysis of a large number of linguistic variables that characterize different texts and registers. First, we look at the relationship between two linguistic variables by means of correlation. Both Pearson's and the non-parametric Spearman's correlations are explained. Next, we explore the classification of words, texts, registers etc. using the technique of hierarchical agglomerative clustering. Several options for cluster identification are considered and the interpretation of the results of the cluster analysis (a tree plot) is laid out in full. Finally, the chapter deals with multidimensional analysis (MD), a methodology which uses factor analysis to extract patterns across multiple variables; the factors are then interpreted as functional dimensions of variation. Multidimensional analysis is described from the initial process of variable selection through to the interpretation of factor loadings and dimension plots.

We'll be exploring answers to four questions:

- How can we test the relationship between two linguistic variables? (Section 5.2)
- How do we classify words, texts, registers etc.? (Section 5.3)
- How can we explore different dimensions of linguistic variation? (Section 5.4)
- How can the techniques discussed in this chapter be used in research? (Section 5.5)

5.2 Relationships between Variables: Correlations

Think about ...

Before reading this section, look at Figures 5.1–5.3 (scatterplots). In each graph, the relative frequencies of two linguistic variables are plotted, one on the x-axis and one on the y-axis. Each point represents an individual text in the BE06 corpus, a one-million-word sample of written British English. Is there any apparent relationship between the two linguistic variables in each graph?

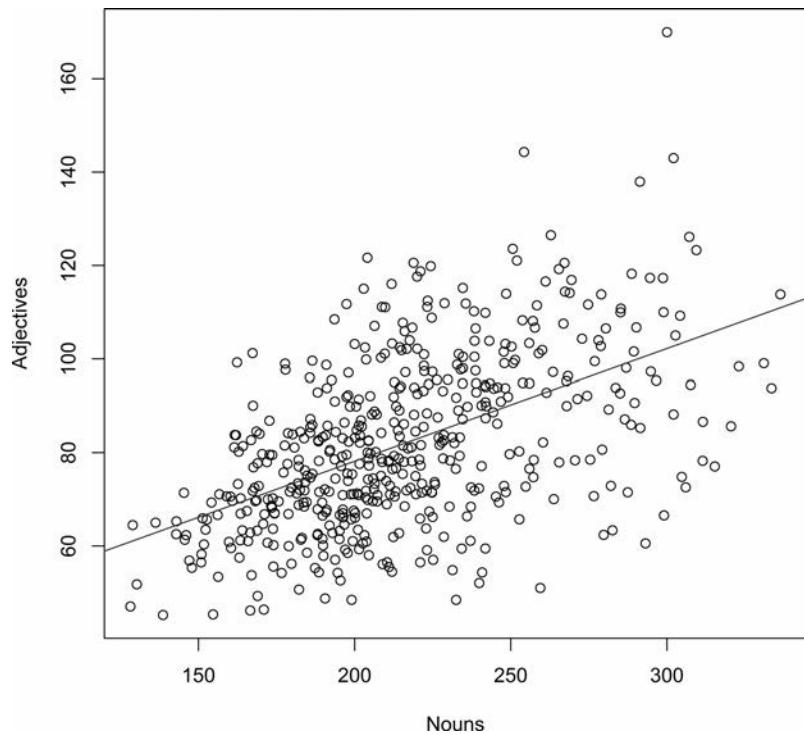


Figure 5.1 *Nouns and adjectives in BE06*

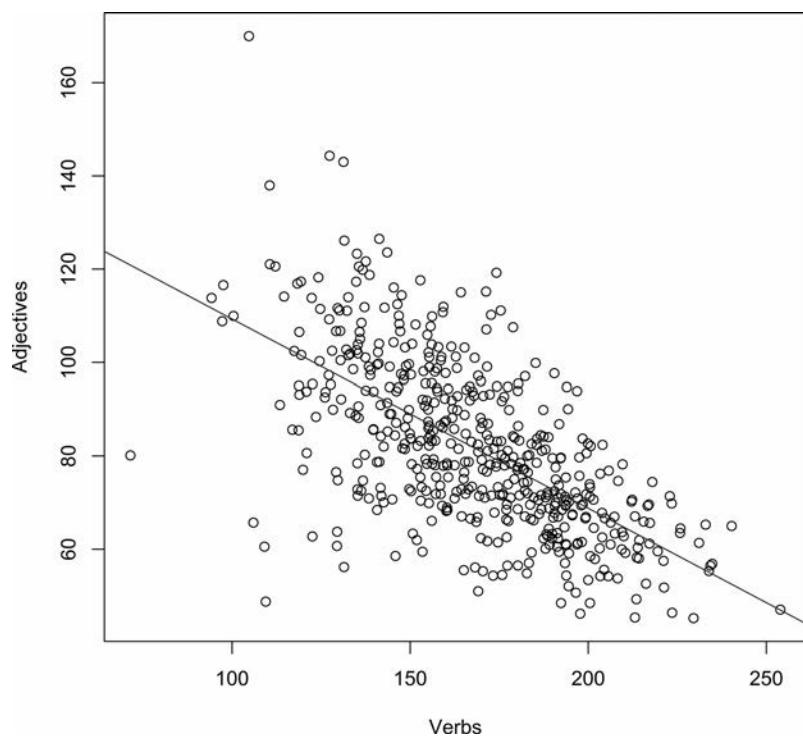


Figure 5.2 *Verbs and adjectives in BE06*

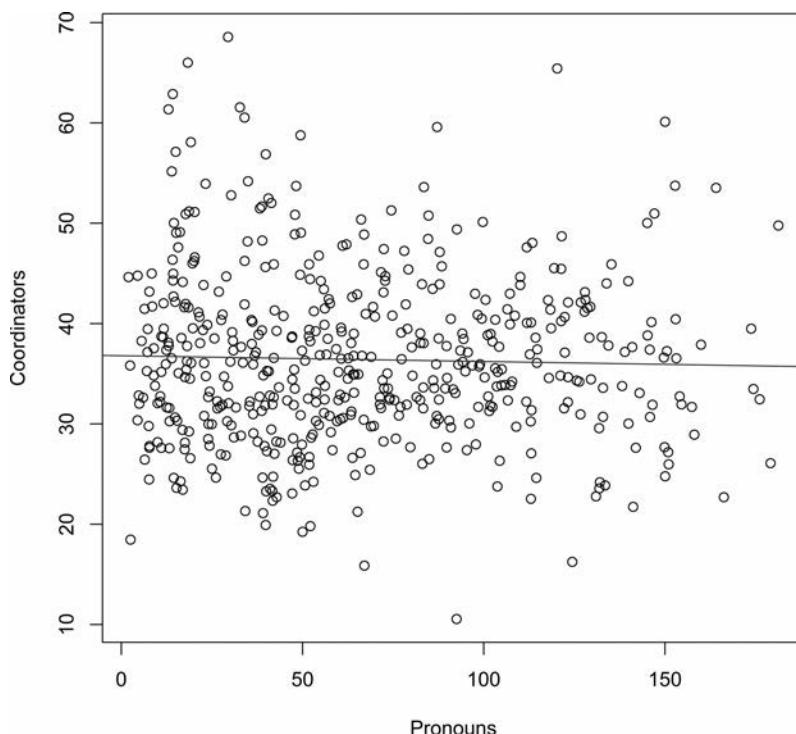


Figure 5.3 *Pronouns and coordinators in BE06*

By virtue of being in a text together, many linguistic variables are related in some way. For example, in English the relative frequency of adjectives in texts is related to the relative frequency of nouns because adjectives modify nouns and therefore typically occur together with them; at the same time, this doesn't mean that nouns do not occur without adjectives. As can be seen from Figure 5.1, texts with a large number of nouns tend also to have a large number of adjectives; this tendency is highlighted by the steep rising regression line (see Section 1.2) that goes through the middle of the data points. We can measure the relationship between two linguistic variables using a technique called correlation. **Correlation** measures whether two typically ordinal or scale variables (see Section 1.3) are related by looking at the extent to which they covary. In other words, we are looking at whether, if one variable increases, the other variable increases, decreases or stays the same. When we see a pattern like that in Figure 5.1, where the increase in the values of the first variable (nouns) means also an increase in the values of the second variable (adjectives), we talk about a **positive correlation**; on the other hand, if as one variable (verbs) increases, the other (adjectives) decreases, as in Figure 5.2, we talk about a **negative correlation**. Finally, when we see data points scattered in an apparently random way as in Figure 5.3, which shows the

relative frequencies of pronouns and coordinators in texts, we can conclude that there is little or no relationship, in other words virtually no correlation, between these two variables.

There are two basic kinds of correlation – **Pearson's correlation** and **Spearman's correlation** (Sheskin 2007: 1219–1368). Pearson's correlation is designed to work with scale variables, while Spearman's correlation assumes ranks (ordinal variables). Note that scale variables can be turned into ranks at the cost of losing some information, but this cannot be done the other way around, e.g. given only rank information, you cannot infer the original observed frequencies of a set of features. So Spearman's correlation can be used even with scale linguistic variables (such as the variables in Figures 5.1–5.3), which would first get converted into ranks. This is sometimes done when the values of the variables are largely skewed, i.e. if the mean is not a good model for the data and the p-values related to Pearson's correlation might be unreliable. However, it has been shown (Edgell & Noon 1984) that this is not a serious issue and therefore Pearson's correlation can be recommended as a default with scale linguistic variables. So how is correlation calculated? **Pearson's correlation (r)** can be expressed as follows:

$$r = \frac{\text{covariance}}{SD_1 \times SD_2} \quad (5.1)$$

It expresses the amount of covariance (variation that the variables have in common) in the data in terms of the standard deviations (SD_1 and SD_2) of the two variables in question; the combination of standard deviations here is used as the standardized measurement unit of ‘statistical distance’. Covariance is calculated by computing the means for variables 1 and 2 (mean_1 and mean_2), taking each single value of variable 1, calculating the distance from mean_1 and multiplying this by the distance of variable 2 from mean_2 . This is expressed by the equation below.

$$\text{covariance} = \frac{\text{sum of multiplied distances from } \text{mean}_1 \text{ and } \text{mean}_2}{\text{total no. of cases} - 1} \quad (5.2)$$

The process might look complicated, but the idea is very simple. For illustration, let's extract five texts from those used to produce Figure 5.1 and focus on the relationship between nouns and adjectives in them. Figure 5.4 shows a clear positive correlation between nouns and adjectives in the five selected texts. Figure 5.5 demonstrates how covariance is measured with the same five data points: for each data point (represented by a circle), we measure the distances from the two means; the means are represented by the dash-dotted vertical and the dash-dotted horizontal line. These distances are first multiplied and then added together. We can see that, with the exception of one data point (H04), all data points in Figure 5.5 are in the 1st or the 3rd Quadrant created by the two means (dash-dotted lines). This implies that

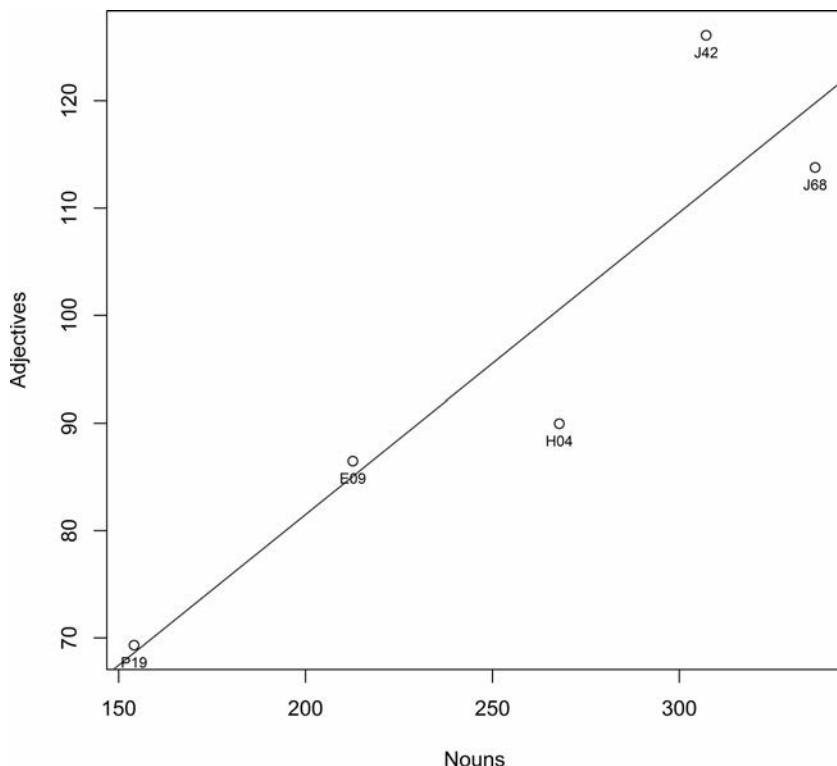


Figure 5.4 Correlation: five data points

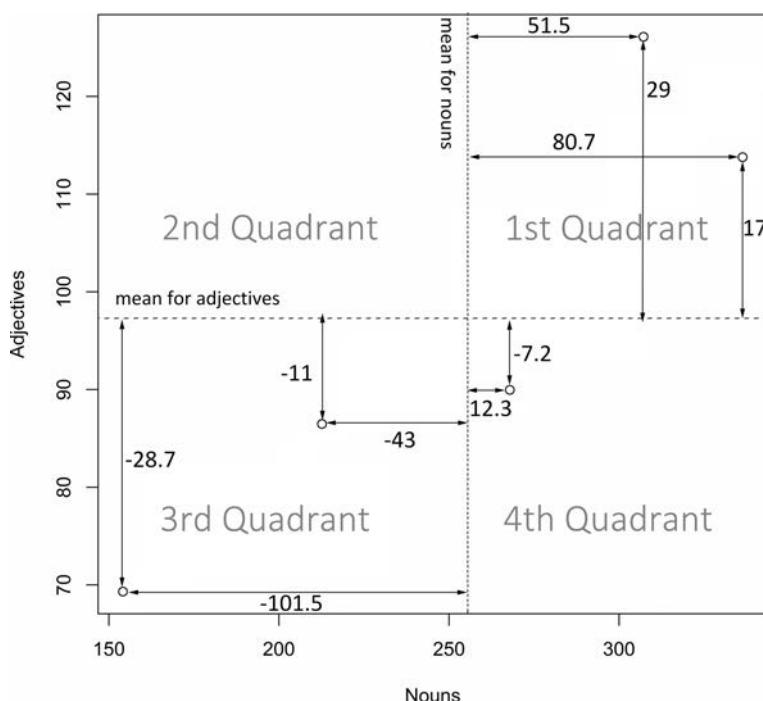


Figure 5.5 Correlation: covariance

the distances from the means are either both positive (1st Quadrant) or both negative (3rd Quadrant) and thus the result of the multiplication of the two distances and hence the overall sum is a positive number. On the other hand, if most data points were in the 2nd and the 4th Quadrants (as in Figure 5.2), one distance would be positive and one negative and hence the result of the multiplication and addition would be a negative number. Finally, if the data points were scattered randomly in all four Quadrants (as in Figure 5.3), we would get both positive and negative results of the multiplication which would cancel each other out when added together, resulting in covariance close to zero.

The covariance of nouns and adjectives in the five texts from Figure 5.5 can be calculated as:

$$\text{covariance} = \frac{(-27.8 \times -101.5) + (-11 \times -43) + (12.3 \times -7.2) + (51.5 \times 29) + (80.7 \times 17)}{5 - 1} = 1518 \quad (5.3)$$

The covariance is then entered into the equation for Pearson's correlation and standardized using standard deviations. The standard deviations are calculated according to equation (2.11) from Chapter 2 (standard deviation sample).

$$r = \frac{1518}{73.3 \times 23} = 0.9 \quad (5.4)$$

In this case, the correlation between nouns and adjectives in the five texts is positive (as is clear from Figure 5.4). It is also very large (0.9) – the correlation coefficient always ranges from –1 to 1, with negative numbers indicating negative correlation and positive numbers indicating positive correlations. Zero means that there is no linear relationship between the two variables in question. As a rough indication, the following cut-off values are used to indicate the size of correlation (Cohen 1988: 79–80); this is a measure of effect size:¹

- 0 no effect
- ± 0.1 small effect
- ± 0.3 medium effect
- ± 0.5 large effect

In addition, the correlation coefficient should be complemented with a p-value or a confidence interval (CI) to indicate whether there is enough evidence in the corpus to generalize the correlation to the population. The p-value is a result of a test that evaluates the null hypothesis which states that the correlation in the population is 0 (i.e. there is no correlation). The statistical significance of a correlation is directly related to the number of observations (cases). With

¹ The suggested interpretation of the r effect size shouldn't be applied mechanically; it should be used as a general guide, with attention being paid to the practical implications of different effect size values for the discipline. For more information see Section 8.4.

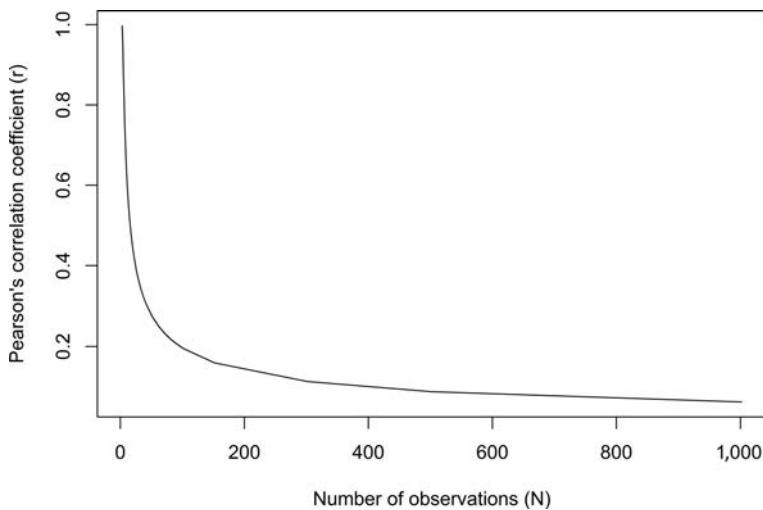


Figure 5.6 *Statistically significant ($p < 0.05$) Pearson's correlations in relation to the number of observations*

large numbers, which are typical in corpus research, small correlations are statistically significant, as can be seen from the graph in Figure 5.3. These correlations are, however, not always necessarily important from the linguistic perspective. For example, the negligible correlation ($r = -.029$) between pronouns and coordinators in Figure 5.3 is not statistically significant with 500 texts; with 5,000 texts, however, the same correlation value becomes statistically significant at $p < .05$ level, although the practical linguistic effect is negligible.

Figure 5.6 shows the relationship between the number of observations and statistically significant ($p < .05$) Pearson's correlations. The general trend is this: with an increasing number of observations we need smaller correlations to reach significance. For example, if we have only ten observations (texts or speakers) we need a large correlation of 0.63 to reach statistical significance; this means that all data points need to show the relationship clearly, with only a small amount of fluctuation being acceptable in order to consider the evidence satisfactory to draw a reliable inference about the population. With 100 observations, more fluctuation is acceptable and the necessary critical value of Pearson's correlation drops to 0.2. With 1,000 observations, a negligible correlation of 0.06 will be statistically significant. From the linguistic perspective, the correlation of 0.06 (or 0.2 for that matter) has very little practical impact in terms of the strength of the relationship between the variables. However, it is still important to report this correlation as a result of the analysis because it can be revealing about how language works. Correlations are thus yet another example of how effect sizes (the correlation coefficient r) and statistical significance (p -value) measure two

very different things which should not be confused (see Sections 1.4 and 8.4 for a detailed discussion of this point).

Let's return to our example of the five texts from BE06 displayed in Figure 5.4. Because the texts show the relationship between the variables very clearly, the p-value connected with the Pearson's correlation is 0.033. This indicates that despite having only five cases, the correlation (0.9) is statistically significant. In other words, we have enough evidence in the data to reject the null hypothesis which states that the correlation in the population is 0. However, if we want to be more specific and estimate how small or large the correlation is likely to be in the population, we need to calculate the **confidence intervals (CI)** for the correlation. You don't need to know the mathematical details² because computers can produce these very easily (see the Correlation tool in Lancaster Stats Tools online). So let us focus on how to interpret these figures. The 95% CI for our example is from 0.128 to 0.994. Note that it does not include 0, which is in line with the significant p-value; the 95% CI thus indicates that whatever the exact correlation is in the population, it will in any case be positive not null. The 95% confidence interval³ is an interval that most likely contains the true value of the correlation in the population (all texts out there). This means that with the evidence we have (only five texts) we cannot be sure if the correlation is truly as high as 0.9 because it can be as low as 0.128 or can also be even higher than 0.9 (0.994). So, we would need to approach this result with caution.

Let's move on to the Spearman's correlation. **Spearman's correlation (r_s)**, sometimes also denoted by the Greek letter *rho* (ρ), is used with ordinal data (ranks) or with scale data when the parametric assumptions are violated; in the latter case, the scale data is converted into ranks. Because we are dealing with ranks, we have no means, *SDs* or distances from the mean at our disposal. Instead, covariance is measured by looking at the differences between the ranks. Spearman's correlation is therefore calculated as follows:

$$r_s = 1 - \frac{6 \times \text{sum of squared rank differences}}{\text{number of cases} \times (\text{number of cases squared} - 1)} \quad (5.5)$$

Let's take again the five texts from Figure 5.4. This time, however, the values on the x and y-axes (i.e. frequencies of nouns and adjectives respectively) need to be converted into ranks, as has been done in Table 5.1.

² For those interested in the equation: Lower limit $_z = z' - 1.96 \times \frac{1}{\sqrt{\text{number of observations} - 3}}$; Upper limit $_z = z' + 1.96 \times \frac{1}{\sqrt{\text{number of observations} - 3}}$ where $z' = 0.5 \ln \left(\frac{1+r}{1-r} \right)$; once computed, Lower/upper limit $_z$ need to be converted back to r .

³ Remember that the number 95 indicates the percentage of samples taken from the same population for which the confidence interval contains the true value (i.e. value in the population) of the measure (see Section 1.3).

Table 5.1 *Ranks of nouns and adjectives in five texts from BE06*

File	Nouns		Adjectives		Nouns – Adjectives	
	RF (per 1,000)	Rank	RF (per 1,000)	Rank	Rank difference	Squared rank difference
BE_E09	212.7	4	86.5	4	4 – 4 = 0	0
BE_H04	267.9	3	90.0	3	3 – 3 = 0	0
BE_J42	307.2	2	126.1	1	2 – 1 = 1	1
BE_J68	336.3	1	113.8	2	1 – 2 = -1	1
BE_P19	154.1	5	69.3	5	5 – 5 = 0	0

When the squared rank differences from Table 5.1 and the number of cases (5) are entered into the equation we get the following:

$$r_s = 1 - \frac{6 \times (0 + 0 + 1 + 1 + 0)}{5 \times (5^2 - 1)} = 0.9 \quad (5.6)$$

In this case, the non-parametric correlation (r_s) is once again very large – the same cut-off values as for Pearson's correlation, i.e. 0.1, 0.3 and 0.5, are used for conventional interpretation of the strength of the effect. However, the p-value is larger than 0.05 ($p = .083$), i.e. by convention we conclude that we don't have enough evidence to reject the null hypothesis which says that the correlation in the population is 0. The five texts therefore do not provide enough evidence that Spearman's correlation is not null – what we are seeing could be the result of chance. The difference between the p-value associated with Pearson's and the p-value associated with Spearman's correlation can be explained by the fact that by converting the actual values to ranks we lost some information and hence also the power to reject the null hypothesis. For this reason, Pearson's correlation is preferable with scale variables, while Spearman's correlation is used with ordinal variables (ranks).

Finally, two remarks need to be made. First, Pearson's correlation coefficient r can be used to account for the amount of variability in one variable shared by the other variable. For this, the coefficient needs to be squared; the product, r^2 , is called the **coefficient of determination** and is calculated using the simple equation below:

$$\text{Coefficient of determination } (r^2) = \text{Pearson's correlation coefficient}^2 \quad (5.7)$$

Let's take, for example, the nouns and adjectives in BE06 (Figure 5.1). The correlation coefficient r is .523 (large effect). When we square this value, we get 0.27 or 27%.

$$r^2 = 0.523^2 = 0.27 \quad (5.8)$$

This can be interpreted as 27% shared variation between the relative frequencies of nouns and the relative frequencies of adjectives in texts. So is 27% a large or a small number? Consider the following: 27% of variation explained leaves 73% of variation unexplained, i.e. explained by other variables that we didn't consider because we looked at the correlation between these two variables only. However, language is a very complex system with a large number of variables that play a role, so if one variable can be used to predict almost 30% of variation in another variable, this is worthy of note.

Second, when dealing with multiple linguistic variables we can calculate pairwise correlations (either Pearson's or Spearman's). These are usually displayed as tables (correlation matrixes), series of scatterplots or visualization matrixes. As an example, take nouns, adjectives, verbs, pronouns and coordinators in BE06 as our linguistic variables of interest. What is the relationship between these five linguistic variables? The following are three different modes of presenting the results of the same analysis.

Figure 5.7 can be used for an initial exploration of the main tendencies in the data. We can see that, with the exception of coordinators (words such as *and*, *but* and *or*), the other four variables when considered pairwise show either positive or negative relationships. On the other hand, coordinators, as can be seen from the last row or the last column, do not predict the occurrence of the other word classes very well.

The initial observation can be confirmed by looking at Pearson's correlations, shown in Table 5.2, which range from small (0.12) to large (0.81). The self-correlations (1.00) on the diagonal can be disregarded because it is obvious that the values of a variable perfectly correlate with themselves. The smallest value in the table is very close to zero (-0.03), which indicates no relationship between the variables. With 500 cases (texts) even small correlations (0.12) are significant at $p<0.01$ level. The only non-significant value is the nearly zero correlation (-0.03).

For those who prefer further visualization of the results reported in Table 5.2, a correlation matrix (Figure 5.8) is available, which uses the depth of

Table 5.2 Correlation table (Pearson's correlations): nouns, adjectives, verbs, pronouns and coordinators

	Nouns	Adjectives	Verbs	Pronouns	Coordinators
Nouns	1.00**	0.52**	-0.65**	-0.79**	0.12**
Adjectives	0.52**	1.00**	-0.63**	-0.59**	0.12**
Verbs	-0.65**	-0.63**	1.00**	0.81**	-0.14**
Pronouns	-0.79**	-0.59**	0.81**	1.00**	-0.03
Coordinators	0.12**	0.12**	-0.14**	-0.03	1.00**

** $p < .01$

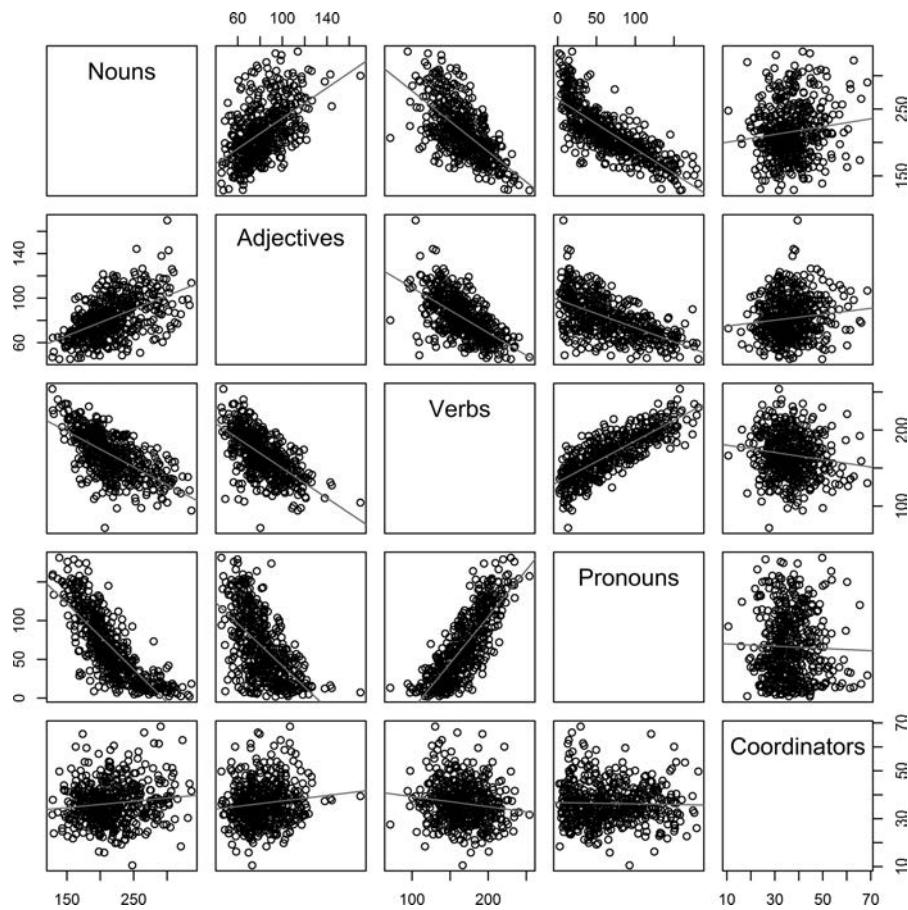


Figure 5.7 Multi-panel scatterplot: nouns, adjectives, verbs, pronouns and coordinators

shading to convey the strength of the correlation – the stronger the correlation the darker the shade. In a colour version of the image, positive and negative correlations are also distinguished by the use of different colours.

To sum up, correlation is a powerful technique for exploring the relationship between variables in corpora. With large corpora, which consist of hundreds or thousands of files, even small correlations (smaller than 0.1) will be statistically significant. We therefore need to consider very carefully the meaning of the correlations for the linguistic relationships between different features of language.

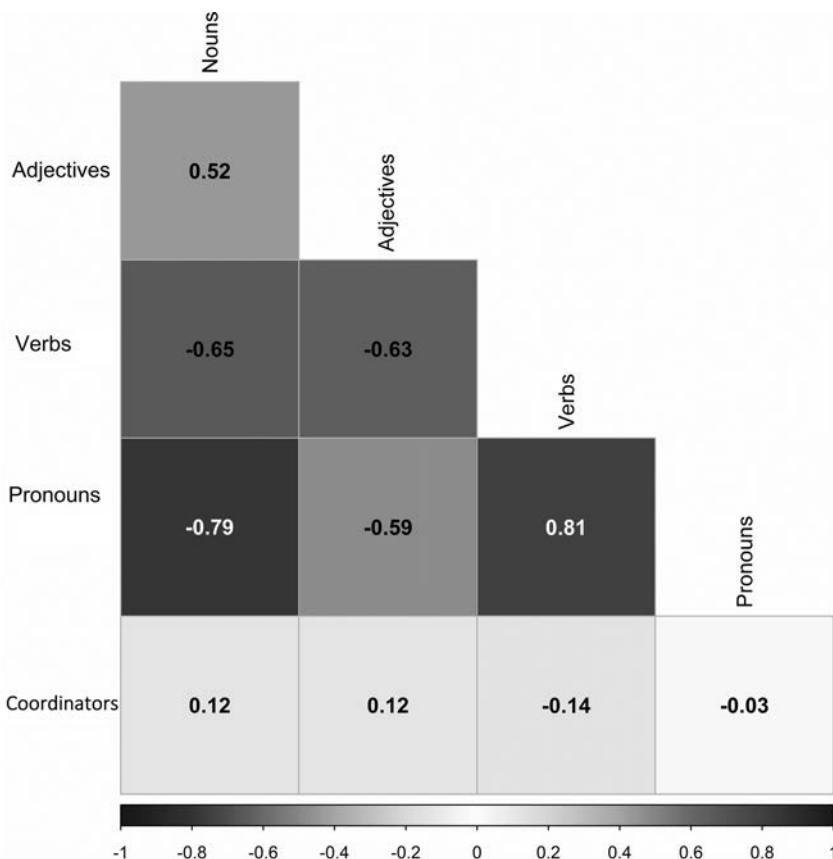


Figure 5.8 *Correlation matrix: nouns, adjectives, verbs, pronouns and coordinators*

Reporting Statistics: Correlations

1. What to Report

Traditionally, the correlation coefficients (r , r_s) are reported together with the related p-values. Often, however, reporting confidence intervals (CIs) instead of p-values is preferable because CIs provide a more precise estimate about the actual value of the correlation in the population. It is also important to interpret the size of the correlation (effect size) – the observed correlation is best compared with similar correlations in the data or those reported in the literature.

The most economical way of reporting correlation (used especially when reporting multiple correlations in a table) is to add a single (*) or double (**) asterisk next to the correlation coefficient. This conventionally means $p < .05$ and $p < .01$ respectively. We can also type out the p-values (although with large corpora with many files these are always very low) or specify the CIs.

2. How to Report: Examples

- There is a strong positive correlation ($r = .52$, 95% CI [.46, .58]) between the number of nouns and the number of adjectives used in English texts. This value, however, is not as large as the correlation between verbs and pronouns ($r = .81$, 95% CI [.775, .836]), which explain each other's occurrence in two-thirds of the cases ($r^2 = 66\%$).
- There is a very strong negative correlation ($r_s = -.83$, $p < .01$) between the number of nouns and the number of pronouns used in English texts. These show a complementary distribution.
- English verbs and adjectives in written texts are in an inverse proportional relationship ($r = -.63^{**}$). So are verbs and nouns ($r = -.65^{**}$). The negative correlation between verbs and coordinators is only small ($r = -.14^{**}$) with little observable impact on the style of writing.

5.3 Classification: Hierarchical Agglomerative Cluster Analysis

Think about ...

Before reading this section, think about the following question: which colour terms in Figure 5.9 belong together based on their frequency in the BNC and word length? Draw circles around the groups.

In the previous section, we looked at the relationship between linguistic variables. We saw that many of them are related to some degree. In this section, we'll shift the focus to 'objects' (words, sentences, texts or speakers etc.) that can be characterized using multiple linguistic variables. Instead of looking at the relationship between linguistic features, we'll be looking at the relationship between objects as characterized by these features. The question we will be asking is relatively simple and is demonstrated in the 'Think about' task: how do we classify objects based on linguistic variables? Here, we used a simple example of colour terms characterized by the frequency of use and word length. We saw that *black* and *white* are the most frequently used colour words, while more specific terms such as *aquamarine*, *turquoise* and *burgundy* are much less frequent. The latter terms are also the longest and for some perhaps more difficult to pronounce. So how would we go about classifying these colour terms based purely on what we know about their frequency and number of letters? An obvious way of going about this would be to use the plot (Figure 5.9) and look at the

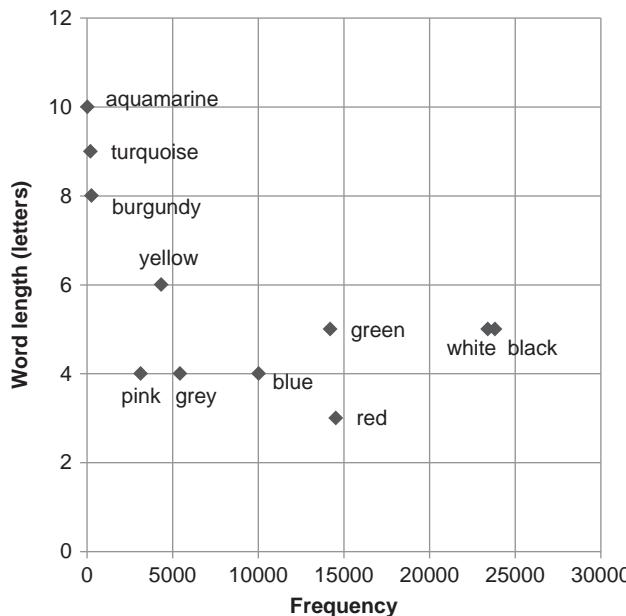


Figure 5.9 Colour terms in the BNC

distances between individual points. Those closest together would belong to the same group. If we try this approach, we'll probably recognize four main groups in the plot: (1) *black* and *white*, (2) *blue*, *red* and *green*, (3) *pink*, *grey* and *yellow*, and (4) *aquamarine*, *turquoise* and *burgundy*. So far, so good. However, how exactly do we measure the distance? Take, for example, the distance between *green* and *blue*. In the graph, *green* has the coordinates [14205, 5] and *blue* has the coordinates [10035, 4]. If we look at the graph carefully, we'll also notice that there is a discrepancy between the scale on the x-axis (ranges from 0 to 30 k) and the y-axis (ranges from 0 to 12). This is because on the x-axis we have plotted word frequencies in the BNC which are very large numbers, while on the y axis we have plotted the number of letters in a colour term with the maximum value of 10. If we used the same scale (0-30 k) for the two axes, we'd give much more weight to word frequencies and almost disregard the length of the words thus virtually reducing the plot to a single dimension that can be displayed as a line. To overcome this incompatibility of scales problem, we can transform the values to z-scores₂. The subscript indicates that we are using the term 'z-score' in a different sense than in Chapter 3; z-scores₂ are standardized values of a scale variable of interest that expresses how many standard deviations a value is from the mean. z-scores₂ are frequently used in situations in which we have variables measured on different scales yet want to give each variable the same weight in the analysis. z-scores₂ are calculated by taking each individual value of a variable, subtracting the mean and dividing it by the standard deviation (the mean is

calculated using equation (1.1) from Chapter 1; the standard deviation is calculated according to equation (2.11) from Chapter 2):

$$z\text{-score}_2 = \frac{\text{value} - \text{mean}}{\text{standard deviation (sample)}} \quad (5.9)$$

The new $z\text{-score}_2$ coordinates for the two colours are: *green* [0.58, -0.32], *blue* [0.11, -0.76].

Returning to the question of how to calculate the distance between *green* and *blue* in the graph in Figure 5.9, we can see that the shortest way from one point (A) to another (B) is via a direct line between the two points. This is called the **Euclidean distance**. It is calculated according to the following formula:

$$\text{Euclidean distance (A, B)} = \sqrt{(x_B - x_A)^2 + (y_B - y_A)^2 + (z_B - z_A)^2} \dots \quad (5.10)$$

where x_A is the first coordinate of point A, x_B is the first coordinate of point B, y_B is the second coordinate of point B etc. We can keep adding variables determining the position of the objects and move from a two-dimensional space to a multi-dimensional space.

In a two-dimensional space created by the two linguistic variables as in our example, we will use only two pairs of coordinates: x_A , x_B and y_A , y_B .

$$\text{Euclidean distance (green, blue)} = \sqrt{(0.11 - 0.58)^2 + [-0.32 - (-0.76)]^2} = 0.64 \quad (5.11)$$

However, there are alternative approaches to calculating the distance between A and B. Imagine you are in a big city such as New York and you want to get from A to B. Unless you have a helicopter, you won't be able to go directly. Instead, you'll have to go down one street, then make a 90 degree turn and go down another street. The distance between A and B measured like this, i.e. following the grids at right angles, is called the **Manhattan distance**. Manhattan distance is calculated according to the following formula:

$$\text{Manhattan distance (A, B)} = |x_B - x_A| + |y_B - y_A| + |z_B - z_A| \quad (5.12)$$

where $|x_B - x_A|$ etc. is the absolute value, i. e. a positive number of the difference between the coordinates of A and B.

For *green* and *blue* in our example, we'll get:

$$\text{Manhattan distance (green, blue)} = |0.11 - 0.58| + |-0.32 - (-0.76)| = 0.91 \quad (5.13)$$

As expected, the Manhattan distance is larger than the direct Euclidean distance, although when used for cluster analysis both distance measures yield fairly similar results; however, Manhattan distance is more robust when dealing

with outliers (see Section 1.3 for a definition of an outlier). Other types of distance measures include Canberra distance (which is a standardized form of Manhattan distance), Squared Euclidean distance (which places more emphasis on objects further apart) and Percent disagreement (used when working with categorical variables) (Everitt et al. 2011).

At this stage, we have all we need to start the cluster analysis. There are many different types of cluster analysis (Everitt et al. 2011) – the one typically used in corpus linguistics for non-diachronic data⁴ is **Hierarchical agglomerative cluster analysis** (Gries 2013a: 336ff). The term, which may sound somewhat intimidating, is in fact a very good descriptive summary of the process. We take the individual data points and in a step-by-step (hierarchical) procedure join (i.e. agglomerate) the closest ones until we create one large cluster containing all the data points. Figure 5.10 illustrates this process, each panel showing an individual step in the procedure.

As can be seen from Figure 5.10, first *black* and *white* are joined in a small cluster followed by *pink* and *grey*, and *aquamarine*, *turquoise* and *burgundy*. After this, a larger cluster is created by combining the *pink* and *grey* mini-cluster with *blue* (specific details about how this is done and how the distance is measured are explained below). If we fast forward the process, we'll get the cluster diagram displayed in Figure 5.11 showing all individual steps in the cluster procedure as well as the final result – one large cluster (cluster 9) that contains all data points.

However, with multiple clusters the cluster diagram quickly becomes too complex and potentially hard to read. For this reason, the results of the cluster procedure are often displayed as a **hierarchical tree plot (or dendrogram)**. The tree plot seen in Figure 5.12 shows the individual steps of the cluster procedure (the numbers are usually not shown) as smaller branches (clusters) gradually converging into one large cluster. The best way to read the tree plot is from the bottom up starting with the clusters that are closest together as defined by the selected distance measure (Euclidean, Manhattan, Canberra etc. distance). The value of the distance between data points/clusters is displayed as the height of the tree plot – the larger the height, the larger the distance between the data points/clusters that are merged. For example, in Figure 5.12 the Euclidean distance between *black* and *white* on z-score₂ scale is close to zero as indicated on the y-axis, whereas the distance between the *black* and *white* cluster and the rest of the data points is close to one.

So far, we have assumed that the method of joining small clusters into larger ones is straightforward. However, in the cluster procedure we need to specify exactly how this is to be done. The question we need to ask is: which of the data points inside a small cluster should be taken as representing the

⁴ For diachronic (historical) data another type of cluster analysis, Neighbouring cluster analysis, is used. This technique is discussed in Section 7.4.

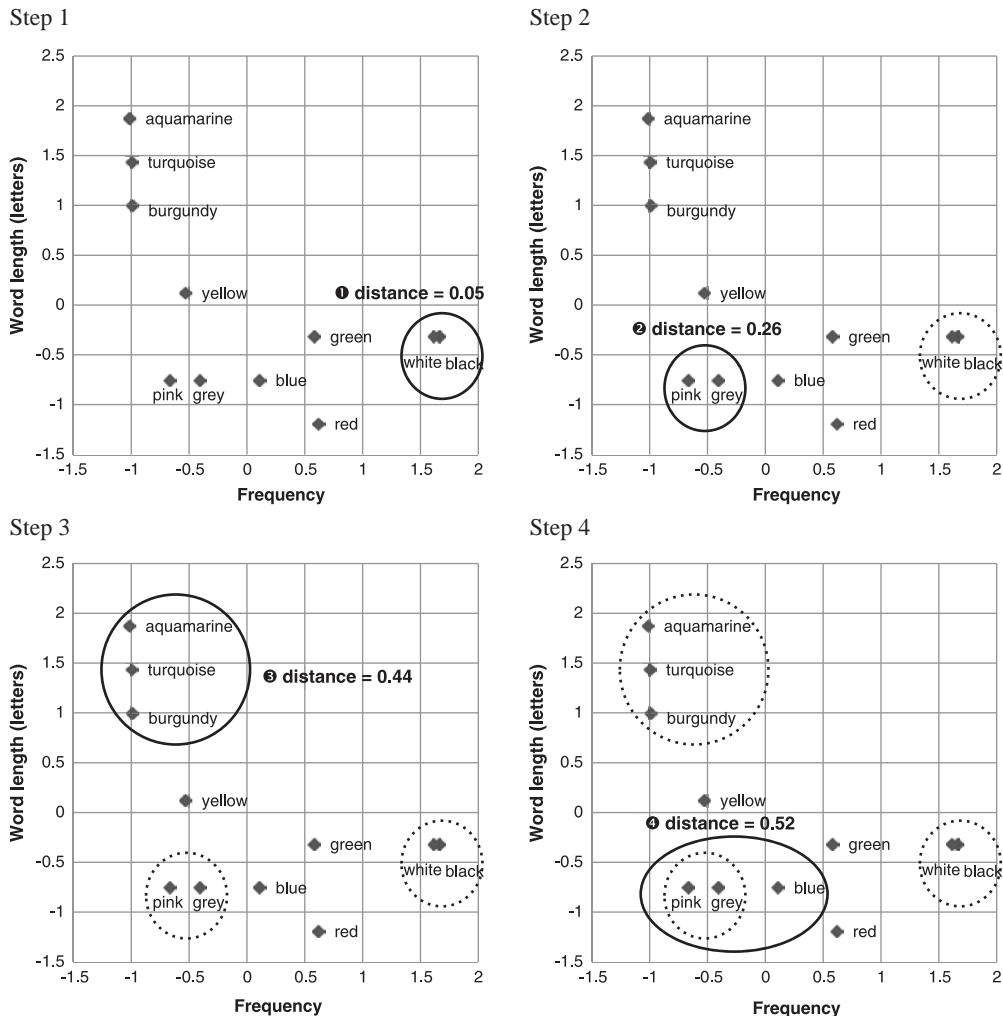


Figure 5.10 *Creating clusters: Steps 1-4*

position of the whole cluster? Typically, four answers can be provided: (1) the closest point to the neighbouring cluster with which we want to merge our original cluster (so-called SLINK method), (2) the furthest point to the neighbouring cluster with which we want to merge our original cluster (so-called CLINK method), (3) none, the mutual distances of all data points are considered by taking their mean value (average linkage method), (4) none, mutual distances of all data points are considered by calculating the sum of squared distances (Ward's method). It is important to realize that different methods of joining clusters produce different results, as can be seen

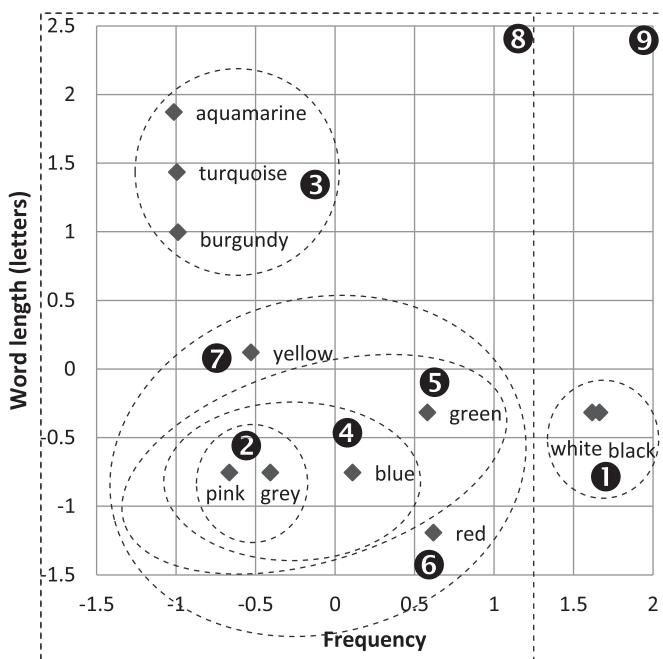


Figure 5.11 *Creating clusters: final result*

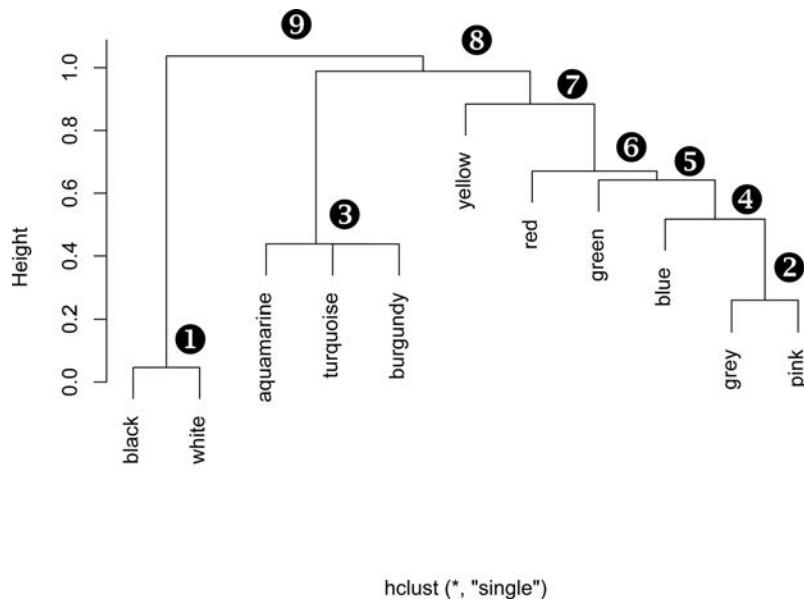


Figure 5.12 *Colour terms: a tree plot (dendrogram) – z-score₂ normalized, Euclidean distance, SLINK method*

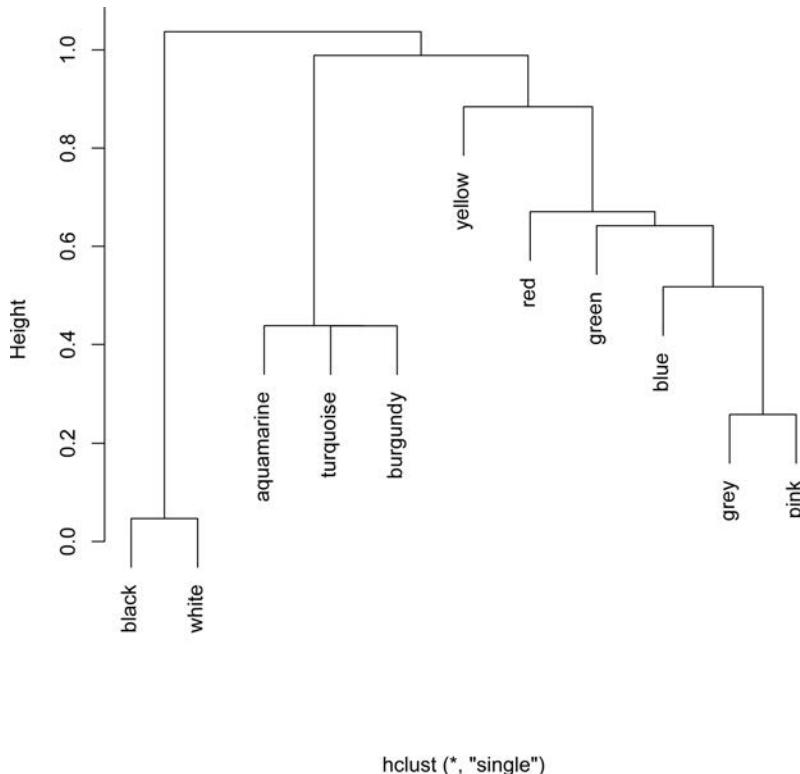


Figure 5.13 Tree plot: SLINK method

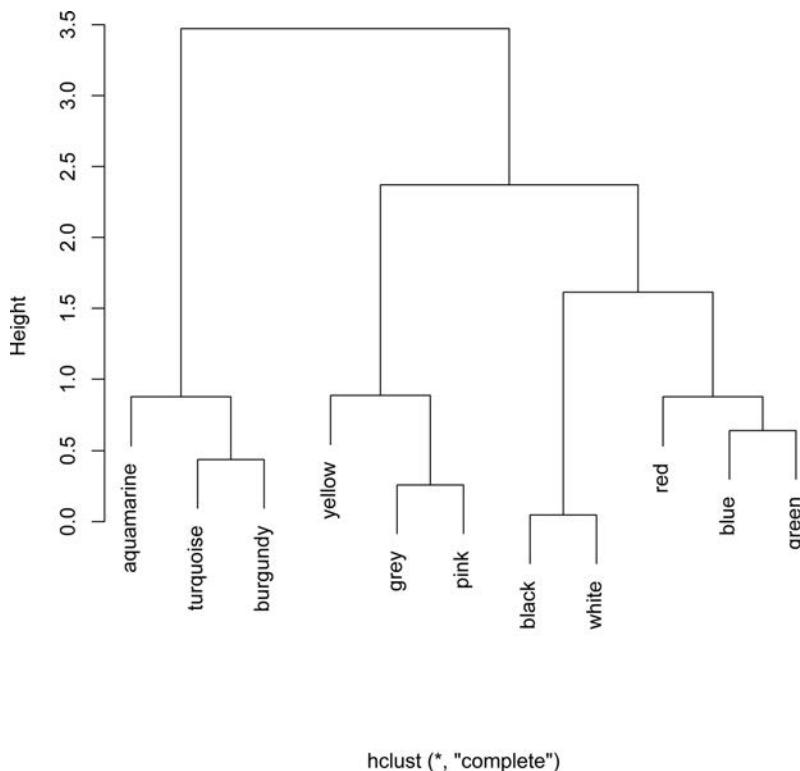


Figure 5.14 Tree plot: CLINK method

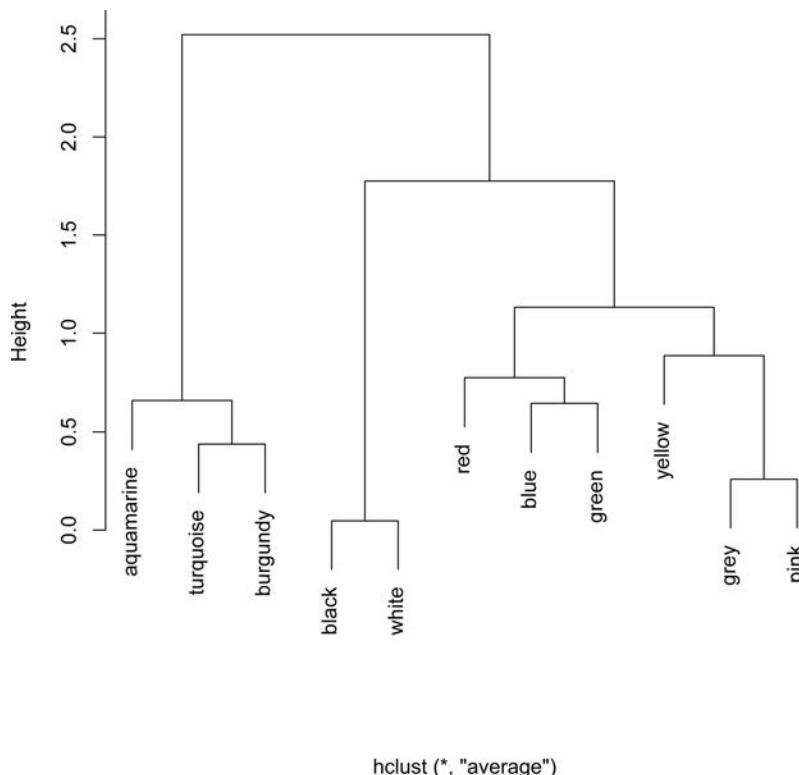


Figure 5.15 *Tree plot: average linkage method*

from Figures 5.13–5.16. Note that the SLINK (Single linkage) method was used to produce the tree plot in Figure 5.12.

As always, there are benefits but also disadvantages to each method. The SLINK method, for instance, is very simple. It leads, however, to the ‘chaining’ effect visible in the right branch of the dendrogram in Figure 5.13, creating a fairly loose cluster structure without prominent cluster groups. In corpus linguistics, Ward’s method seems to be the most popular because it creates compact clusters that are relatively easy to interpret (cf. Divjak & Gries 2006; Gries et al. 2009).

One final point about clusters needs to be made. So far in the example we have considered two linguistic variables in a two-dimensional space. The cluster technique can also be used with multiple linguistic variables. The procedure is similar to the simple example described above with the only difference being the fact that with multiple variables we are looking at the distances between data points in a multidimensional space. An example of the use of the cluster technique with 44 linguistic variables is offered in Section 5.5.

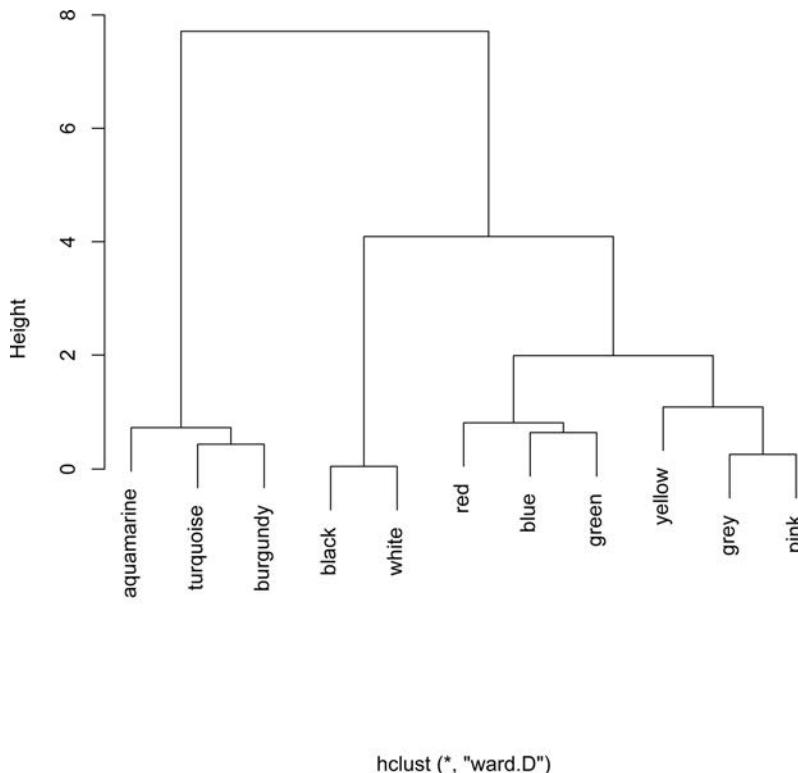


Figure 5.16 Tree plot: Ward's method

Reporting Statistics: Hierarchical Agglomerative Cluster Analysis

1. What to Report

Cluster analysis is largely an exploratory visual method to show patterning in the data. Both the parameters for cluster identification (data transformation, distance measure, cluster combination method) as well as the results of the analysis (tree plot) need to be reported. The Method section of the research report describes the analytical procedure and the parameters used. In the Results and Discussion, each tree plot (dendrogram) needs to be carefully discussed. The main question to be addressed is: how many meaningful factors can be observed in the plot?

2. How to Report: An Example

- The data was analysed using the hierarchical agglomerative cluster technique (z -score data transformation, Manhattan distance, Ward's method).

For an example of reporting and discussing results of the cluster analysis see Section 5.5.

5.4 Multidimensional Analysis (MD)

Think about . . .

Before reading this section, think about the following question:

How do the two excerpts below taken from the BNC differ in terms of the language they use?

Excerpt A

MARGARET:	We shall go back, erm after Easter
BOB:	Yes
MARGARET:	hoping, permitting, you know, if it's not too expensive, it got very dear you know
BOB:	Yes, that is also a thing to
MARGARET:	Yeah
BOB:	erm, I'm, I'm feeling a bit hard up at the moment, I had a
MARGARET:	Yes
BOB:	a bill for repairing the car for two hundred and thirty pounds and <pause> so

Excerpt B

MacCulloch (1819), and later, Murchison and Geikie (1861), provided the first accounts of the geology of Lewis and Harris and the other islands which constitute the north east–south west chain of islands called the Outer Hebrides, which lies some 70 km west of the northern Scottish mainland. In 1923, Jehu and Craig produced the first detailed account of the geology of this region, and followed it up with further accounts between 1925 and 1934. Thereafter, several research papers were published on the Lewisian complex, including Dearnley (1962), Myers (1970, 1971), Coward (1972, 1973) and Graham (1973);

As users of a language, we are multi-stylistic. This means that we can employ a variety of styles of speech and writing depending on the situation. For example, we use a different type of language when talking informally to friends than when we are asked to write a research report. This ability to change the style of speaking/writing is directly reflected in the genres or registers of language we produce, as can be seen in the two examples from the ‘Think about’ task. For instance, you might have noticed that the spoken dialogue (Excerpt A) uses shorter syntactic structures (so called utterances) than the academic text (Excerpt B), which employs relatively long and complex sentences. At the same time, Excerpt B is much more ‘polished’ – it lacks hesitations (*erm*), repetitions (*I'm, I'm feeling*) and false starts (*Yes, that is also a thing to ... erm, I'm*) that are typical of informal speech. In addition, there are numerous other linguistic features

that characterize informal speech such as frequent use of personal pronouns, contractions and discourse markers, while academic writing typically uses a large number of nouns, prepositions and passives.

The issue we are thus presented with in the linguistic analysis of registers is how to make sense of the large amount of functional variation in the data and how to characterize individual registers by looking at the underlying principles of systematic variation (Biber 1988; Conrad & Biber 2001). Different **registers**, which are functionally defined types of language use, employ a multitude of linguistic features to achieve the communicative goal, which is to interact successfully given the expected (social, academic etc.) conventions. **Multidimensional analysis** is a complex procedure developed by Biber (1988) that deals with a large number of linguistic variables and identifies underlying principles of functional variation by looking at how individual linguistic variables co-occur in texts. It starts with a simple observation that registers differ in multiple ways because they have different functions.

The full multidimensional analysis has four main stages: (1) Identification of relevant variables, (2) Extraction of factors from the variables, (3) Functional interpretation of factors as dimensions and (4) Placement of registers on the dimensions. However, we can also perform a simple comparison with Biber's (1988) original dimensions, skipping steps 1–3 and focusing on the comparison of the registers in our corpus with the registers identified in Biber's (1988) original study (Conrad & Biber 2001). Let's examine the four stages one by one (see also Frigolin & Hardy 2014).

Step 1: Identification of relevant variables. Before performing the statistical analysis (step 2), we need to identify a large number of linguistic variables in the texts of our corpus. The corpus needs to include different registers (e.g. informal speech, news reporting, academic writing, popular fiction etc.) and we need to identify relevant variables, that is, those that can distinguish between the registers in our corpus. For example, all the variables mentioned in Section 5.2 play an important role in register identification. Usually, several dozen (40 – over 140) variables are identified; Biber (1988) used 67 variables, as listed in Table 5.3.

A more extensive list (141 items) combining lexico-grammatical and semantic features can be found in Xiao (2009). Multidimensional analysis follows the individual text/speaker research design (see Section 1.4); Figure 5.17 is an example of a part of a dataset used for the analysis. Each row represents an individual file; columns 3–8 show the relative frequencies of Biber's (1988) variables in the files. When preparing the data for the multidimensional analysis we need to make sure that we have enough files (i.e. sampling points)

Table 5.3 *The full set of Biber's (1988) features based on Conrad & Biber (2001: 18–19)*

1. past tense	24. infinitives	47. hedges (e.g. <i>almost, maybe, sort of</i> [except as true noun])
2. perfect aspect	25. present participial adverbial clauses (e.g. <i>Screaming with rage, he ran up the stairs.</i>)	48. amplifiers (e.g. <i>completely, totally, utterly</i>)
3. present tense	26. past participial adverbial clauses (e.g. <i>Given these characteristics, it is not surprising that . . .</i>)	49. emphatics (e.g. <i>a lot, for sure, really</i>)
4. place adverbials (e.g. <i>behind, downstairs, locally</i>)	27. past participial postnominal (reduced relative) clauses (e.g. <i>the exhaust air volume required by the 6-ft. x 4-ft. grid</i>)	50. discourse particles (e.g. sentence initial <i>anyhow, now, well</i>)
5. time adverbials (e.g. <i>eventually, immediately, nowadays</i>)	28. present participial postnominal (reduced relative) clauses (e.g. <i>the currents of dissent swirling beneath the surface</i>)	51. demonstratives
6. first-person pronouns	29. <i>that</i> relative clauses in subject position (e.g. <i>the papers that are on the table</i>)	52. possibility modals (<i>can, could, may, might</i>)
7. second-person pronouns	30. <i>that</i> relative clauses in object position (e.g. <i>the papers that she thought would be interesting</i>)	53. necessity modals (<i>must, ought, should</i>)
8. third-person personal pronouns (excluding <i>it</i>)	31. <i>Wh</i> -relatives in subject position (e.g. <i>people who know him</i>)	54. predictive modals (<i>shall, will, would</i>)
9. pronoun <i>it</i>	32. <i>Wh</i> -relatives on object position (e.g. <i>people who he knows</i>)	55. public verbs (e.g. <i>complain, explain, promise</i>)
10. demonstrative pronouns (<i>that, this, these, those</i> as pronouns)	33. pied-piping relative clauses (e.g. <i>the way in which food is digested</i>)	56. private verbs (e.g. <i>believe, think, know</i>)
11. indefinite pronouns (e.g. <i>anyone, everybody, nothing</i>)	34. sentence relatives (e.g. <i>We waited for six hours, which was ridiculous.</i>)	57. suasive verbs (e.g. <i>command, propose, recommend</i>)
12. pro-verb <i>do</i>	35. causative adverbial subordinator (<i>because</i>)	58. <i>seem</i> and <i>appear</i>
13. direct Wh-questions	36. concessive adverbial subordinators (<i>although, though</i>)	59. contractions (<i>don't</i>)
14. nominalizations (ending in <i>-tion, -ment, -ness, -ity</i>)	37. conditional adverbial subordinators (<i>if, unless</i>)	60. complementizer <i>that</i> deletion (e.g. <i>I think [Ø] he's gone already.</i>)

Table 5.3 (cont.)

15. gerunds (participial forms functioning as nouns)	38. other adverbial subordinators (e.g. <i>insomuch as, such that, while</i>)	61. stranded prepositions (e.g. <i>the person that I was talking to</i>)
16. total other nouns	39. total prepositional phrases	62. split infinitives (e.g. <i>I want to completely convince you that</i>)
17. agentless passives	40. attributive adjectives (e.g. <i>the small room</i>)	63. split auxiliaries (e.g. <i>they have apparently sold it all ...</i>)
18. <i>by</i> -passives	41. predicative adjectives (e.g. <i>the room is small</i>)	64. phrasal coordination (NOUN and NOUN; ADJ and ADJ; VERB and VERB; ADV and ADV)
19. <i>be</i> as main verb	42. total adverbs	65. independent clause coordination (clause initial and)
20. existential <i>there</i>	43. type/token ratio	66. synthetic negation (e.g. <i>No evidence was found.</i>)
21. <i>that</i> verb complements (e.g. <i>We felt that we needed a financial base.</i>)	44. mean word length	67. analytic negation (e.g. <i>That's not true.</i>)
22. <i>that</i> adjective complements (e.g. <i>It's quite obvious that certain things can be sexlinked.</i>)	45. conjuncts (e.g. <i>alternatively, nevertheless, therefore</i>)	
23. <i>Wh</i> -clauses (e.g. <i>I wondered what to do.</i>)	46. downtoners (e.g. <i>mildly, partially, somewhat</i>)	

in the corpus to look at the variation meaningfully. Ideally, the number of files should be at least five times the number of linguistic variables but sometimes fewer files can be sufficient⁵ (Friginal & Hardy 2014: 303).

⁵ This is when there are a number of strong correlations between linguistic variables (>0.5) and the patterns are clear even with fewer sampling points.

Filename	Register	PAST	PERF	PRES	PLACE	TIME	1PRON
BE_A01	News_reportage	4.35	0.89	3.86	0.45	0.59	1.58
BE_A02	News_reportage	3.15	0.94	3.64	0.33	0.77	0.72
BE_A03	News_reportage	4.12	1.52	5.11	0.34	0.34	0.39
BE_A04	News_reportage	6.68	0.95	2.79	0.4	0.75	0.35
BE_A05	News_reportage	4.36	0.62	4.15	0.67	0.88	1.66
BE_A06	News_reportage	3.45	1.18	4.28	0.05	0.41	0.31
BE_A07	News_reportage	2.99	1.42	4.35	0.3	0.51	0.56
BE_A08	News_reportage	4.67	0.68	3.46	0.47	0.84	1.89
BE_A09	News_reportage	5.44	0.64	5.79	0.25	0.74	1.68
BE_A10	News_reportage	6.85	0.77	3.76	0.39	0.82	1.74
BE_A11	News_reportage	3.96	0.84	3.96	0.5	0.74	0.45
BE_A12	News_reportage	5.32	1.27	5.27	0.61	0.2	2.68
BE_A13	News_reportage	4.16	1.57	4.31	0.51	0.25	0.66
BE_A14	News_reportage	2.78	0.36	4.84	0.36	0.21	1.08

Figure 5.17 *A dataset for multidimensional analysis (a small extract)*

Step 2: Extraction of factors from multiple variables. The extraction of factors is performed using a statistical technique called factor analysis. **Factor analysis** is a complex mathematical procedure that reduces a large number of linguistic variables to a small number of factors, each combining multiple linguistic variables. This is done by considering correlations between variables (see Section 5.2 for the explanation of correlation); those that correlate – both positively and negatively – are considered components of the same factor because they have a connection. Positive correlations mean that the variables show the same pattern of occurrence in the data, while negative correlation indicates complementary distribution, that is, if one variable appears with a high frequency the other appears infrequently and vice versa. A **factor** is thus a group of related linguistic variables summarizing a more general tendency (underlying dimension – see step 3) in the data. This interconnectedness of linguistic variables that the factors capture is a direct reflection of core functional features of registers; it helps us understand how speakers/writers combine different linguistic features to adopt an appropriate style for a given situation. For instance, we know that personal pronouns, contracted forms, present tense verbs, and verbs such as *believe*, *think* and *know* ‘team up’ to express informality and speaker’s personal involvement, which is typical of registers such as fiction and informal speech.

Figure 5.18 illustrates the main principle of factor analysis with real linguistic data but fewer variables and factors than usually considered (see step 1 above).⁶ It shows ten linguistic variables combined into two factors. Factor 1 consists of private verbs (e.g. *believe*, *think*), *that* deletion, present tense verbs, contractions and second-person pronouns, while factor 2 comprises past tense verbs, third-person pronouns, perfect aspect, public verbs (e.g. *complain*, *promise*) and

⁶ The illustration is inspired by Field et al. (2012: 753ff).

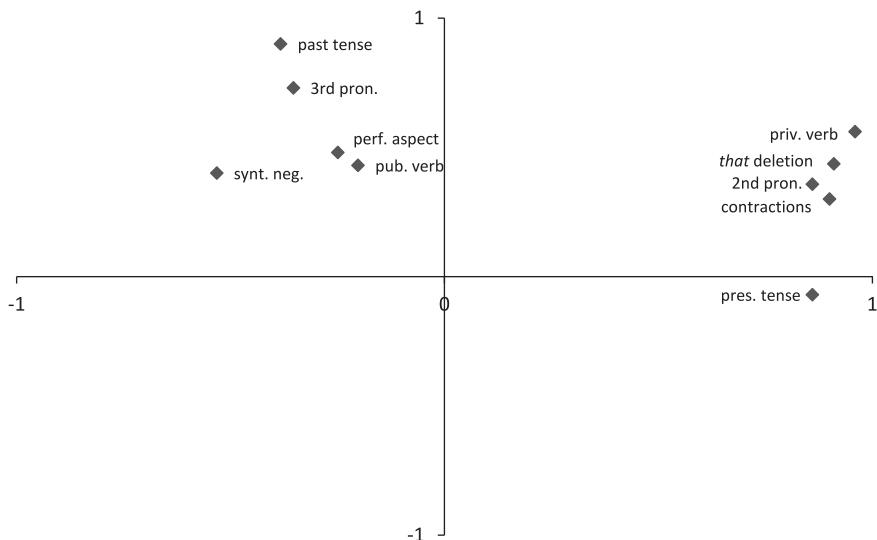


Figure 5.18 *Data reduction: ten variables into two factors*

synthetic negation. Factor 1 is represented by the horizontal line (x-axis) while factor 2 is represented by the vertical line (y-axis). The data points (diamonds) show individual linguistic variables.

Three important details need to be mentioned before we move to step 3: first, prior to carrying out factor analysis, we need to check that the individual linguistic variables correlate reasonably well (above 0.3) with some other variables in the dataset, otherwise we wouldn't be able to combine variables into factors. The checking can be done by looking at the Pearson's correlations between pairs of variables; the correlations are usually displayed in a form of a correlation matrix where each variable is correlated with the rest of the variables in the dataset.⁷ Those variables that do not correlate with other variables can be excluded from further analysis.

Second, to optimize the results of the factor analysis, the factors (represented by the two axes in Figure 5.18) are usually rotated so that they pass through the middle of the relevant variables; in this way, the factors better represent the core variables. Different kinds of rotation are possible (varimax, promax, etc.). Because we are dealing with linguistic data, we can assume that the factors are related (as everything in language is to some degree) so the appropriate rotation for this type of analysis is promax (oblique type of rotation), where there is an oblique angle between the lines representing the factors. It is, however, also possible to choose

⁷ Sometimes, in addition to looking at the correlation matrix, Bartlett's test, a formal statistical test that checks whether there is a minimal required relationship between the variables (i.e. one significantly different from zero), is performed.

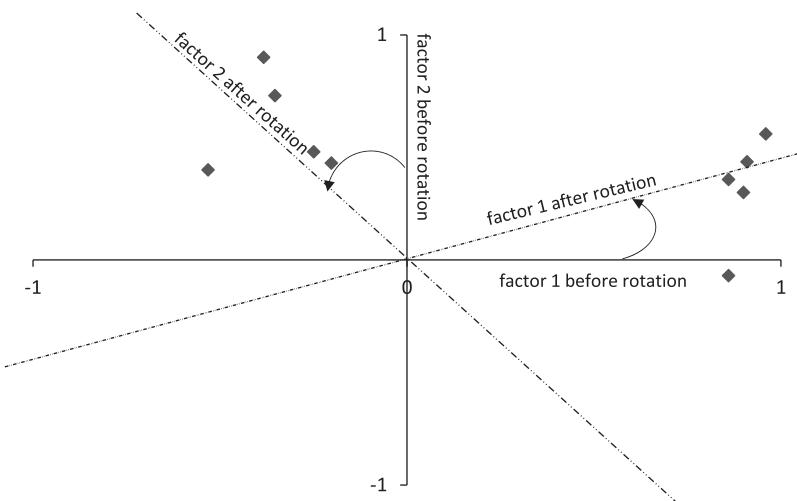


Figure 5.19 *Promax factor rotation*

varimax (perpendicular rotation), which is the default in non-linguistic types of analysis. The principle of the rotation can be seen in Figure 5.19.

Third, we have to decide how many factors we want to extract. The idea is to extract as few factors that explain as much variation in the data as possible, because the whole point of factor analysis is to reduce a large number of linguistic variables into a few underlying factors. To help us decide, a visualization technique producing a scree plot is used. A **scree plot** is a graph exemplified in Figure 5.20 which provides an indication of how many factors we should extract. It displays the number of factors (each represented by a small triangle) on the x-axis and eigenvalues on the y-axis. **Eigenvalue** is a measure of how much variation in the data a factor explains – the larger the value the better. Mathematically, eigenvalue is a sum of squared factor loadings for all variables. If a variable has high factor loading, it is an important part of the factor. So how do we know how many factors to extract?

There are different answers to this question: often factors with eigenvalues above one (Kaiser 1960) are considered for further analysis. If we apply this criterion to the scree plot in Figure 5.20 we will extract five factors. Another criterion is to look at the point of inflection in the scree plot – this is where the sharp drop in eigenvalues ends and the curve of the plot starts levelling off. In Figure 5.20 there are two clear points of inflection indicating a three- or six-factor solution. Finally, Parallel analysis (PA) is sometimes used to establish a baseline of eigenvalues obtained in computer simulation (Monte Carlo) with unrelated (random) variables. The factors extracted need to have eigenvalues clearly above this baseline (Ledesma & Valero-Mora 2007; Hayton et al. 2004). In Figure 5.20, we can see up to nine factors satisfying this criterion. More generally, we have to

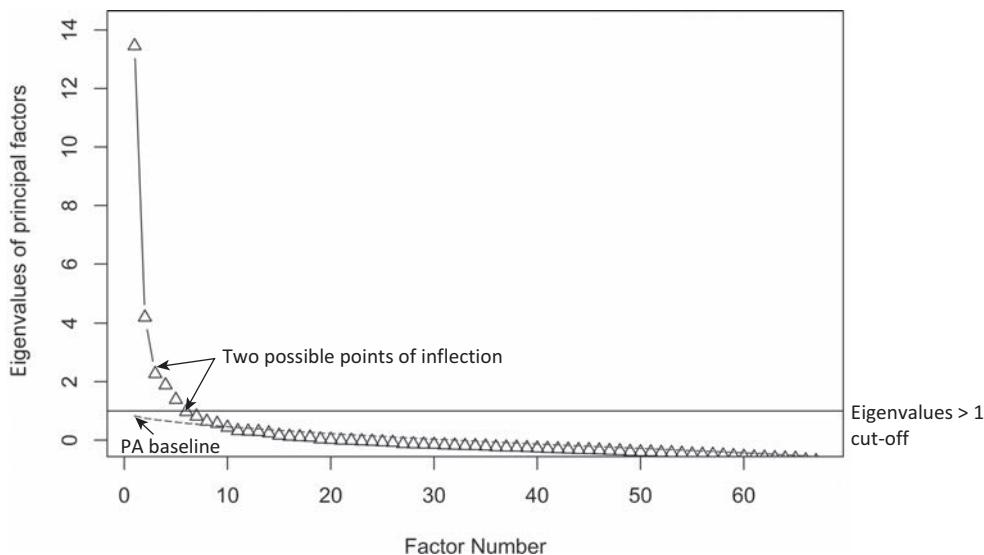


Figure 5.20 *Factor extraction: scree plot*

realize that we are dealing with a trade-off situation. If we extract more factors, we'll account for more variation in the data. On the other hand, factors with smaller eigenvalues are notoriously difficult to interpret (see step 3 below).

Step 3: Functional interpretation of factors as dimensions. This step involves functional interpretation of the extracted factors as dimensions. Here, we need to take each of the extracted factors and look at the factor loadings of individual variables that the factor analyses outputs. Factor loadings are measures of importance of the individual variables in relation to a particular factor usually ranging from -1 to $+1$.⁸ Let's demonstrate the interpretation of factors with factor 1 displayed in Table 5.4. We can see three types of variables contributing to this factor. Group A: variables with high positive loadings larger than 0.35 (1-5); group B: variables with high negative loadings larger than 0.35 in absolute terms (8-12); group C: variables with low factor loadings both positive and negative, i.e. smaller than 0.35 in absolute terms (6 and 7).

Factor 1 can be interpreted as follows: first we look at the variables with high positive loadings (Group A). These variables co-occur in texts and we can assume that they have a common communicative function. When we look at the range of variables in this group we can see that these variables mark informal, highly interactive language with frequent use of first- and second-person pronouns, contractions, adverbs and *that* deletion. Next, we look at the variables with

⁸ Note that with promax rotation, where factors are allowed to correlate, factor loadings can be greater than 1 in absolute terms.

negative factor loadings (Group B) which have a complementary distribution to the variables with positive loadings. This means that they occur infrequently in texts in which the variables from Group A occur with high frequency and vice versa. Again, variables from Group B share a communicative function, which can be labelled as ‘academic-type description’ with passives and many modified nouns and nominal forms ending in *-tion*, *-ment*, *-ness* and *-ity*. Biber (1988) interprets the first factor (based on similar findings) as a dimension of INVOLVED (Group A variables) vs INFORMATIONAL (Group B variables) production. Variables with low factor loadings (Group C) can be ignored because they don’t contribute substantially to the factor; such variables, however, usually have high loadings on other factors. In fact, as a general principle, each variable is considered only once as part of the factor on which it has the highest loading above 0.35 in absolute terms.

Table 5.4 *Factor 1: loadings of individual variables^a*

Features	Factor loadings	Variables
1. contractions	0.906	Group A
2. 1st pers. pron.	0.67	
3. 2nd pers. pron.	0.605	
4. adverbs	0.576	
5. <i>that</i> deletions	0.546	
6. possessive pronouns	0.219	Group C
7. gerunds	-0.227	
8. nominalizations	-0.408	Group B
9. attributive adjective	-0.465	
10. passives	-0.518	
11. prepositional phrases	-0.607	
12. other nouns	-0.778	

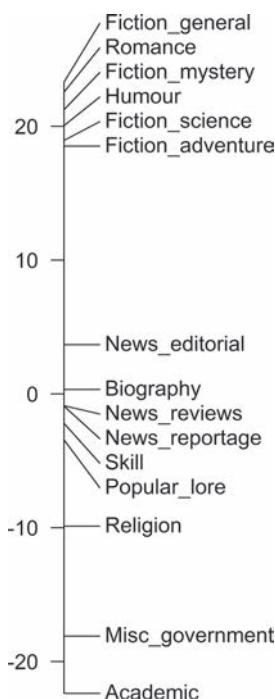
^a For illustration purposes the list of variables was shortened.

Step 4: Placement of registers on the dimensions. The final step involves placement of the registers on the dimensions. This step can also help interpret the communicative functions of the dimensions from step 3 by looking at the types of texts in which the variables that load high on each factor occur. First, we need to standardize the dataset (see Figure 5.17) by calculating z -scores₂. As we already know, z -scores₂ are standard values that use standard deviation as the unit of measurement; they are used to give each variable the same weight (influence) regardless of whether the linguistic feature is frequent or rare and whether it is measured on the same or a different scale. z -scores₂ are calculated using equation (5.9) from Section 5.3.

After this, dimension scores for each text are calculated as follows:

$$\text{Dimension score}_{\text{text}} = \text{variable 1}_{\text{Group A}} + \text{variable 2}_{\text{Group A}} \dots \\ - \text{variable 1}_{\text{Group B}} - \text{variable 2}_{\text{Group B}} \quad (5.14)$$

A dimension score is the sums of z-scores₂ for the variables with high positive loadings (Group A) minus the z-scores₂ for the variables with high negative loadings (Group B). Finally, the mean dimension score for each register is calculated by taking all text dimension scores belonging to the same register and calculating the average value. Each register is then placed on a one-dimensional scale according to the resulting mean value of the dimension score. This can be demonstrated with Figure 5.21.



One-way ANOVA: $F(14, 485) = 64.42; p <.001, r^2 = 65.03\%$

Figure 5.21 *Mean scores of registers placed on Dimension 1: Involved vs Informational*

We can see that registers with high Dimension 1 scores are different types of fiction in which a lot of interaction takes place. They are therefore placed towards the Involved end of the dimension. On the other hand, government documents and academic writing have a low dimension score clustering towards the Informational end of the dimension. The other registers are somewhere in the middle, some closer to the Involved end and others closer to the Informational end.

In addition, to test whether there is a statistically significant difference between the registers placed on Dimension 1, **one-way ANOVA** is computed (see Section 6.3 for more information); R^2 is also reported showing the amount of variation in the registers explained by this factor. The process of register placement is repeated with each of the extracted factors.

Reporting Statistics: Multidimensional Analysis

1. What to Report

Multidimensional analysis involves a number of complex steps and statistical procedures which need to be reported for full replicability of the results. The success of MD also directly depends on the reliability of the automatic identification of linguistic variables (tagging) in corpora. The following information should be reported:

- (1) Linguistic variables used: all linguistic variables should be listed together with the information on how much they overlap with Biber's (1988) original 67 variables. As a unique reference to Biber's (1988) variables, a number from the list of these variables in Table 5.3 can be used, e.g. present tense (3). However, note that MD is a universal procedure not specific to English; similar relevant variables can be identified in other languages as well.
- (2) The tagging procedure (i.e. which tagger and, where appropriate, which version of the tagset was used) and whether its reliability has been checked.
- (3) Type of multidimensional analysis performed: Full MD or Comparison with Biber's (1988) dimensions.
- (4) Factor analysis rotation (e.g. promax or varimax) and number of factors extracted.
- (5) Results: (a) factor loadings, (b) dimension plots, (c) ANOVA and r^2

The Method section of the research report describes the analytical procedure and the parameters used. In the Results and Discussion, each dimension plot needs to be carefully discussed. The main question to be addressed is: what is the underlying functional variation in registers that the method revealed?

2. How to Report: An Example

For a contextualized example see Section 5.5.

5.5 Application: Registers in New Zealand English

This micro-study is directly inspired by Richard Xiao's Multidimensional exploration of world Englishes (Xiao 2009). While Xiao looked at the features of English as used in Britain, Hong Kong, India, the Philippines and Singapore,

Table 5.5 *Registers in ICE-NZ*

Register	No. of texts
1. Spoken-dialogue-private (conversations)	100
2. Spoken-dialogue-public	80
3. Spoken-monologue-unscripted	70
4. Spoken-monologue-scripted	50
5. Written-non-printed-student writing	20
6. Written-non-printed-letters	30
7. Written-printed-academic	40
8. Written-printed-popular	40
9. Written-printed-reportage	20
10. Written-printed-instructional	20
11. Written-printed-persuasive	10
12. Written-printed-creative	20

this study focuses on New Zealand English. It is based on the ICE-NZ corpus (Vine 1999), a one-million-word corpus of both written (40%) and spoken (60%) English produced in the New Zealand context. Table 5.5 shows the distribution of the registers in this corpus.

As a starting point, Biber's (1988) 67 variables were considered. Out of these, 44 variables were selected which do not depend on the sentence structure in the corpus, to avoid problems with identifying variables in the spoken subcorpus of ICE-NZ.⁹ The following is a list of these variables (numbers in brackets refer to Biber's (1988) order):

past tense (1), perfect aspect (2), present tense (3), place adverbials (4), time adverbials (5), first-person pronouns (6), second-person pronouns (7), third-person personal pronouns (8), pronoun *it* (9), demonstrative pronouns (10), indefinite pronouns (11), pro-verb *do* (12), nominalizations (14), total other nouns (16), agentless passives (17), *by*-passives (18), *be* as main verb (19), existential *there* (20), causative adverbial subordinator (35), concessive adverbial subordinators (36), conditional adverbial subordinators (37), other adverbial subordinators (38), total prepositional phrases (39), attributive adjectives (40), predicative adjectives (41), total adverbs (42), type/token ratio (43), mean word length (44), conjuncts (45), downtoners (46), hedges (47), amplifiers (48), emphatics (49), demonstratives (51), possibility modals (52), necessity modals (53), predictive modals (54), public verbs (55), private verbs (56), suasive verbs (57), *seem* and *appear* (58), contractions (59), synthetic negation (66), analytic negation (67)

Several other linguistic variables typical of the NZ context were also considered, but eventually dropped. For example, the presence of Māori words in the

⁹ For example, variables such as sentence relatives, sentence-initial discourse particles and prepositions in the final position all depend on clear marking of sentence boundaries. These are, however, not available in the spoken subcorpus of ICE-NZ, where the unit is speaker turn.

spoken part of the corpus was investigated. However, this variable was not related to the other 44 variables; all correlations were very small, i.e. below 0.1.¹⁰

Before part-of-speech (POS) tagging, the corpus files were pre-processed and all meta-tags and comments were deleted. In addition, in the spoken subcorpus, paralinguistic sounds (*um*, *er*, *mhm*, *mm* . . .) and incomplete words were also removed to prevent POS tagging errors and inflation of particular variable counts.¹¹ Finally, in the spoken part all instances of the first-person pronoun, which, according to the transcription convention, was transcribed as a lower case *i*, were capitalized. This was again done to reduce the number of POS tagging errors because lower case *i* is systematically tagged as a foreign word by the tagger used in this study, the Stanford tagger.

MAT analyser (Nini 2015), which is a module built on the Stanford tagger, was used to identify the 44 variables in the corpus. The results of the tagging procedure were checked for any systematic tagging errors and errors in variable identification (see Leech et al. 1994, Manning 2011 and Vine 2011 for a discussion of tagging errors). The data was then processed using the Correlation, Cluster and MD analysis tools from Lancaster Stats Tools online.

When dealing with multiple linguistic variables, the first step is to look at the correlation between these variables. For this purpose, a 44×44 correlation matrix was produced (Figure 5.22) visualizing the Pearson's correlations among the 44 variables. The intensity of the shading shows the strength of the correlation. As can be seen by visual inspection, a large number of the correlations are moderate or strong, which is one of the prerequisites for successfully performing the factor analysis. In the correlation matrix, the strongest negative correlation was identified between mean word length and the use of contractions ($r = -.885$, 95% CI $[-.903, -.865]$).

This negative correlation indicates that texts in the corpus with on average longer words (as measured by the number of characters) have very few, if any, contracted forms and vice versa. This correlation makes sense linguistically – longer vocabulary items occur in formal written texts, which rarely use contracted forms. On the other hand, informal spoken registers have many contractions and shorter words. The relationship between contractions and mean word length can be visualized using a scatterplot (Figure 5.23).

Here, we can see that these two variables alone can help distinguish between different registers. In the top left corner of the two-dimensional space, cluster informal conversations. When we move down the diagonal, we can see overlapping clusters of public dialogues, unscripted monologues, letters and pieces of creative writing (left) as well as scripted monologue (right). These are

¹⁰ However, the presence of Māori words correlates with the speaker's ethnicity, i.e. whether the speaker is of Māori origin ($r = .436$, 95% CI [.34, .524]).

¹¹ Hesitation marks, *er* and *um*, are, for instance, often mis-tagged as nouns, adjectives or verbs depending on the syntactic context.

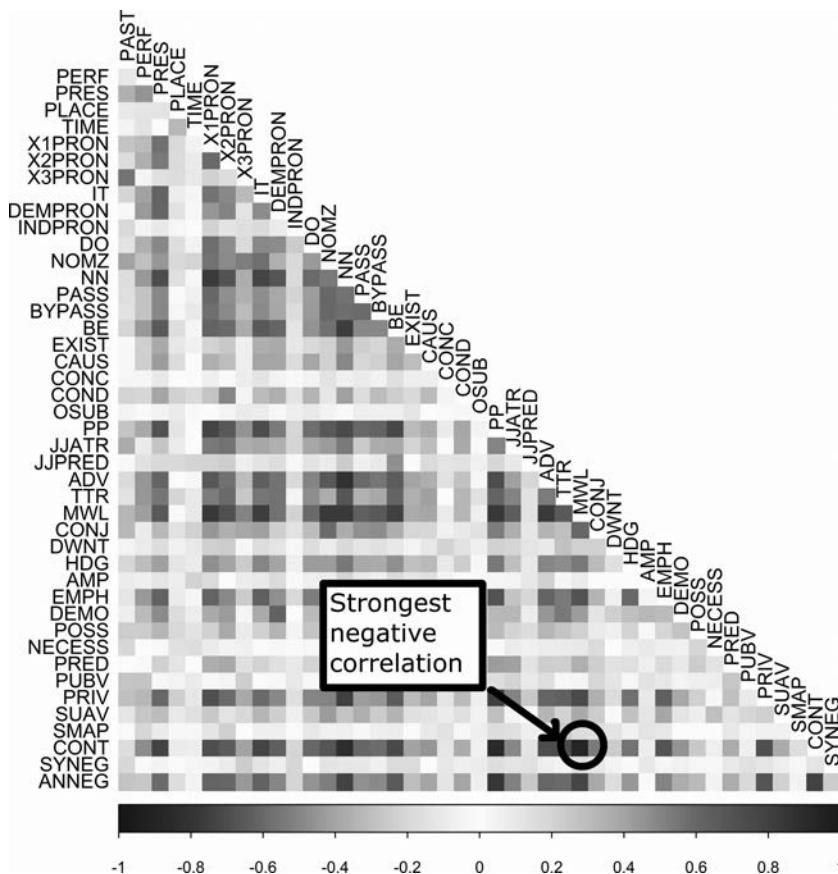


Figure 5.22 Correlation matrix: 44 variables

marked by a gradually decreasing number of contractions and increasing mean word length.

When we move beyond the two-dimensional space from Figure 5.23 and use all 44 variables, we can distinguish between the registers with even more precision. To do this, first the hierarchical agglomerative cluster technique was used¹² with the whole dataset (44 variables, no transformation, Manhattan distance, Ward's method). The resulting cluster plot is shown in Figure 5.24.

In Figure 5.24, we can see a clear split between speech and writing, with the exception of spoken scripted monologue, which clusters together with letters. The closeness of the scripted monologue to writing is to be expected because scripted speech is a specific written-to-be-spoken text type. Overall, three main clusters emerge: (1) a spoken cluster (private and public dialogue and unscripted

¹² The clusters are based on the mean values for each register, which were calculated prior to the cluster analysis.

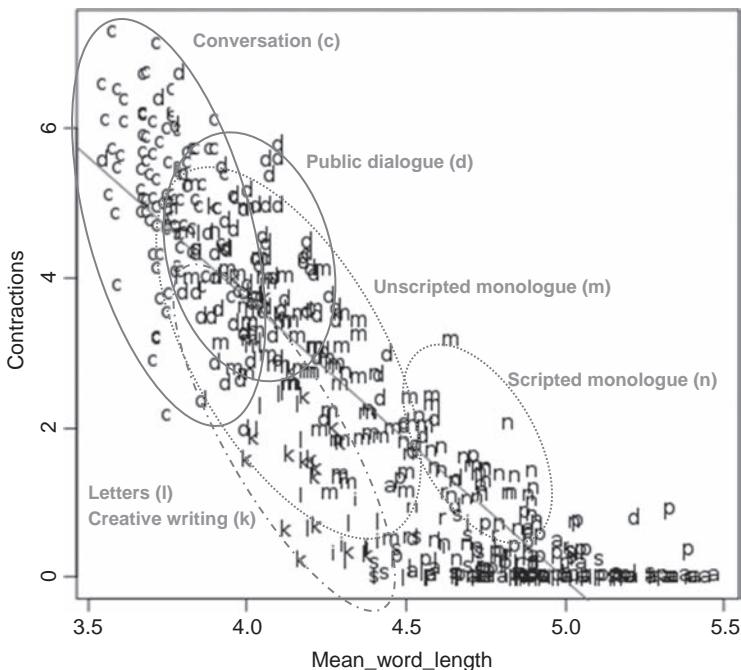


Figure 5.23 *Correlation between mean word length and contractions: register clusters*

monologue), (2) a more formal writing cluster (student, instructional, academic and reportage writing) and (3) a less formal writing + scripted monologue cluster (creative, letter, popular, persuasive writing and scripted monologue). We can also see the internal structure of these clusters by looking at the height on the y-axis of the plot which signifies the distance between the registers in a 44-dimensional space. Although the cluster technique is very powerful, it does not tease apart different groups of linguistic variables (factors) that characterize individual registers. In fact, the cluster analysis reduces a 44-dimensional space to only one dimension which combines the contributions of all the 44 variables and is displayed as the height of the tree plot.

To explore more dimensions in the data, multidimensional analysis was performed. Because of space constraints, only a subset of the results of this analysis (first two factors/dimensions) will be reported. Rather than a comparison with Biber's (1988) original dimensions (the full range of Biber's 67 variables was not available), a full multidimensional analysis was performed. Factor analysis (rotation promax) was used to extract four factors which were then interpreted as dimensions. The decision about the number of factors to be extracted was based on the point of inflexion in the scree plot. Table 5.6 shows Factors 1 and 2, including the factor loadings of the individual variables above 0.35 in absolute terms.

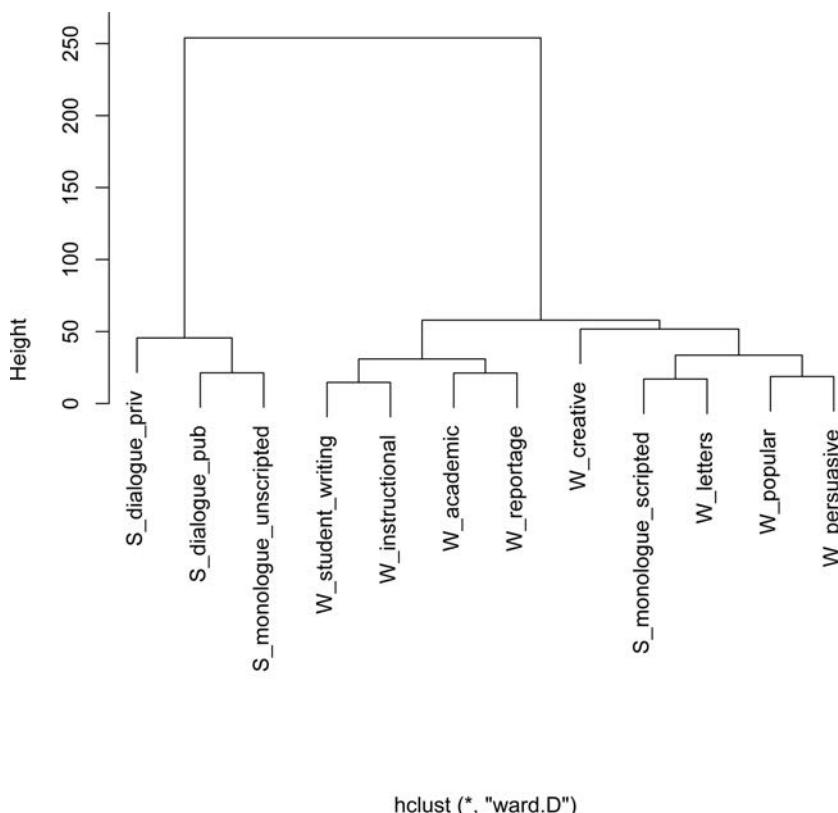


Figure 5.24 Cluster plot: registers in New Zealand English

We can see that the first factor includes more variables with relevant (above 0.35 in absolute terms) loadings than the second factor. The positive loadings on Factor 1 include present tense verbs, *be* as the main verb, contractions, adverbs, private verbs, pronouns etc. Negative loadings include heavy noun and preposition use, longer words, larger lexical richness (type/token ratio), use of passives etc. Functionally, the positive features indicate high speaker's involvement and informal style. The negative features, on the other hand, appear in more formal academic registers rich in informational content. Because the distribution of the features corresponds well with Biber's (1988) Dimension 1, Biber's label 'Involved vs informational' was retained for this dimension. Figure 5.25 shows the distribution of individual registers along Dimension 1, with the spoken registers (with the exception of scripted monologue¹³) being closer to the Involved end of the dimension than the written registers. As expected, the most involved register is spoken private dialogue (conversations) and the most informational register is newspaper reporting together with academic writing. It is also worth noting that contractions and mean word length,

¹³ Note that scripted monologue also clusters with written registers in the tree plot in Figure 5.24.

Table 5.6 *Results of factor analysis of NZ English: factor loadings*

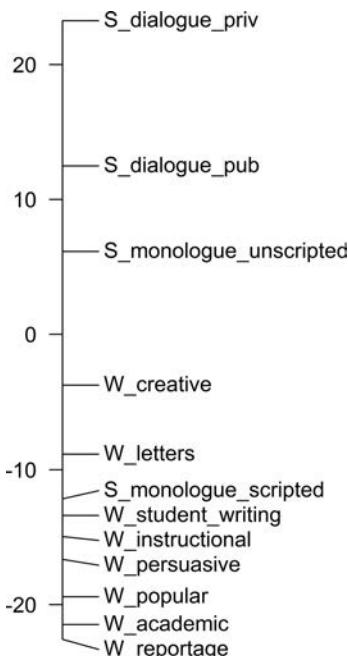
Features	Factor 1 loadings	Features	Factor 2 loadings
present tense (3) ^a	0.879	predictive modals (54)	0.75
<i>be</i> as main verb (19)	0.848	conditional adv. subordinators (37)	0.7
contractions (59)	0.846	suasive verbs (57)	0.57
total adverbs (42)	0.802	possibility modals (52)	0.56
emphatics (49)	0.798	necessity modals (53)	0.53
pronoun <i>it</i> (9)	0.746	[second-person pronouns (7)]	0.358]
demonstrative pronouns (10)	0.739	attributive adjectives (40)	-0.453
private verbs (56)	0.713		
first-person pronouns (6)	0.67		
analytic negation (67)	0.631		
demonstratives (51)	0.629		
hedges (47)	0.596		
pro-verb <i>do</i> (12)	0.572		
second-person pronouns (7)	0.537		
causative adv. subordinator (35)	0.511		
existential <i>there</i> (20)	0.472		
<i>by</i> -passives (18)	-0.492		
nominalizations (14)	-0.512		
suasive verbs (57)	-0.519		
agentless passives (17)	-0.558		
perfect aspect (2)	-0.632		
type/token ratio (43)	-0.692		
total prepositional phrases (39)	-0.693		
mean word length (44)	-0.74		
total other nouns (16)	-0.875		

^a The numbers in brackets refer to the order of presentation of Biber's (1988) original variables on which this list is based (see Table 5.3).

two variables with the strongest negative correlation considered above (see Figure 5.22), have very high loadings on Dimension 1. This means that these two variables can effectively distinguish between the individual registers. In very simple terms, Figure 5.25 is thus an extension of the simple scatterplot in Figure 5.23.

Looking at Factor 2, the largest positive loadings include different types of modals, conditional adverbials *if* and *unless* as well as verbs such as *command*, *propose* and *recommend*. Second-person pronouns were disregarded because they have a higher loading on Factor 1. The only negative loading is attributive adjectives. The underlining function of the positive features is modalized production typical of written instructional texts, while the other end of the dimension is descriptive production typical of academic writing and popular informational texts. We can therefore use the label 'Modalized vs descriptive' production. Note that this

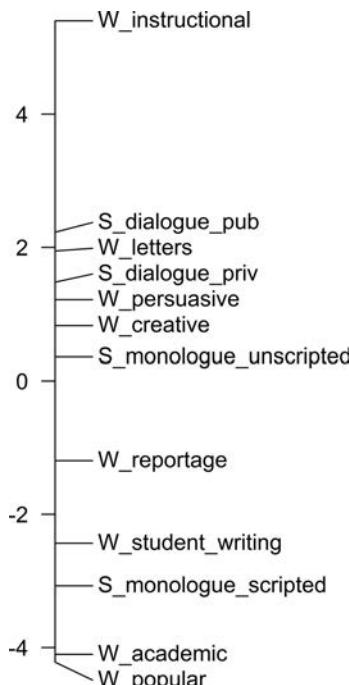
Involved vs informational



$F(11, 488) = 211.67; p < .001$,
 $r^2 = 82.67\%$

Figure 5.25 Dimension 1: New Zealand English – full MD analysis

Modalized vs descriptive



$F(11, 488) = 27.27; p < .001, r_2 = 38.07\%$

Figure 5.26 Dimension 2: New Zealand English – full MD analysis

dimension is completely different from Biber's (1988) Dimension 2: Narrative vs non-narrative discourse. Biber's narrative/non-narrative features are, however, salient also in New Zealand English; they appear in Factor 3 (not discussed here).

Finally, it is worth mentioning that both dimensions show statistically significant differences among the registers as established by one-way ANOVA. However, in the case of the Dimension 1 almost 83% of the variation in the dimension scores of the individual texts is explained by their register membership, whereas in the case of Dimension 2 this number is less than 40%. Dimension 1 is thus a more powerful predictor of register variation than Dimension 2.

5.6 Exercises

- Manually calculate the Pearson's and Spearman's correlations between verbs and adjectives in ten randomly selected texts from BE06. The data is provided below:

Verbs	169.9	135.0	161.7	183.0	163.1	190.8	140.7	213.9	218.0	165.2
Adjectives	96.0	102.6	91.9	76.5	98.8	77.6	68.4	60.3	74.4	76.5

- What can you tell about the relationship between the variables in the four graphs in Figures 5.27–5.30?

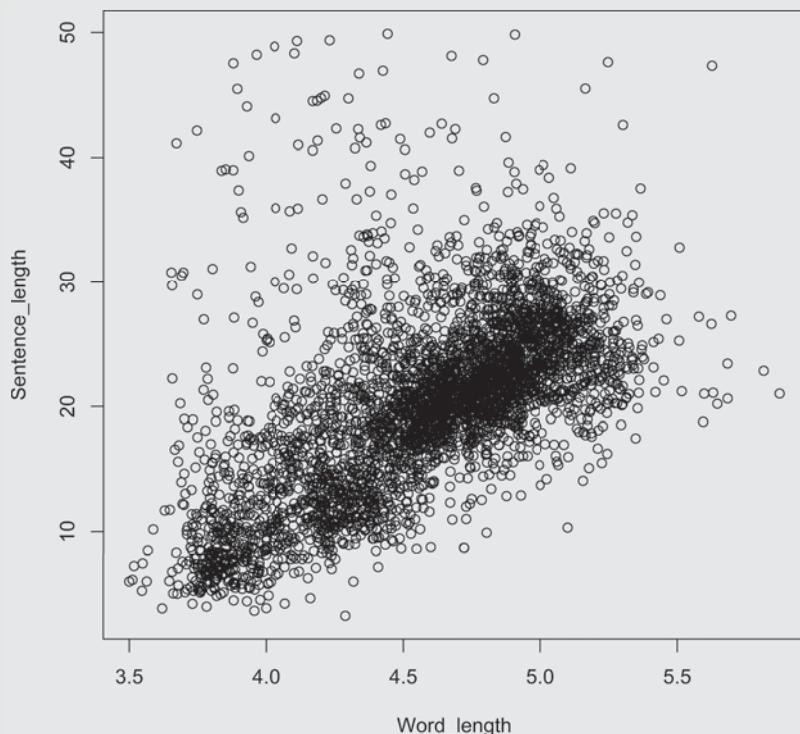


Figure 5.27 *Relationship between mean word length (number of characters) and mean sentence length (number of words) in BNC*

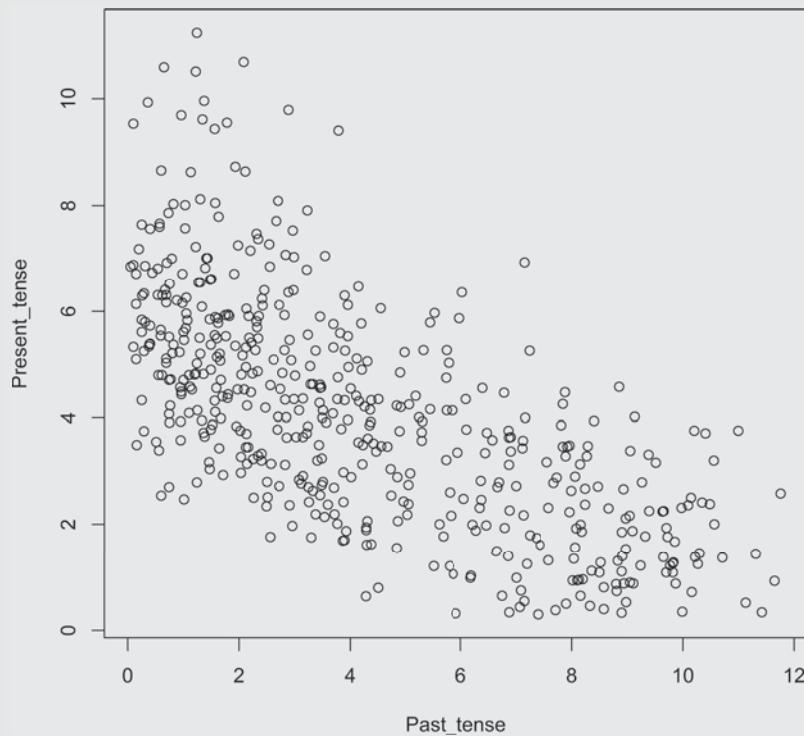


Figure 5.28 *Relationship between the use of the past and the present tense in BE06*

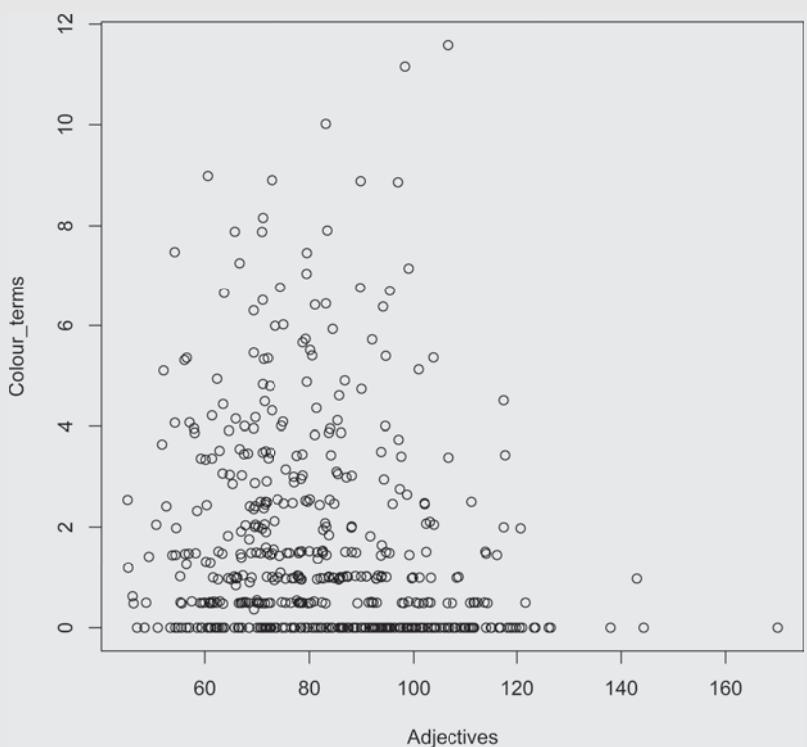


Figure 5.29 *Relationship between the use of adjectives and colour terms in BE06*

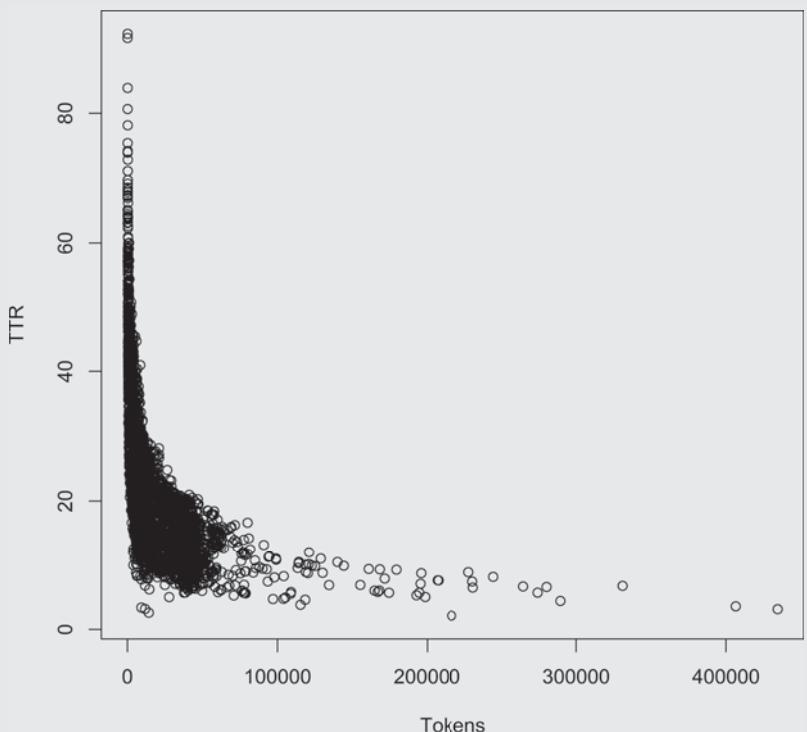


Figure 5.30 *Relationship between text length (tokens) and type–token ratio (TTR) in BNC*

Downloaded from <https://www.cambridge.org/core>. University of Warwick, on 18 Sep 2018 at 00:11:39, subject to the Cambridge Core terms of use, available at <https://www.cambridge.org/core/terms>. <https://doi.org/10.1017/9781316410899.006>

3. Each Brown family corpus is divided into 15 different types of texts listed below (see also Section 1.4).

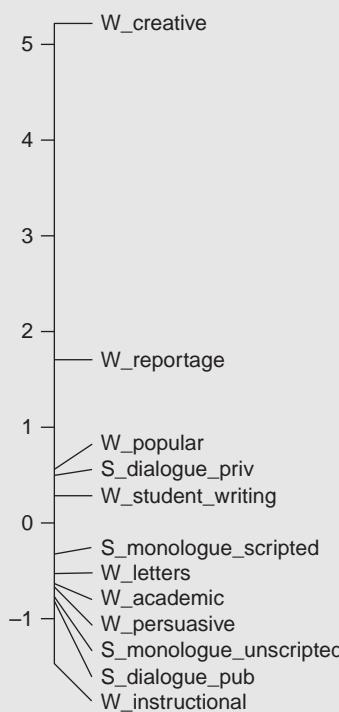
A (Press: reportage), **B** (Press: editorial), **C** (Press: reviews), **D** (Religion), **E** (Skills, trades and hobbies), **F** (Popular lore), **G** (Belles lettres, biography, essays), **H** (Miscellaneous government documents, foundation reports, industry reports, college catalogue, industry house organ), **J** (Learned and scientific writings), **K** (General fiction), **L** (Mystery and detective fiction), **M** (Science fiction), **N** (Adventure and western fiction), **P** (Romance and love story), **R** (Humour).

This classification is very useful; however, for some purposes it might be too detailed. Group the individual text types into larger categories based on their functional similarity. Then design a study in which you could verify your grouping.

4. Table 5.7 presents factor loadings of Factors 3 and 4 based on the multidimensional analysis of New Zealand English from Section 5.5. The dimension plots are also provided (Figures 5.31 and 5.32). Interpret each factor functionally as a dimension. Create labels for these dimensions.

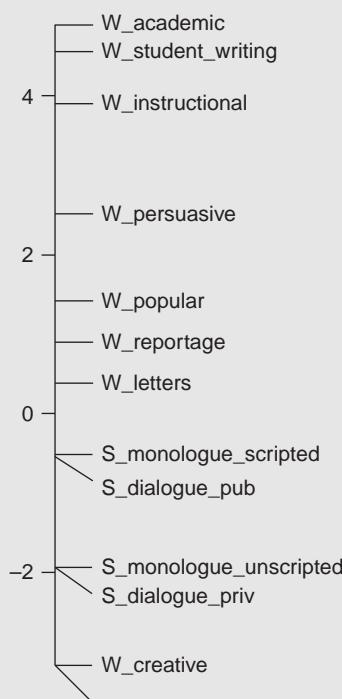
Table 5.7 *Results of factor analysis of NZ English: factor loadings of Factors 3 and 4*

Features	Factor 3 loadings	Features	Factor 4 loadings
past tense (1)	1.099	nominalizations (14)	0.618
third-person personal pronouns (8)	0.461	conjuncts (45)	0.499
attributive adjectives (40)	-0.304	agentless passives (17)	0.36
present tense (3)	-0.583	by-passives (18)	0.347
		time adverbials (5)	-0.444
		place adverbials (4)	-0.504



$F(11, 488) = 25.25; p <.001, r^2 = 36.27\%$

Figure 5.31 Dimension 3



$F(11, 488) = 44.69; p <.001, r^2 = 50.18\%$

Figure 5.32 Dimension 4

5. Use the data provided on the companion website and the MD tool to compare registers in current British and American English.

THINGS TO REMEMBER

- Correlations are used for the investigation of the relationship between two variables at a time.
- Pearson's correlation is suitable for scale variables, while Spearman's correlation assumes ordinal variables (ranks). Spearman's correlation can also be used with scale variables if the means as the measures of central tendency do not represent the data well (extremely skewed distributions).
- Hierarchical agglomerative cluster analysis is used for classification of words, texts, registers etc. The result of this analysis is a tree plot (dendrogram).
- The most complex type of analysis out of the three discussed in this chapter is multidimensional analysis (MD). MD analyses a large number of linguistic variables and reduces them to a small number of factors which are interpreted as dimensions of variation. Along these dimensions, different registers can be placed.

Advanced Reading

- Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.
- (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, 8(1), 9–37.
- Biber, D., Conrad, S., Reppen, R., Byrd, P. & Helt, M. (2002). Speaking and writing in the university: a multidimensional comparison. *TESOL Quarterly*, 36(1), 9–48.
- Chen, P. Y. & Popovich, P. M. (2002). *Correlation: parametric and nonparametric measures*. Thousand Oaks, CA: Sage.
- Conrad, S. & Biber, D. (2001) Multidimensional methodology and the dimensions of register variation in English. In S. Conrad & D. Biber (eds.), *Variation in English: multidimensional studies*, pp. 18–19. Harlow: Pearson Education.
- Divjak, D. & Gries, S. Th. (2006). Ways of trying in Russian: clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory*, 2(1), 23–60.
- Romesburg, C. (2004[1984]). *Cluster analysis for researchers*. Raleigh, NC: Lulu Press.
- Sardinha, T. B. & Pinto, M. V. (eds.) (2014). *Multi-dimensional analysis, 25 years on: a tribute to Douglas Biber*. Amsterdam: John Benjamins.
- Sheskin, D. J. (2007). *Handbook of parametric and nonparametric statistical procedures*. Boca Raton, FL: Chapman & Hall/CRC.
- Xiao, R. (2009). Multidimensional analysis and the study of world Englishes. *World Englishes*, 28(4), 421–50.

Companion Website: Lancaster Stats Tools Online

1. All analyses shown in this chapter can be reproduced using the statistical tools available from the companion website. In particular the available tools include:
 - Correlation Calculator
 - Clusters
 - MD analysis
2. The website also offers additional materials for students and teachers.

6 Sociolinguistics and Stylistics

Individual and Social Variation

6.1 What Is This Chapter About?

This chapter discusses different statistical procedures available for the analysis of stylistic and sociolinguistic variation in corpora. It reviews different approaches to variation, pointing out the common connection to the notion of ‘style’ understood as a particular way of speaking and using language. The statistics discussed include the t-test, ANOVA, the Mann–Whitney *U* test, the Kruskal–Wallis test, correspondence analysis and mixed-effects models. Each of these focuses on a different type of sociolinguistic analysis and requires a specific research design: for group comparisons, the t-test, ANOVA, the Mann–Whitney *U* test and the Kruskal–Wallis test are used. Individual linguistic style can be explored using correspondence analysis. Traditional (Labovian) sociolinguistic analysis with a focus on variation in a carefully defined linguistic context can be carried out using mixed-effects models, a technique that is still fairly new in corpus-based sociolinguistics.

We’ll be exploring answers to five questions:

- How can we approach and operationalize (sociolinguistic) variation (Section 6.2)?
- How can we compare groups of speakers based on different social variables such as gender or social class (Section 6.3)?
- How can we summarize multiple linguistic variables and create speaker profiles (Section 6.4)?
- How can we explore Labovian sociolinguistic variables (Section 6.5)?
- How can the techniques discussed in this chapter be used in research (Section 6.6)?

6.2 Individual Style and Social Variation: Where Does a Sociolinguistic Variable Start?

Think about ...

Before reading this section, look at the four samples of speech from four speakers, two real speakers and two fictional characters. Based on what you read, can you guess their gender, approximate age and other social characteristics? We will consider these examples shortly.

SPEAKER 1

He's a fucker, I can't stick him he's the
 most snobbish little cunt I've ever
 known I'd like to see what stuff he had
 if he had to pay for it himself
 hard luck, no I'd like to see what he'd do
 with his stuff if he had to pay with it, pay
 for it himself.
 Cos he's well out of order, when you, were
you there when he was
 were you there when he was like slamming
 his nine iron into one of those brick
 posts, that's sad
 he's a twat

SPEAKER 2

What boy?
 What in my school?
 Nick?
 I went out with him.
 I dumped him.
 Yeah.
 I dumped him on Wednesday.
I said hey, Mac!
 Not this Wednesday, last Wednesday I
 was saying, I just said I don't think we
 should go out no more.
He said alright.
I said bye.
I said I hope we can still be friends.
He said yeah.
I said alright I'll meet you for school,
 bye.
 <laugh>I went erm I hope we can still be
 friends.

SPEAKER 3

Come hither Nurse. What is yond
 gentleman?
 ...
 Go ask his name. If he be married,
My grave is like to be my wedding bed.
 My only love sprung from my only hate.

SPEAKER 4

And fire-ey'd fury be my conduct now!
 Now, Tybalt, take the villain back again
 That late thou gav'st me, for Mercutio's
 soul
 Is but a little way above our heads,
 Staying for thine to keep him company.
 Either thou, or I, or both, must go with
 him.
 This shall determine that.
 O, I am fortune's fool.

The notion of **style** is central to the analyses described in this chapter. Following Coupland's (2007: 2) broad definition of style as 'ways of speaking that are indexically linked¹ to social groups, times and places', we will be looking at the role of speaker background and speech community in the language that speakers produce. Style is a unifying notion linking sociolinguistics (social style), stylistics (literary style) and forensic linguistics (individual style). Regardless of whether we are looking at naturally occurring data or fiction, the statistical procedures discussed in this chapter will help us quantify and make sense of

¹ i.e. are pointing to.

variation in speaking/writing style. Linguistic variables involved in this variation show systematic differences according to both individual speakers (distinguishing **individual styles**) and groups (distinguishing social dialects or **sociolects**). But how can we identify such variables?

In variationist sociolinguistics as pioneered by Labov (e.g. 1966, 1972), a sociolinguistic variable is defined as ‘different ways of saying the same thing’ (Labov 2010: 368). For example, looking at Speaker 1 from the ‘Think about’ task, assuming we had access to a phonemic transcription or that we could hear the original recording, we could make an educated guess about which part of Britain the speaker comes from by considering the pronunciation of ‘u’ in words such as *luck*, *ucker* and *cunt*, which could be generally pronounced as /ʊ/ (some northern dialects in England) or /ʌ/. Note that we are considering the same words with the same meaning; only the (phonological) form is different. Another example of this type of variable is the use of double negative as in Speaker 2’s *I don’t think we should go out no more*, which could have been uttered as a single negative *I don’t think we should go out anymore* without any change of meaning. Because the propositional meaning of these two variants is the same, we can focus on the social meaning, that is, the social implications of the use of one form over the other. Similarly, when we extend the variation to historical data (see Chapter 7 for analysis of language change), *you* (used by Speaker 1) and *thou* (used by Speaker 4), both referring to the second person singular, satisfy Labov’s definition of a sociolinguistic variable. To define a **Labovian sociolinguistic variable**, we need to carefully **circumscribe the variable context**, in other words, define the **envelope of variation** (see Section 4.2). This means that we need to find all places in the text where variation between two or more variants (linguistic forms) is possible while the meaning remains the same (Tagliamonte 2006). A sociolinguistic variable of this type is a closed category which has several, typically two, levels (variants). For more discussion, see Section 6.5.

Although this *formal* definition of a sociolinguistic variable gives us complete control over the linguistic and social processes behind variation, it reaches its limits fairly soon when we look at other types of systematic variation beyond phonology and simple grammar (Lavandera 1978). For example, Speaker 1 uses a number of strong swearwords, which are socially stigmatised linguistic features. They therefore clearly carry social meaning. Yet swearwords do not satisfy Labov’s definition of a sociolinguistic variable – it is, for example, not clear what the words *ucker* and *cunt* are replaced with in more polite speech. Moreover, many swearwords are used as intensifiers that can appear anywhere in the discourse (see the discussion on **ambient variables** in Section 4.2).

In contrast to Labov’s formal approach, Biber (e.g. Biber & Conrad 2009) examines functional variation, where the analytical focus is on speakers’/writers’ choices ‘from the entire inventory of lexico-grammatical characteristics of a language, selecting the linguistic features that are best suited functionally to their situations and communicative purposes’ (Biber & Conrad 2009: 265). For example, we can observe a number of reporting structures (*I said, he said, I went*) in the sample

from Speaker 2. They clearly mark the style of speaking of this speaker, although there is no clear meaning-preserving competition between linguistic variants that could be easily analysed in Labovian terms. Similarly, the metaphorical language in Speaker 3 (*My grave is like to be my wedding bed*) unquestionably has implications for the particular speaker style, but can only be analysed functionally. It is important to realize that the functional approach, although it allows us to investigate a larger variety of features than Labov's formal approach, is not always straightforward and requires critical reflection on the type and sources of evidence in the data (corpora). This approach relies on direct comparability of the (sub)corpora we use; it assumes that the corpora that we compare are similar in all important aspects except the variables that we control for and focus on such as gender, age and register. The functional approach thus delegates the issues which Labov's approach deals with at the level of variable definition/selection to corpus sampling. If, for instance, the corpora we compare are biased due to over-/underrepresentation of a particular topic, the validity of such comparisons might be at stake.

Finally, to provide a full answer to the question in the 'Think about' task (many individual aspects of the language have already been discussed): Speakers 1 and 2 are real people taken from the *British National Corpus* (BNC). Speaker 1 is a 17-year-old male student from the Midlands from a working-class background. Speaker 2 is a 14-year-old female student from London also from one of the lower social classes. Speaker 3 (Juliet) and Speaker 4 (Romeo) are fictional characters from Shakespeare's famous play. They are of comparable age and gender with Speakers 1 and 2, but differ in the historical variety of English they speak, social class (Romeo and Juliet come from wealthy families in Shakespeare's imaginary Verona) and, most importantly, the fact that Romeo and Juliet are products of Shakespeare's imagination, not real persons. Interestingly, the speakers deal with somewhat parallel topics: anger and aggression (Speakers 1 and 4) and dating and courtship (Speakers 2 and 3).

6.3 Group Comparison: T-Test, ANOVA, Mann–Whitney U Test, Kruskal–Wallis Test

Think about ...

Before reading this section, think about the following question: Can you tell the gender of Speaker 1 from the following short conversation?

SPEAKER 1: So, she says, oh it don't matter. She says er <pause> will you help me fill them in now?

I says, can I ask you a question? She says what? I says, these forms things, I says <pause> d'ya know how to fill them in? She says

SPEAKER 2: Ah.

SPEAKER 1: No. She said, somebody's always done it for me.

SPEAKER 2: Annette's done them.

The conversation above comes from BNC64 (Brezina 2013), a 1.5-million-word sample of informal speech extracted from the *British National Corpus*. It is remarkable for a number of reasons, especially the frequent pronoun use by Speaker 1. The frequency of personal pronouns as a (socio)linguistic variable deserves some attention. Can it help distinguish the gender of the speaker? Let's hypothesize that personal pronoun use is a sociolinguistic marker and test this hypothesis statistically.

BNC64, which consists of spoken samples from 32 female and 32 male speakers, allows us to easily trace the linguistic performance of individual speakers² and provides a large amount of evidence for each speaker in terms of the words that can be analysed. Before we start with the analysis, it is important to note that in this section, we are drawing on the functional approach to variation discussed in Section 6.2. Personal pronouns cannot be construed as Labovian sociolinguistic variables as we are not dealing with a strict competition between two (or more) forms that express the same propositional meaning. Instead, we are interested in the comparison of the style that uses many personal pronouns as demonstrated in the ‘Think about’ task with other styles of speaking with fewer pronouns. Instead of assuming constant meaning in linguistic variables (which we can’t) we assume that we are working with a balanced corpus which samples the same type of interaction (informal speech in this case) from every speaker and that, in principle, the speakers have an equal opportunity to use personal pronouns in their speech.

So which statistical test should we use in this situation? The simplest option we have is the **t-test**, and in this case we'll use a version of the t-test that is called **Welch's independent samples t-test**. This test compares two groups of speakers (e.g. male and female speakers). The t-test compares the mean values of the linguistic variable (in our example, the relative frequency of personal pronouns in individual speaker samples) and takes into consideration the internal variation in each group expressed as variance. **Variance (S^2 or SD^2)** is the sum of squared distances of individual values from the group mean divided by the degrees of freedom. In fact, it is the squared version of sample standard deviation (see Section 2.4), hence SD^2 .

Variance is calculated according to the following formula:

$$\text{Variance} = \frac{\text{sum of squared distances from the mean}}{\text{degrees of freedom}} \quad (6.1)$$

² The group of statistical methods discussed in this section employ the ‘Individual text/speaker’ research design; they trace the occurrence of linguistic variables in the speech/writing of individual speakers/writers. In written language, this is made easier by the fact that texts can usually be taken as observations because they are often (but not always – think of multi-author academic papers) written by a single author. In dialogic speech, on the other hand, we need to carefully separate the turns by different speakers.

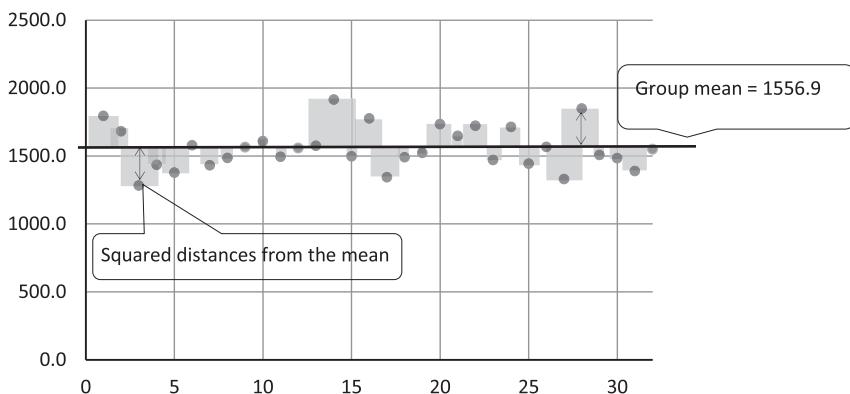


Figure 6.1 *Distribution of personal pronouns in BNC64 female speakers*

Figure 6.1 illustrates how variance is computed. It displays all 32 speakers from the female subcorpus of BNC64 and their use (relative frequency per 10k words) of personal pronouns. Each circle is an individual speaker, the thick horizontal line is the mean for the group (see Section 1.2 for explanation of the mean) and the shaded area represents the squared distances of individual values from the mean.

The logic of computing variance is simple: to see how much overall variation there is, we add the squared distances from the mean (we use squared rather than simple distances because distances below the mean are negative and would cancel out the positive distances) and divide this by the degrees of freedom. The degrees of freedom is a complex concept, which is used when dealing with calculations based on a sample (corpus) rather than a population; in corpus linguistics, this is the case most of the time. **The degrees of freedom (*df*)** signifies a number of independent ('free') components in our calculation, i.e. components that are not predictable from the previous components. In practice, it is the number of cases (texts/speakers) or groups (when looking at group variance) minus one. We subtract one from the number of cases (groups) because the last case is always predictable from the previous cases.³ For example, when we deal with 32 speakers, the degrees of freedom will be 31 (32–1).

When applied to our example, we'll get:

³ For successful use of the statistical techniques discussed here, understanding the mathematical details of the notion of 'degrees of freedom' is not crucial; what is important is to know that *df* figures in the equations for these statistical measures and is routinely reported in research reports (see the 'Reporting statistics' box at the end of this chapter). As a metaphor, imagine a classroom with 15 seats where students look for a place to sit. Only 14 (15–1) of the students have some 'freedom' to choose a place; the last student to come will have to sit on the last empty seat.

$$\text{Variance(females)} = \frac{(1794.6 - 1556.9)^2 + (1681.3 - 1556.9)^2 + \dots (1550.0 - 1556.9)^2}{32 - 1}$$

$$= 23820.5 \quad (6.2)$$

Let's return to the t-test, which uses variance as one of its components. Like every statistical test, the t-test has assumptions, which need to be reviewed before running the test. The most important assumption is the independence of observations. In the sociolinguistic context, the **independence of observations** means that each observation (text or speech sample) comes from a different (randomly sampled) speaker⁴ and that the use of language by one speaker in the sample is not affected by the use of language by another speaker. We therefore need to make sure that we include each speaker as an observation (case) only once even if multiple texts/transcripts are available, otherwise we are counting the same person twice/multiple times and violating the first part of the assumption. The second part of the assumption is easy to maintain in written texts or monologues, but in a conversation, this assumption will be violated to some extent if speakers interact with each other (something known as a 'priming effect' in psycholinguistics and more generally as 'accommodation' in sociolinguistics). We have to be mindful of this effect and evaluate its impact in each individual case. Two other assumptions listed in textbooks (and often misunderstood) are the normal distribution and homoscedasticity. **Normal distribution** of the linguistic variable in the population is an assumed symmetrical distribution visualised as a bell-shaped curve (see Section 1.3). **Homoscedasticity**⁵ is a technical term for the equality of variances, i.e. amount of variation, in two groups that we want to compare. As shown in the literature (Boneau 1960; Lumley et al. 2002), the t-test is robust and can therefore be used even with very skewed (not normally distributed) samples. Moreover, the version of the t-test known as the Welch's independent sample t-test also compensates for unequal variances (different amounts of variation in each of the compared groups) by readjusting the degrees of freedom (this is a technical detail, which you don't have to worry about). So violating these two assumptions is rarely a problem.

Welch's independent sample t-test is calculated according to the following formula:

$$\text{Welch's independent sample t-test} = \frac{\text{Mean of group1} - \text{Mean of group2}}{\sqrt{\frac{\text{Variance of group1}}{\text{Number of cases in group1}} + \frac{\text{Variance of group2}}{\text{Number of cases in group2}}}} \quad (6.3)$$

⁴ If we have two or more samples from each speaker and are interested in the difference in their language between sample 1 and sample 2 etc. (e.g. linguistic change/development), we are dealing with a so-called repeated measures design, which requires a different version of the statistical test (see below for more information).

⁵ Homoscedasticity is a bit of a tongue twister; you can use the more descriptive term 'equality of variances' instead.

As can be seen from the equation, there are three factors that have an effect on whether the test will be significant: (i) size of the mean difference, (ii) variance in each of the two groups and (iii) sample size (number of cases, i.e. speakers or texts, in both groups). The t-test value is large (and the test is significant) if there is a large difference between the means, small variance in the groups and a large number of cases; these factors combined show that there is enough evidence in the data that the two groups are different with respect to the use of the linguistic variable in question.

In our example, the means of relative frequencies (per 10 k) of personal pronouns for the male and the female subcorpora are 1,451.8 and 1,556.9 respectively, the variances are 22,256.5 and 23,820.5 and we have 32 speakers in each group (see the description of BNC64 above). When we enter this in the equation, we get:

$$\text{Welch's independent sample t-test} = \frac{1451.8 - 1556.9}{\sqrt{\frac{22,256.5}{32} + \frac{23,820.5}{32}}} = \frac{105.1}{\sqrt{744.4 + 695.5}} = 2.8 \quad (6.4)$$

In this case, the t-test statistic (2.8), with the appropriate degrees of freedom (61.93) calculated using an equation that will not be introduced here,⁶ is statistically significant: the p-value (calculated automatically by statistical packages) is 0.007, which is smaller than the 0.05 (the usual 5% alpha level⁷) and hence we can say that there is enough evidence in the corpus to reject the null hypothesis (which says that there is no difference between the two gender groups); we thus conclude that there is a statistically significant difference between male and female speakers.

In addition to the statistical test, we also need to calculate an **effect size measure** to evaluate in standardized terms (i.e. units comparable across linguistic variables and corpora) the size of the difference between the two groups. With the t-test, we have several options of effect size measures that include Cohen's *d* and *r* as two typically used effect size measures. **Cohen's *d*** is calculated as the difference between the two means expressed in standard deviation units. Cohen (1988: 40) recommends the following standard interpretation of the *d* measure (see also Section 8.4):

Interpretation of *d*: *d* > 0.3 small, *d* > 0.5 medium and *d* > 0.8 large effect

This is regardless of the plus or minus sign in front of the value. The plus or minus sign indicates whether the first group considered in the comparison has a

⁶ There is a (slightly complicated) equation that can be used for calculating the degrees of freedom and hence establishing the cut-off point of the t-test statistic for significance. This cut-off point would traditionally be listed in statistical tables under the appropriate degrees of freedom. In modern statistical packages, this step can be skipped because degrees of freedom as well as the p-value (statistical significance) are calculated automatically.

⁷ 'Alpha level' is the statistical term for the level of statistical significance that we choose. In social science, this is usually .05 or 5%.

larger mean value (+) or a smaller mean value (−) than the second group, so this is an indicator of the **direction of the effect**. The interpretation of Cohen's d values can vary according to a specific discipline: something that is considered a large effect in physics might not be such a large effect in language studies or vice versa (see Section 9.2 for more discussion). Cohen's d is calculated as follows:

$$\text{Cohen's } d = \frac{\text{Mean of group1} - \text{Mean of group2}}{\text{pooled } SD} \quad (6.5)$$

where

$$\text{pooled } SD = \sqrt{\frac{SD1^2 \times (\text{cases in group1} - 1) + SD2^2 \times (\text{cases in group2} - 1)}{\text{all cases} - 2}}$$

In our example, the calculation of Cohen's d is as follows:

$$\text{Cohen's } d = \frac{1451.8 - 1556.9}{\sqrt{\frac{22,256.5 \times (32 - 1) + 23,820.5 \times (32 - 1)}{64 - 2}}} = \frac{1451.8 - 1556.9}{151.8} = 0.69 \quad (6.6)$$

This can be interpreted as a medium effect.

As with any other effect size measure, we also need to look at the 95% confidence interval for Cohen's d , which, in our example, is 0.18 to 1.21, as calculated automatically by statistical packages such as Lancaster Stats Tools online. This shows a likely range of the effect in the population (all male and female speakers of British English). Because this 95% CI is extremely wide ranging from a minimum to a large effect, we cannot be sure about the actual size of the effect in the population; this is due to a relatively small sample size (64 speakers). A small aside: for extremely skewed data (i.e. data severely violating the normality assumption) a robust version of Cohen's d has been proposed (Erceg-Hurn & Mirosevich 2008). This version uses the same equation as the standard version (equation 6.5), but the means and standard deviations entering it are modified: means are trimmed and standard deviations are winsorized.⁸

Another popular option for effect size with the t-test is r , a well-known measure from correlation analysis (see Section 5.2). Cohen's d can be easily converted into r and vice versa (see Section 8.4). The advantage of r is a standard range of values (0 to 1); Cohen's d doesn't have a minimum or maximum value. Cohen (1988: 79–80) recommends the following interpretation of the r measure:

Interpretation of r : $r > 0.1$ small, $r > 0.3$ medium and $r > 0.5$ large effect

⁸ This process is introduced to eliminate outliers (see Section 1.3). While trimming means involves deleting a certain portion (e.g. 20% on each side of a distribution) of extreme values before the mean is calculated in the standard way (see equation 1.1), winsorization replaces these extreme values with the last non-extreme value before SD is calculated in the standard way (see equation 2.11). For more information see Erceg-Hurn & Mirosevich (2008).

In our case, r is 0.33, 95% CI [.08, .53], which, again, can be interpreted as a medium effect with the caveat that the true population effect may range from minimum (0.08) to large (0.53).

So far, we have compared two groups of speakers. However, what if we need to compare more? In such a case a test called **one-way ANOVA**⁹ (acronym for ANalysis Of VAriance) can be used. For instance, imagine that you are interested in the use of the non-standard form *ain't* in British speech and whether there is an effect of the socio-economic status of the speakers (social class). For illustration, we'll again use data from BNC64. Here are some examples of the use of *ain't* in BNC64:

- (1) cos there ain't [= isn't] no sign of them, is there ? (BNC64, M2)
- (2) I've won twice ain't [= haven't] I ? (BNC64, F5)

BNC64 classifies speakers into four categories reflecting the socio-economic status of the speakers: AB – Managerial, administrative, professional; C1 – Junior management, supervisory; professional; C2 – Skilled manual; DE – Semi- or unskilled.

ANOVA has the following assumptions, which are similar to the assumptions of the t-test discussed above: (i) independence of observations, (ii) normality and (iii) homoscedasticity. The most important one is the independence of observations because as was shown in the literature the ANOVA test is robust against the violation of the normality assumption (Schmider et al. 2010). In addition, if the homoscedasticity assumption is violated, Welch's version of ANOVA, which is robust against the violation of this assumption (Minitab n.d.), can be used. The equation of the one-way ANOVA¹⁰ is:

$$\text{One-way ANOVA (F)} = \frac{\text{Between-group variance}}{\text{Within-group variance}} \quad (6.7)$$

To illustrate the logic behind ANOVA, Figure 6.2 plots 60 speakers from BNC64 according to their use of *ain't* (relative frequency per 10 k words) and their social-class membership (shapes of data points); four speakers whose social class was unknown were excluded from the analysis. The top panel shows the between-group variance (shaded area multiplied by the cases in each group, i.e. how far each group's mean is from the grand mean), while the bottom panel shows within-group variance (shaded area). The dotted horizontal line in the top panel is **the grand mean**, i.e. the mean for the whole corpus. The four shorter horizontal lines represent means for the individual social-class groups. The squares in the bottom panel show the squared

⁹ One-way in the statistical jargon refers to the number of explanatory variables we consider, that is one. With the one-way ANOVA we divide speakers into multiple groups based on a single criterion such as socio-economic status – see the discussion below.

¹⁰ This is the equation of the traditional Fisher's one-way ANOVA, which is simpler to explain; Welch's ANOVA follows the same principles, but the equation includes a number of adjustments (see the companion website for the equation of Welch's ANOVA).

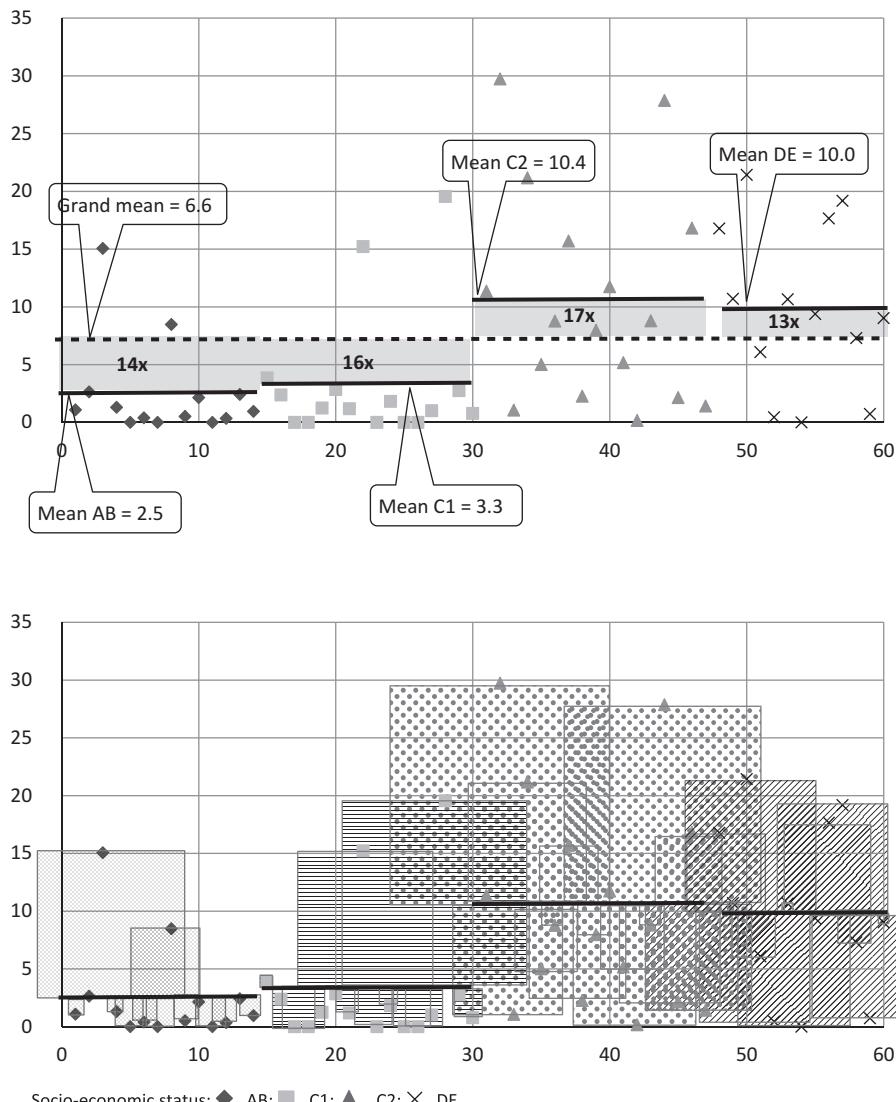


Figure 6.2 ANOVA calculation: between-group variance (top), within-group variance (bottom)

distances (shaded area) from the group means (shorter horizontal lines) for each data point.

To calculate between-group variance, that is the variation in the data explained by social class as the explanatory variable, we need to take the group means and calculate their squared distances from the grand mean and multiply these numbers by the number of individual cases in each group. This is then divided by the appropriate degrees of freedom: number of groups minus one.

$$\text{Between-group variance} = \frac{\text{cases group1} \times (\text{mean1} - \text{grand mean})^2 + \text{cases group2} \times (\text{mean2} - \text{grand mean})^2 + \dots}{\text{number of groups} - 1} \quad (6.8)$$

Within-group variance is the sum of individual variances for each of the groups, each calculated according to equation (6.1). The degrees of freedom is the number of cases minus number of groups.

$$\text{Within-group variance} = \frac{\text{sum of squared distances for group1} + \text{sum of squared distances for group2} + \dots}{\text{number of cases} - \text{number of groups}} \quad (6.9)$$

In our example, between-group variance, within-group variance and ANOVA (F) are calculated as follows:

$$\text{Between-group variance} = \frac{14 \times (2.5 - 6.6)^2 + 16 \times (3.3 - 6.6)^2 + 17 \times (10.4 - 6.6)^2 + 13 \times (10.0 - 6.6)^2}{4 - 1} = 267.35 \quad (6.10)$$

$$\text{Within-group variance} = \frac{231.2 + 485.1 + 1341.4 + 621.7}{60 - 4} = 47.8 \quad (6.11)$$

$$\text{One-way ANOVA } (F) = \frac{267.35}{45.7} = 5.6 \quad (6.12)$$

The related p-value is $p = .002$ and hence we can conclude that the result is statistically significant.

Because ANOVA is an omnibus test, it detects statistically significant difference anywhere in the data (between any of the groups), but it does not tell us where exactly the difference lies. For this, we need to carry out so-called post-hoc tests. **Post-hoc tests** are pair-wise comparisons of individual group means (similar to the t-test discussed above) with a correction for multiple testing; with multiple testing the probability of a falsely positive result (a so-called type I error) increases. This is because with each test that uses a p-value we are willing to accept that in a small number of cases (5%) the result will be statistically significant, even if the null hypothesis is true (there is no effect of the explanatory variable). With multiple testing, this probability dramatically increases. For instance, with four groups (as in our example), we can run six pair-wise comparisons, which will increase the type I error to 26.5%.¹¹ This is not an acceptable error rate – we therefore apply corrections for multiple testing such as Bonferroni's correction (Toothaker 1993; Shaffer 1995; Cabin & Mitchell 2000).

Like post-hoc tests, t-tests with Bonferroni correction are often reported; Bonferroni correction is fairly strict and hence the test is fairly conservative, i.e. it may not have the statistical power to detect small differences. Other options

¹¹ This is called the **family-wise error**, because it is connected with multiple testing based on a family of (several related) statistical inferences. The family-wise error is calculated as: $1 - 0.95^{\text{number of tests}}$, which is, in our example, $1 - 0.95^6 = 0.265$

ID	FEMALE		MALE	
	Relative freq.	Rank	Relative freq.	Rank
1	1794.6	4	1296.1	57
2	1681.3	10	1402.0	48
3	1282.5	59	1427.1	44
4	1435.1	42	1394.4	50
5	1377.9	53	1513.7	29
6	1577.8	16	1529.6	25
7	1432.1	43	1395.8	49
8	1485.7	34	1283.9	58
9	1564.5	20	1413.7	45
10	1609.6	14	1547.7	24
11	1493.4	32	1439.0	41
12	1558.1	22	1277.0	61
13	1575.8	17	1465.2	38

Figure 6.3 Dataset from BNC64 – relative frequencies and ranks: use of personal pronouns

include Tuckey's HSD or different tests based on bootstrapping, a method of multiple resampling (see Section 7.3).

Finally, when reporting ANOVA, in addition to statistical significance, the effect size needs to be reported. The overall (omnibus) effect size that is sometimes reported is eta squared (η^2). The standard interpretation of this effect size is as follows (Richardson 2011; Kirk 1996):

Interpretation of partial η^2 : $\eta^2 > 0.01$ small, $\eta^2 > 0.06$ medium and $\eta^2 > 0.14$ large effect.

A better (more specific) option, however, is to report effect sizes for differences between individual groups alongside the post-hoc tests. Cohen's d or r are the most commonly used effect sizes in this situation.

The t-test and ANOVA also have their **non-parametric counterparts** called the **Mann–Whitney U test** (also known as Wilcoxon rank-sum test) and the **Kruskal–Wallis test**. Non-parametric means that we don't need to know (assume) any knowledge about the parameters (such as the mean or standard deviation) of the variable of interest in the population (Dodge 2008). These tests thus do not have the assumptions about the underlying normal distribution of the (socio)linguistic variable in the population, equality of variances etc., and can be used in cases of ‘unruly’ (i.e. extremely skewed) scale data or data that is ordinal. Instead of relative frequencies, these two tests work with ranks; relative frequencies (scale variable) can easily be transformed into ranks, although this is at the cost of losing some information about the exact differences between speakers (cases). Figure 6.3 demonstrates a dataset transformation from relative frequency to ranks; this dataset (only partially displayed) shows the distribution of personal pronouns in two gender groups in BNC64 discussed above in relation to the t-test.

To calculate the Mann–Whitney *U* test and the Kruskal–Wallis test the following steps need to be taken:

1. If the linguistic variable is a scale variable, all data is ranked regardless of the group membership with the highest value receiving rank 1 and tight scores both/all receiving an average rank (e.g. if all three top values are 10.5, the ranking will be 2, 2, 2 calculated as $(1+2+3)/3$).
2. The data is then divided into groups according to the explanatory variable, e.g. gender or socio-economic status, and the sum of ranks is calculated for each group.

For the Mann–Whitney *U* (Wilcoxon rank-sum) test, two *U* values are calculated,¹² one for each group, from which the smaller value is then taken (Mann & Whitney 1947; Kerby 2014):

$$U_1 = \text{sum of ranks for group1} - \frac{\text{cases in group}_1 \times (\text{cases in group}_1 + 1)}{2} \quad (6.13)$$

$$U_2 = \text{sum of ranks for group2} - \frac{\text{cases in group}_2 \times (\text{cases in group}_2 + 1)}{2}$$

The idea behind the test is this: we take into account the actual sum of ranks for each group from which we subtract the minimum possible sum of ranks that the group can achieve [$\text{cases in group}_1 \times (\text{cases in group}_1 + 1) / 2$]. Suppose that we have two groups of five speakers, one of which (group₂) has all the smallest ranks (1–5). The sum of ranks for this group is $1 + 2 + 3 + 4 + 5 = 15$ and the minimum possible sum of ranks for this group is also 15 [$(5 \times 6)/2$]. The resulting *U*, the smaller of the two options from equation (6.13), is 0 ($15 - 15$). From the logic of the test follows that small *U* values show strong differences between groups as exemplified above.

The Kruskal–Wallis test (producing *H* values) works on a similar principle but takes into account ranks in multiple (3+) groups (Kruskal & Wallis 1952). The information about statistical significance of *U* and *H* is automatically provided by statistical packages and can be also obtained from statistical tables.

Different effect size measures have been proposed for the non-parametric tests (Kerby 2014). As discussed above, an effect size that specifically quantifies the difference between two groups (rather than an omnibus effect size measure) is probably most useful to report. The Mann–Whitney *U* test can thus be supplemented by the rank biserial correlation (Glass 1965). **Rank biserial correlation (*r_{rb}*)** is calculated as follows:

$$\text{Rank biserial correlation } (r_{rb}) = \frac{\text{mean rank group1} - \text{mean rank group2}}{\text{number of all cases}}$$

¹² Note that equation (6.13) presents a simpler version of the one presented in the original paper (Mann & Whitney 1947: 51).

Like the Pearson's or Spearman's correlation coefficients (see Section 5.2), r_{rb} can take on values from -1 to 1. The larger the value is in absolute terms, the stronger the correlation; a negative value signifies that group 2 has a larger mean rank value. Another option for effect size measure is to use probability of superiority (PS) discussed in Section 8.4.

A final note: if we have multiple samples for each individual speaker and are interested, for example, in how their language develops over time or how it was affected by another explanatory variable between two or more sampling points, a so-called repeated measures version of the tests discussed in this chapter needs to be used. **Repeated measures tests** match individual speakers across conditions and do not assume random distribution of speakers in different groups (see Verma 2016). Another extension of the procedures described in this section is a **multi-way ANOVA (factorial ANOVA)**, a test that takes into account two or more explanatory variables at the same time (e.g. gender and social class) as well as their interactions (see Field et al. 2012: chapter 12 for a practical example).

Reporting Statistics: T-Test, ANOVA, Mann–Whitney *U* Test and Kruskal–Wallis Test

1. What to Report

When comparing groups of speakers using the t-test, ANOVA, Mann–Whitney *U* test and Kruskal–Wallis test, we report (i) the test statistic (t , F , U and H respectively), (ii) the degrees of freedom (for t-test and ANOVA), (iii) p-value and (iv) effect size (including 95% CI).

Exact p-values are reported unless they are smaller than 0.001; after this point $p < .001$ is reported. For ANOVA and the Kruskal–Wallis test, where multiple groups are involved, post-hoc tests and their relevant effect sizes should also be reported, where relevant. It is also possible to accompany the tests by data visualization such as boxplots and/or error bars showing 95% CIs.

2. How to Report: An Example

- | | |
|---------------|--|
| t-test | <ul style="list-style-type: none"> • There was a statistically significant effect of gender on the use of personal pronouns, $t(61.93) = 2.77, p = .007$, with women ($M = 1,556.9, SD = 154.3$) using personal pronouns more than men ($M = 1451.8, SD = 149.2$). The size of the effect is medium, $d = .69$, 95% CI [.18, 1.21]. • In BNC64, there was a significant effect of gender on the use of personal pronouns: $t(61.93) = 2.77, p = .007, d = .69$, 95% CI [.18, 1.21]. • There was a significant effect of social class on the use of the form <i>ain't</i>, $F(3, 56) = 5.59, p = .002, \eta^2 = .432$. |
|---------------|--|

(cont.)

ANOVA

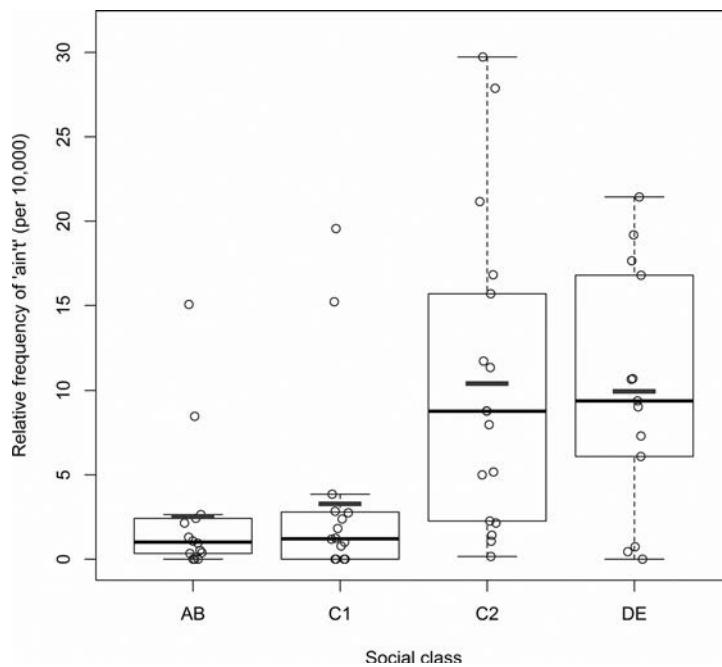


Figure 6.4 *Distribution of ain't in BNC64 speakers: social-class effect*

95% confidence limits

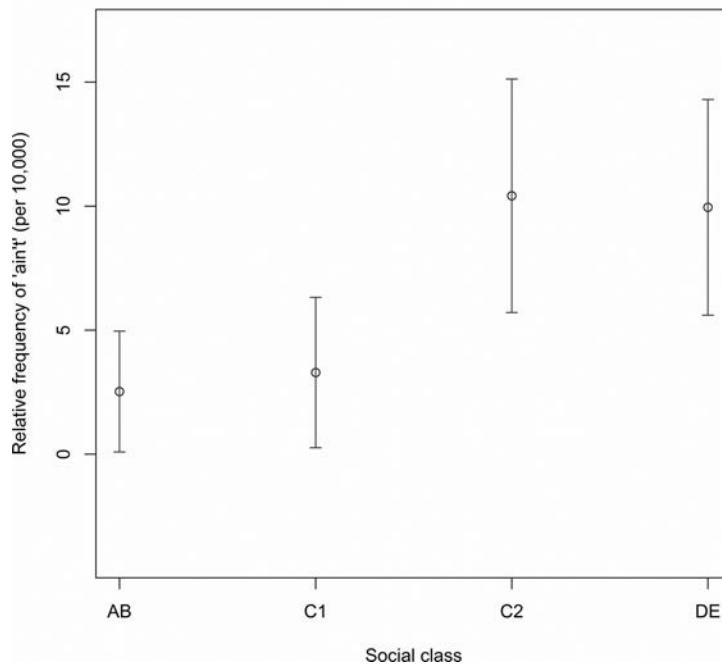


Figure 6.5 *Ain't in BNC64: 95% CI*

(cont.)

- As we can see from Figure 6.4, there is a clear split between the AB and C1 (i.e. middle-class) speakers and the C2 and DE (working-class) speakers. This observation is also valid for the population as seen from the 95% Confidence intervals (CIs) in Figure 6.5.
 - There was a significant effect of social class on the use of the form *ain't*, $F(3,56) = 5.59, p = .002, \eta^2 = .432$. As indicated by post-hoc tests (Bonferroni), the significant difference is between AB and C2/DE as well as between C1 and C2/DE (all $p < .05$).
- Mann-Whitney *U***
- The use of pronouns by female speakers ($Mdn^{13} = 1,536.6$) differed significantly from the use by male speakers ($Mdn = 1433.0$), $U = 705, p = .009$, $r_{rb} = -.19$.
- Kruskal-Wallis**
- There was a significant effect of social class on the use of the form *ain't*, $H(3) = 15.57, p = .001$.

6.4 Individual Style: Correspondence Analysis

Think about ...

Before reading this section, look at the samples of transcribed speech. These come from two different speakers; three samples belong to one speaker, the remaining sample to another speaker. Can you tell which samples belong together? Are there any linguistic clues that can help you tell the speakers apart?

Sample 1

You lot ain't supposed to know I'm taping. Oh *you wanna listen* to something, but *you don't know* what *you wanna listen*. No! *It's nothing to do* with this school. No one in this school **listens** to *it*. **Can I have** a look at the bottle please?

Sample 2

Yes, exactly. Yeah. Yeah. But no *it*, *it doesn't bother me*. *I don't mind you know* on camera, *I don't mind being on video*, anything. Warts and all. Oh no no. *I know I can't see* Mike and Robin **making** a cake, **can you?** Not, no no no.

¹³ Mdn stands for median, the middle value (see Section 1.3). With non-parametric tests, medians are reported instead of means.

Sample 3

And *you were saying* all this stuff. *You were. I've got* it on tape. That's what *you were saying*. And *you were saying*, oh yeah, *we found* these porno magazines and *we're selling them* off to *perverts*. And *I said* yeah but *you're looking* at *them* and, and *you started laughing*.

Sample 4

You're fifteen years old and *you still wanna watch* the Jungle Book. Honestly! Some people! Oh don't start on me you know, *saying* *I can't*. there on Tuesday! *Don't start* because *I'll, I'll smash your face in!* What *am I doing* down here?

Word-class code: nouns, verbs, pronouns

Individual style has been investigated with a range of linguistic features both lexical and grammatical. In this section, we take a look at a technique which is somewhat similar to factor analysis discussed in Chapter 5 (Section 5.4), but that has been primarily developed for the analysis of cross-tabulation tables with categorical data, the type of data which was discussed in Chapter 4 (Section 4.3).¹⁴ The technique is called **correspondence analysis**, sometimes also known as **optimal scaling** and **homogeneity analysis** (Clausen 1998: 1). Correspondence analysis is a summary technique which outputs a correspondence plot. A **correspondence plot** is a visual depiction of a cross-tabulation table which is projected in a (typically) two-dimensional space using the chi-squared distance as a measure of closeness/remoteness of the categories listed in the table. A 2D correspondence plot (e.g. Figure 6.6) is the most useful depiction of complex reality because it reduces the number of dimensions of variation to the manageable two dimensions represented by the x-axis and the y-axis.¹⁵

The unique feature of the correspondence plot is the fact that it captures both the column and row categories of the cross-tabulation table in the same space. For example, the correspondence plot in Figure 6.6 shows three individual speakers (F1, F6, M7 and M28) clustered according to their use of different word classes. Both the speaker samples and the word classes are displayed in the plot. For each speaker, three samples of exactly 1,000 words each were used in which the frequencies of nouns, verbs, adjectives etc. were traced. We can see that each time the three speech samples from the same speaker cluster closely together, thus distinguishing the particular speech style of the speaker. The full dataset on which the correspondence plot in Figure 6.6 is based can be seen in the cross-tabulation table (Table 6.1).

¹⁴ Although this technique was designed for categorical data such as frequency counts of different linguistic features or lexico-grammatical categories, Greenacre (2007: 182–4) shows that the technique is more versatile and can be used even with scale data (e.g. relative frequencies). Scale data is often transformed into ranks or z-scores, if different scales are involved.

¹⁵ 3D correspondence plots are also possible, but their interpretation is more difficult because analysing three dimensions at the same time becomes fairly complex; moreover, two dimensions are often sufficient to capture most of the variation in the data.

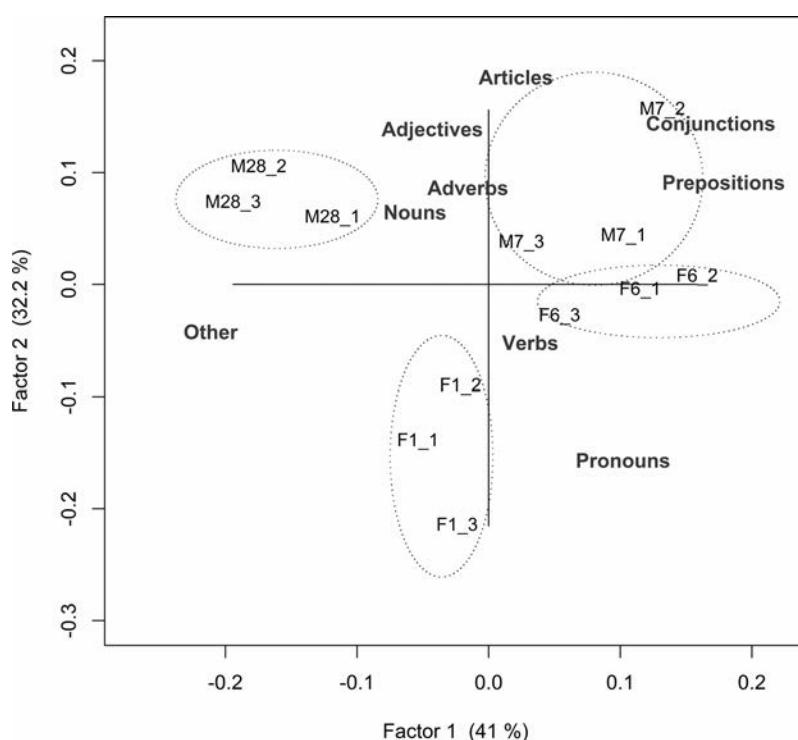


Figure 6.6 A correspondence plot: word classes in the speech of individual speakers

Table 6.1 Cross-tabulation table: word classes in the speech of individual speakers

Sample	Verbs	Nouns	Pronouns	Adjectives	Adverbs	Prepositions	Conjunctions	Articles	Other word classes
F1_1	257	154	197	42	81	39	26	39	165
F1_2	240	139	182	19	82	65	42	49	182
F1_3	267	101	208	30	83	53	39	26	193
F6_1	256	109	179	29	103	61	69	61	133
F6_2	257	115	181	48	79	78	68	54	120
F6_3	257	152	171	23	97	64	49	49	138
M7_1	250	121	163	31	134	68	57	55	121
M7_2	218	161	148	47	92	87	70	67	110
M7_3	248	130	151	47	107	76	41	53	147
M28_1	235	154	124	41	108	48	52	49	189
M28_2	219	154	106	42	114	57	42	58	208
M28_3	219	141	120	48	100	40	47	66	219

Table 6.2 *Cross-tabulation table: verbs and articles in the speech of two speakers*

Sample	Verbs	Articles	TOTAL
F1_1	257	39	296 (row total)
M28_3	219	66	285 (row total)
TOTAL	476 (column total)	105 (column total)	581 (grand total)

Let us briefly go back to the ‘Think about’ task. Samples 1, 3 and 4 represent speaker F1 from Figure 6.6 and Table 6.1. Sample 2 represents speaker M28. F1 is a young female speaker (14 years old) from London; M28 is an older male speaker (65 years old) from Lancashire. In the analysis, these speakers are clearly distinguished by different frequencies with which they use the standard word classes. While this tendency might not be immediately apparent when looking at small samples (about 50 words) in the ‘Think about’ task, it is possible to recognize the patterns in larger samples (1,000 words). This example thus shows a situation where quantification is clearly beneficial because it helps us uncover regularities in language use which may remain hidden if only a qualitative approach was used.

Conceptually, correspondence analysis is related to the chi-squared test discussed in Section 4.3, which can be performed on a two-way cross-tabulation table (e.g. Table 6.1) and which tests the homogeneity null hypothesis, i.e. establishes if the values are proportionally equally distributed in the table within a chance variation or if there are systematic differences between the categories displayed in the table (Greenacre 2007: 26–9; Benzécri 1992: 54–5). In our example, using the chi-squared test we would be asking if there is a difference between the speakers in terms of the distribution of the individual word classes in their speech. Instead of a p-value, which the chi-squared test produces, the correspondence analysis shows the relationship between categories visually by plotting both the speakers and the word classes in the same correspondence plot. While the chi-squared test can only answer a simple YES/NO question about statistical significance without indicating where exactly the difference lies (which is especially problematic with large cross-tabulation tables), the correspondence analysis can show us the larger picture of complex relationships (similarities and differences).

So how is the information in the table ‘translated’ into the visual representation in the correspondence plot? Let’s start by taking two rows and two columns from Table 6.1 (indicated by shading); these are displayed separately in Table 6.2. Note also that F1 and M28 are the speakers from the ‘Think about’ task. Table 6.2 reports also the row and column totals as well as the grand total.

The first step in the correspondence analysis is to turn the information in the table into what is called profiles. **Profiles** are proportions (percentages) based

Table 6.3 *Row (speaker) profiles*

Sample	Verbs	Articles	TOTAL
F1_1	257/296 = 0.87	39/296 = 0.13	296/296 = 1
M28_3	219/285 = 0.77	66/285 = 0.23	285/285 = 1
Average profile	476/581 = 0.82	105/581 = 0.18	581/581 = 1

on row and column totals (**row and column profiles**); the profiles are expressed as decimal numbers, e.g. 0.87, rather than percentages (see Section 4.3). A profile as a ‘system of proportions’ (Benzécri 1992: 4) is thus calculated by dividing the number in the cell by the row and column totals respectively for the row and column profiles. For example, the row profiles (representing the two speakers, F1 and M28) are calculated as shown in Table 6.3.

In a similar way, column profiles are also calculated and analysed in the correspondence analysis (see below). The row profiles from Table 6.3 can be graphed using a simple scatterplot (see Section 1.5) which displays in a two-dimensional space the speaker profiles; these profiles capture the proportions of use of verbs and articles by the two speakers. In our example, the verb proportions will form the x-axis coordinates (0.87 and 0.77 for F1 and M28 respectively) and the article proportions the y-axis coordinates (0.13 and 0.23 for F1 and M28 respectively). The **average profile** will be placed in the middle according to the average profile coordinates (0.82; 0.18). The result of the plotting can be seen in Figure 6.7.

The scatterplot in Figure 6.7 uses a so-called **Euclidean distance** between the points in the graph, which is the most direct distance between the points, that is, what we usually mean by ‘distance’ in the everyday use of the word. As explained in Section 5.3, there exist other types of distance (Manhattan, Canberra etc.) that can be used for statistical purposes. In correspondence analysis, we use a so-called chi-squared distance that is derived from an elaboration of the chi-squared test equation (Greenacre 2007: 26–9; Dodge 2008: 68). **Chi-squared distance** is a weighted form of the Euclidean distance giving proportionally more weight to categories with fewer instances. Chi-squared distance is calculated as follows:

$$\text{Chi-squared distance}(A, B) = \sqrt{\frac{(x_B - x_A)^2}{\text{Average profile } x} + \frac{(y_B - y_A)^2}{\text{Average profile } y} + \dots} \quad (6.14)$$

where x_A is the first coordinate of point A, x_B is the first coordinate of point B, y_B is the second coordinate of point B etc. We can keep adding categories (different

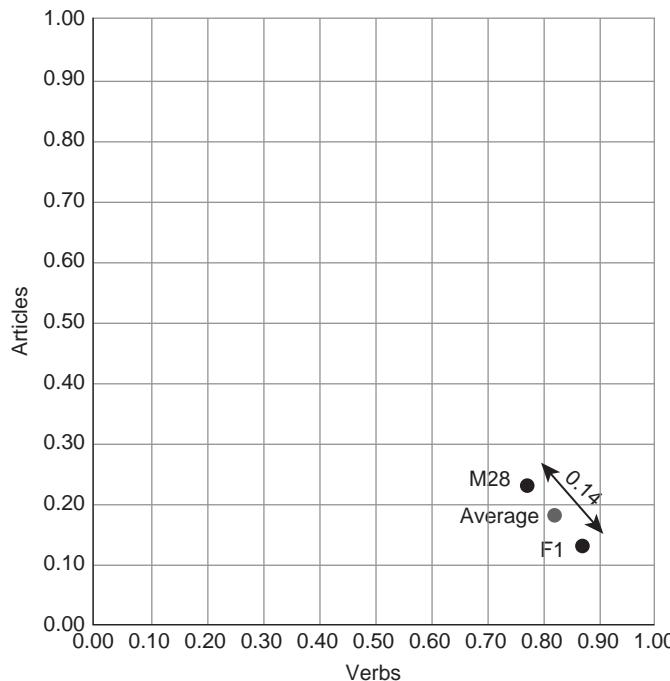


Figure 6.7 *Speaker (row) profiles: Euclidean distance*

word classes in our example) determining the position of the objects and move from a two-dimensional space to a multidimensional space.

In our example, the chi-squared distance is:

$$\text{Chi-squared distance}(F1, M28) = \sqrt{\frac{(0.77 - 0.87)^2}{0.82} + \frac{(0.23 - 0.13)^2}{0.18}} = 0.26 \quad (6.15)$$

In effect, the chi-squared distance between two points is always larger than the simple Euclidean distance as can be seen in Figure 6.8, which redraws the data from Table 6.3 with the chi-squared distance.¹⁶ There are many mathematical reasons for using the chi-squared as opposed to the Euclidean distance which are beyond the scope of this introduction (see Benzécri 1992: 44–58; Greenacre 2007). The best way to understand the use of the chi-squared distance in the correspondence analysis is to realize its relationship with the chi-squared test: the chi-squared test could be used as a simple alternative to the correspondence analysis for the same cross-tabulation tables.¹⁷

¹⁶ As can be seen from Figure 6.7, the equivalent Euclidean distance is 0.14.

¹⁷ Note that while the chi-squared test can be used only with two-way contingency tables (as the ones in our example), correspondence analysis is possible with multiple variables (multi-way tables). For multiple correspondence analysis see Greenacre (2007: 137ff).

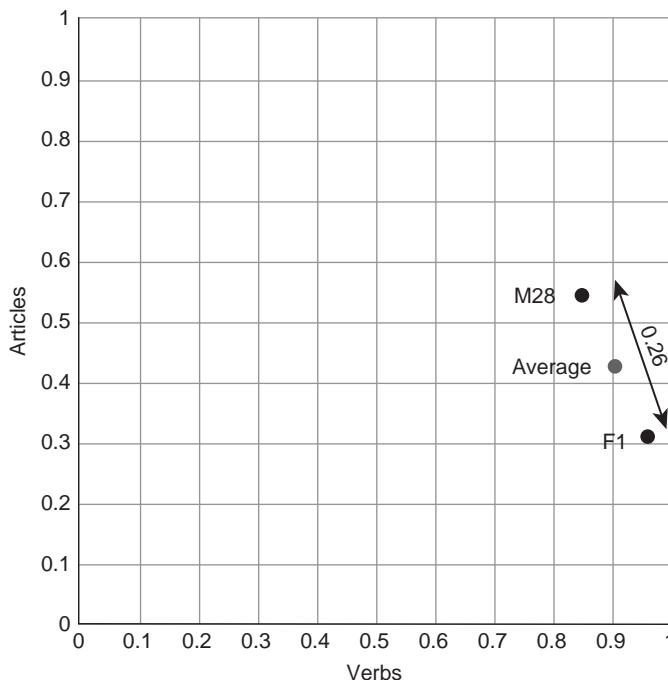


Figure 6.8 Speaker (row) profiles: chi-squared distance

The final (most complex) step of the correspondence analysis involves dimension reduction. While Table 6.2 includes two categories of word classes (verbs and articles) that can be displayed in a two-dimensional scatterplot, Table 6.1 includes nine categories, which would be impossible to display in the same simple scatterplot. For this reason, the correspondence analysis includes a dimension reduction procedure, somewhat similar to the one in factor analysis discussed in Section 5.4. The individual categories (articles, verbs etc.) in the scatterplot in Figure 6.8 are thus replaced by **factors** that combine multiple categories and that explain as much variation in the data as possible. The resulting correspondence plot, as we saw in Figure 6.6, is similar to a scatterplot in that it has two axes (x and y), which, however, represent the factors rather than single categories. To indicate the relationship between the single categories and factors, the correspondence plot displays also the single categories. The correspondence plot also shows how much of the original variation in the data each factor explains. In Figure 6.6 Factor 1 explains 41% and Factor 2 explains 32.2% of variation, which means that overall the correspondence plot explains 73.2% of variation, which is a very reasonable amount; the trade-off of the dimension reduction process is the fact that 26.8% of variation is lost ‘in the translation’ – we would need more dimensions (factors) to capture more variation, but the overall picture would get more complicated. For our purposes, the two-dimensional result is satisfactory because it shows us clearly the main trends in the data

while explaining a large amount of variation (over 70%). You can think about the correspondence plot as a map which simplifies and reduces the features of the terrain to the most important ones that help us navigate around a city or through a countryside and find our way to the destination.

When we look at a correspondence plot such as the one in Figure 6.6, we notice clusters of both linguistic categories (word classes) and speaker samples. The linguistic categories help us interpret the factors and their linguistic meaning. For instance, Factor 1 stretches from the other grammatical word classes to verbs and prepositions and conjunctions – the last two categories also have a high score on Factor 2; Factor 2, on the other hand, stretches from pronouns to articles. We can also observe the close (chi-squared) distances between prepositions and conjunctions as well as between nouns, adjectives, adverbs and articles. Most importantly, we can see that the samples drawn from the speech of the same speakers (F1, F6, M7, M28) cluster relatively closely together – again we can measure the chi-squared distances between the samples. In the plot, the speaker samples are displayed in relation to their preferences for different word classes. The plot thus suggests that there is strong evidence in the data for distinct styles of speaking in these four speakers.

Reporting Statistics: Correspondence Analysis

1. What to Report

Correspondence analysis is best summarized by the correspondence plot, which is based on a cross-tabulation table; both the plot and the table should be reported. We should note the overall percentage of variation explained by the two factors as well as the linguistic interpretation of these factors. We should then attempt to identify meaningful clusters of speakers in the plot and comment on their relationship. Note that the chi-squared distance can be interpreted only within row/column categories and the distances between row and column categories cannot be directly interpreted. For instance, in Figure 6.6 we can measure the exact (chi-squared) distance between the individual speaker data points; however, we cannot relate the speaker points directly to the individual word classes via measuring the distance in the plot because row and column categories operate on different scales.

2. How to Report: An Example

- The data in Table 6.1 was analysed using correspondence analysis. The resulting correspondence plot is displayed in Figure 6.6. Overall, the correspondence plot explains 73.2% of the variation in the data. We can see four main clusters of speaker samples ...

6.5 Linguistic Context: Mixed-Effects Models

Think about ...

Before reading this section, look at the examples below. Is there a meaning difference between the two utterances? Can you imagine contexts in which you would say these utterances?

- (i) This is really good.
- (ii) This is very good.

One of the crucial tenets of the traditional variationist sociolinguistics is what Labov calls the **principle of accountability**¹⁸ (e.g. Labov 1972: 27; Tagliamonte 2006: 12–13). This principle states that for the analysis to be meaningful, we have to find all contexts in which all variants of a sociolinguistic variable occur and include them in our analysis. In Section 6.2, we discussed the notion of the Labovian sociolinguistic variable and its potential limitations. Here, we take Labovian sociolinguistic variables as our starting point and look at ways in which we can analyse such variables. As an example, we will take the variation between utterances (i) and (ii) in the ‘Think about’ task; this example is inspired by Ito and Tagliamonte (2003). At first sight, there is no difference between the two utterances. In terms of the main content (sometimes called the ‘propositional meaning’) these utterances say the same thing, which is an important criterion for defining a Labovian sociolinguistic variable (see Section 6.2). Following the principle of accountability, our next step is to find all examples in the corpus where this type of variation occurs or can occur (a zero variant such as *This is good* is also a variant); this is, as we already know, called **circumscribing the variable context**. In our example, we have to search for all adjectives in the corpus that can be modified by intensifiers such as *really* and *very*, that is, adjectives in contexts in which they are used as gradable adjectives (see Ito and Tagliamonte 2003: 263–4 for more information about this step). Because in the case study below we want to focus only on the variation between the intensifiers *really* and *very*, the two most common intensifiers in speech, we have selected only a subset of all adjectives in the corpus: those adjectives that are modified by *really* or *very* (1,952 cases from 58,127 instances of adjective occurrence in the BNC64, a 1.5-million-word corpus of British speech (Brezina & Meyerhoff 2014)). This makes the circumscribing of the variable context more specific, narrowing down the research question, which in our case study is: what factors have an effect on the use of *very* as opposed to *really*?

¹⁸ Compare Leech’s principle of ‘total accountability’ (Leech 1992: 112).

ID	Speaker	Gender	Age	Class	Syntax	Outcome	L	Node	R
1 M31	A_male		77	A_DE	B_predicative	B_very	on the weekend # You s very partial	to your mum 's apple p	
2 F17	B_female		40	C_C1	B_predicative	B_very	not found very difficult	very difficult	. # Computer studies P
3 M30	A_male		70	D_AB	B_predicative	B_very	the Police Station ! # Ha very funny	! # Yellowish you mean	
4 F32	B_female		75	B_C2	B_predicative	B_very	every body # knew that	very bright	# or he did n't try , wel
5 F17	B_female		40	C_C1	B_predicative	B_very	my point of view I have	very punctual	and helpful . # If anyone
6 M30	A_male		70	D_AB	B_predicative	B_very	n't know whether we o	very difficult	, you know ? # They ma
7 F19	B_female		41	A_DE	A_attributive	B_very	with not moving that th	very bad	cough down your ches
8 F15	B_female		37	A_DE	B_predicative	B_very	that film # is n't it ? # is	very safe	, is it safe ?, no its not
9 F1	B_female		14	B_C2	B_predicative	A_really	do , choo , do do do	really good	. # I like shooting , sho
11 F4	B_female		20	C1	B_predicative	A_really	just tell everyone I'm r	really new	in the area . # So my ta
12 M22	A_male		51	D_AB	A_attributive	B_very	me with bells on my an	very good	point there Lynda . # E
13 F21	B_female		46	C_C1	A_attributive	B_very	, I said we did n't even	very professional	people # at the time , I
14 F9	B_female		30	B_C2	B_predicative	B_very	was very good cutting b	very difficult	to cut , I thought I did e
15 F11	B_female		34	B_C2	B_predicative	A_really	we 're gonna practice #	really interesting	, let me switch this thi
16 F24	B_female		54	B_C2	A_attributive	B_very	Right ! I'll leave that the	very peculiar	character , I just find it
17 M12	A_male		33	A_DE	B_predicative	B_very	# That's true , that 's tr	very true	# Do n't know , I do n't
18 F11	B_female		34	B_C2	B_predicative	B_very	, either way she # amaz	very practical	that 's what I find of he

Figure 6.9 *Sociolinguistic dataset: internal and external factors (an excerpt)*

The first analytical step is to extract the data from the corpus and classify all the examples in the dataset (see Figure 6.9); we need to look at both the linguistic contexts in which *very* and *really* occur (e.g. the syntactic position) – these are called the **internal factors**, and also at the speaker and situation-related variables (e.g. age, gender, genre, etc.) – these are called the **external factors**. The internal and external factors are known as **predictor variables**.

Figure 6.9 shows the first 18 lines from a sociolinguistic dataset based on the BNC64. The whole dataset has 1,646 valid entries (1,952 lines minus erroneous entries). Because typically the internal variables (in our case the syntactic position) have to be coded manually, we need to decide which predictor variables are relevant before the whole dataset is coded to avoid data recoding. If one or more judgement variables are involved, it is a good practice to double-code a random sample of the data and report an inter-rater agreement statistic (see Section 3.5).

The dataset in Figure 6.9 contains one internal ('Syntax') and four external predictor variables ('Gender', 'Age' and 'Class'). The outcome variable ('Outcome') is a binary variable with the variants 'very' and 'really'. The right part of the dataset ('L', 'Node' and 'R') shows the full concordance line, which was the basis for the coding but is not used for the statistical analysis. Note that the categorical variables in the dataset are coded with a prefix: A_, B_, C_ etc. This is to ensure a straightforward interpretation of the results against baseline values coded with the prefix A_ (see Section 4.4).

The statistical technique introduced in this section, whose application will be demonstrated with the dataset in Figure 6.9, is called the mixed-effects models. **Mixed-effects models** are a group of powerful multivariate statistical techniques. We will focus here on only one specific use of mixed-effects

```

Random effects:
Groups   Name        Variance Std.Dev.
Speaker (Intercept) 0.9902   0.9951
Number of obs: 1646, groups: Speaker, 60

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.89807  0.55256 -1.625  0.1041
GenderB_female -0.18425  0.32193 -0.572  0.5671
ClassAB       0.13502  1.19296  0.113  0.9099
ClassB_C2      0.19664  0.46684  0.421  0.6736
ClassC_C1      0.07452  0.45810  0.163  0.8708
ClassC1        -0.63974  1.13243 -0.565  0.5721
ClassD_AB      0.72511  0.46981  1.543  0.1227
Age            0.03913  0.00940  4.163 3.14e-05 ***
SyntaxB_predicative 0.33611  0.15168  2.216  0.0267 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figure 6.10 *Mixed-effects models: output*

models that is somewhat similar to logistic regression (see Section 4.4). Mixed-effects models build a model from the data which best predicts the outcome (linguistic) variable, taking into account both so-called fixed and random effects, hence mixed-effects models (Gries 2013a; Johnson 2009).

Fixed effects are explanatory variables that are the direct object of the socio-linguistic investigation (e.g. gender, social class, age or genre). **Random effects** represent other sources of variation that need to be taken into account in the model such as individual speaker preferences, but are not part of the research question. Mixed-effects models can thus analyse a complex linguistic situation and account for both systematic variation due to different external and internal factors (fixed effects) as well as individual variation due to speaker preferences etc. (random effects).

Because of the similarity with logistic regression described in detail in Section 4.4, readers are referred to Chapter 4 for a detailed explanation of the basic terms and principles. In what follows, the main differences between logistic regression and mixed-effects models (mixed-effects logistic regression) are highlighted and an example of the use of this technique is provided. When we run the mixed-effects model technique we get an output such as that displayed in Figure 6.10.

The interpretation of mixed-effects models has two main parts: (i) evaluation of the model characteristics and (ii) interpretation of the significance of the fixed effects. In our example, the model is statistically significant and shows the effect of age and syntactic position on the use of *very* as opposed to *really* (our baseline variant marked with the prefix A_). Older speakers prefer *very* and syntactic contexts with a predicative adjective (e.g. *you're very lucky*) also favour this intensifier. The other predictors are non-significant. This result is in line with Ito and Tagliamonte's (2003) study, which was based on a

different (smaller) corpus; the study also offers a more detailed discussion of the results.

The main practical difference between a simple logistic regression and the mixed-effects model described above is the ability of the mixed-effects model to take into account individual variation between speakers. This is a largely desirable feature, which makes the mixed-effects model superior to a simple logistic regression known as the VARBRUL method in sociolinguistics (Johnson 2009). To bridge the terminological gap between sociolinguistics and corpus linguistics, below is a small dictionary of terms used in these two disciplines when talking about sociolinguistic data:

A Small Dictionary of Sociolinguistic Terms

In corpus linguistics and variationist sociolinguistics, different terms are used to mean the same thing or the same terms are used to mean different things (Johnson 2009); all of this may cause confusion. The following mini-dictionary should help clarify these terminological differences.

Variationist term	Meaning	Corpus linguistics equivalent
factor	A type of speaker (e.g. male or female, young or old) or a type of context (e.g. syntactic position) which favours the use of a particular variant of the sociolinguistic variable.	level of a predictor variable
factor group	A variable that is used to explain the (socio) linguistic variation.	predictor variable or factor
factor weight	This number expresses the importance of the effect of a factor as a probability; the alternative term ‘coefficient’ is usually expressed in log odds units.	coefficient
principle of accountability	A general scientific principle, which states that we need to account for all relevant instances in the data; the opposite of ‘cherry picking’.	total accountability
sociolinguistic variable	The linguistic variable that we are interested in; sociolinguistic variable has multiple (usually two) variants that express the same meaning.	outcome variable

Variationist term	Meaning	Corpus linguistics equivalent
token	An occurrence of a linguistic feature of interest. In contrast, a ‘token’ in corpus linguistics is any running word in a corpus.	case, observation
VARBRUL analysis	An acronym for VARiaBle RULE analysis; a typical statistical procedure in variationist sociolinguistics.	(a form of) logistic regression
variant of a sociolinguistic variable	One of two or more linguistic forms which compete with each other in a particular context and which have the same meaning.	level of an outcome variable

Reporting Statistics: Mixed-Effects Models

1. What to Report

With mixed-effects models, it is important to report the type of mixed-effects model used as well as the details of the model. The focus of the reporting will be on the effect of the individual predictors.

2. How to Report: An Example

- Mixed-effects logistic regression was used with the individual speakers as random effects and gender, class, age and syntactic position as fixed effects. The model, which was overall significant ($p < .001$), showed a significant effect of age and syntactic position: older speakers and predicative contexts favour the intensifier *very*.

6.6 Application: Who Is This Person from the White House?

I have a friend (they wish to remain anonymous), who every year, instead of a birthday card sends me a linguistic puzzle. This year, the puzzle had the form of a ‘sociolinguistic riddle’ with an attachment, which contained a file with transcribed speech. This is what appeared in my mailbox:

From: [REDACTED]
Sent: 19 December 2015 17:13
To: Brezina, Vaclav
Subject: Happy birthday

 Message  Attachment: Speech_transcript.zip

Sociolinguistic Riddle

I am a man or a woman.
I am young or old.
I work in Washington DC and
speak for the President.
Who am I?
(Find out without Google.)

Attachment: Speech_sample.zip

The good news is that the filibuster has been broken on national service. The Senate has decided that there is no need for a second cloture vote. Mitchell and Dole just announced that national service will be the first order of business on Tuesday and we are fully confident that the Senate bill will now pass. So good news breaking out all over. [...]

What follows is a brief description of my work over the following few days after receiving the email, which I spent solving the riddle. In forensic linguistics (my quest closely resembled the detective work of a forensic linguist), there are two basic approaches, which depend on the amount of evidence available: if the amount of evidence is small (a few sentences or paragraphs), close reading for signs of idiosyncratic language use (or shibboleths) is appropriate. If, on the other hand, as was my case, there is a lot of data available (the sample was approximately 200,000 words), the statistical approach is called for. The second part of the riddle was clear and matched the type of language in the sample. ‘I work in Washington DC and speak for the President’ indicated that the speech sample comes from a White House press secretary. Luckily, I could use a comparative corpus of White House press conferences (WH) (Barlow 2013), to try to match the sample in the attachment, which I called ‘transcript X’, with the WH corpus data. Table 6.4 is an overview of the corpus I used.

The first step was to decide whether transcript X came from a male or a female speaker. There are two women in the corpus (Dee Dee Myers and Dana Perino) and six men. The women contribute one sample each, while there are 33 speech samples from the male group. It is not possible to use traditional gender-distinguishing words such as *lovely* (preferred by women in informal British speech), because in the whole corpus (over 6 million words) there are only 20 occurrences of *lovely*. This is no surprise: we are dealing with a very specific formal spoken register of American English. We

Table 6.4 *WH corpus*

Press secretary	Born	In office	No. of samples (200 k words)	No. of tokens	File names
Dee Dee Myers	1961	1993–4	1	0.2 million	DDM
Mike McCurry	1954	1995–8	6	1.2 million	M1, M2, M3, M4, M5, M6
Ari Fleischer	1960	2001–3	4	0.8 million	A1, A2, A3, A4
Scott McClellan	1968	2003–6	3	0.6 million	S1, S2, S3
Tony Snow	1955	2006–7	1	0.2 million	T1
Dana Perino	1972	2007–9	1	0.2 million	D1
Jay Carney	1965	2011–14	7	1.4 million	JC1, JC2, JC3, JC4, JC5, JC6, JC7
Josh Earnest	1975	2014–16	8	1.6 million	J1, J2, J3, J4, J5, J6, J7, J8

therefore need to look at more general features such as the use of personal pronouns, which can be hypothesized to distinguish speaker's gender (Argamon et al. 2003; Rayson et al. 1997). To compare two groups of speakers, the t-test is typically used. Although it might be tempting to use all the data in the corpus (33 speech samples), this cannot be done. The assumption of the independent samples t-test is that each of the texts was sampled independently of the other texts; this assumption would be violated if we included multiple texts per speaker. If we reduce the sample to one text per speaker we end up with 9 texts (2 for women and 7 for men). This is a very small and uneven sample (with many more male than female speakers) and traditional statistical wisdom (e.g. Siegel 1956) would advise against using the t-test in this situation. However, new research (de Winter 2013) shows that the t-test performs well even with very small and uneven samples as long as the equality of variances assumption is not severely violated; we can take care of this assumption by using the robust Welch's version of the t-test.

The t-test returns a statistically significant result [$t(3.87) = 4.47; p < .05$] for the difference between male and female White House press secretaries when looking at the use of personal pronouns. The mean for the male group was 14,232.86 ($SD = 1,672.08$), while the mean for the female group was 18,065.50 ($SD = 820.95$). The value for transcript X was 18,044, clearly much closer to the mean for the female group. It can be hypothesized that transcript X belongs to a female speaker. In the corpus, there are two female speakers: Dee Dee Myers and Dana Perino. Which one will it be?

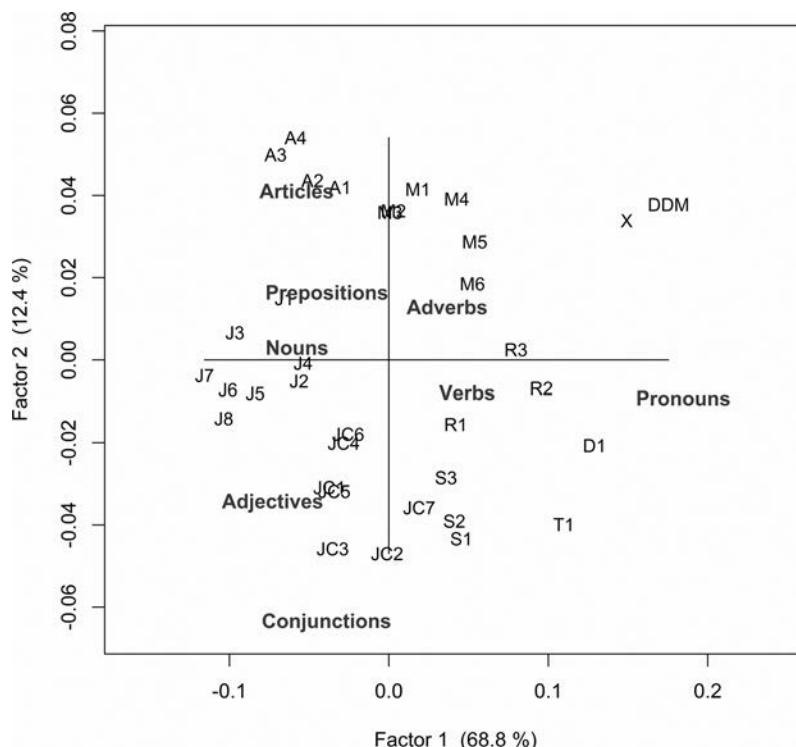


Figure 6.11 *Correspondence analysis: use of word classes by White House press secretaries*

To answer this question, correspondence analysis was used with all the speech samples in the corpus plus transcript X (34 + 1). The analysis looked into the proportions of different word classes in these samples as the linguistic variables which are both frequent and independent of the topic discussed (see Figure 6.11).

The correspondence analysis clearly grouped individual speech samples from the WH press secretaries together. For instance, all speech samples from Ari Fleischer (A1–A4) cluster at the top left, while speech samples from Scott McClellan (S1–S3) cluster at the bottom right closer to the centre than speech samples from Tony Snow (T1) and Dana Perino (D1). What about the mysterious transcript X? In the graph, it clusters in the proximity of the sample from Dee Dee Myers and far apart from the other samples. With a high probability it can thus be assumed that transcript X comes from this WH press secretary. The speech of Dee Dee Myers, the first ever woman to hold the post, is characterized by the frequent use of pronouns, a relatively infrequent use of nouns, adjectives and conjunctions. These stylistic differences can be identified only by taking large speech samples and analysing them quantitatively.

Following the investigation described above, I replied to my friend's email:

From: Brezina, Vaclav
Sent: 1 January 2016 00:30
To: [REDACTED]
Subject: Re: Happy birthday

Was it Dee Dee Myers (no p-value needed)? I'll go and google it ;)

For those of you who are still not convinced, you can follow my example and try to search for the first sentence from transcript X on the internet: 'The good news is that the filibuster has been broken on national service.'

6.7 Exercises

1. Which linguistic disciplines use the notion of 'style' as one of their core concepts? How can it be operationalized?
2. Which of these cases of variation satisfy Labov's definition of a sociolinguistic variable and can thus be investigated using Labov's methods? Justify your answer.
 - (a) *h*-dropping in different social contexts: i.e. the variation between the pronunciation of e.g. the word *hair* as /heə/ or /eə/.
 - (b) The variation between naming strategies for soft drinks in the US, e.g. *soda*, *pop* or *coke*. (<http://popvssoda.com/>)
 - (c) The variation between the active and the passive construction, e.g. *I did it* vs *It was done*.
 - (d) The variation between verbs expressing preferences with a different intensity, e.g. *adore*, *love*, *like* and *don't mind*
 - (e) The use of hedges in discourse and the variation between a hedged and unhedged utterance, e.g. *I sort of agree* vs *I agree*.
3. Look at the following ways of reporting the t-test, ANOVA and the Mann–Whitney *U* test. Identify erroneous or missing pieces of information.
 - (a) $t = 2.77, p = .007, d = .69$.
 - (b) $F(56) = 5.59, p < .00201359$.
 - (c) $U = 705, p = 1.3$
4. Use the equations in this chapter to calculate the test statistics (i.e. not the p-values) for:
 - (a) The use of *lovely* in male and female speech [t-test, Mann–Whitney *U*].
 Male group ($n1=10$): 0.91, 1.4, 2.18, 6.21, 2.63, 1.2, 0, 1.06, 6.49, 5.43
 Female group ($n2=10$): 8.84, 1.09, 12.47, 1.65, 3.93, 1.1, 4.11, 21.21, 2.51, 0.47
 - (b) The use of *innit* in the speech of speakers from the South, Midlands and the North [one-way ANOVA].
 South ($n1 = 10$): 4.19, 29.29, 5, 30.43, 6.09, 12.77, 25.93, 0.61, 28.08, 15.94
 Midlands ($n2 = 10$): 9.68, 3.65, 1.2, 0, 2.07, 2.26, 5.18, 0, 0, 0
 North ($n3 = 10$): 9.09, 9.09, 0, 7.38, 0, 5.77, 0, 4.47, 0, 3.23

5. Calculate Cohen's d for the dataset (a) in Exercise 4.
6. Use the Group comparison tool to check results of Exercises 4 and 5.
7. Interpret the correspondence plot in Figure 6.12. It is based on BNC64, a corpus of informal British speech. 16 male and 16 female speakers are plotted on the graph according to their use of different semantic types of certainty markers (*certainly, maybe, perhaps, possibly, etc.*). Each speaker is labelled according to their gender (F1a and M1a), number identifier (F1a) and sample number (F1a, F1b); there are two samples per speaker.

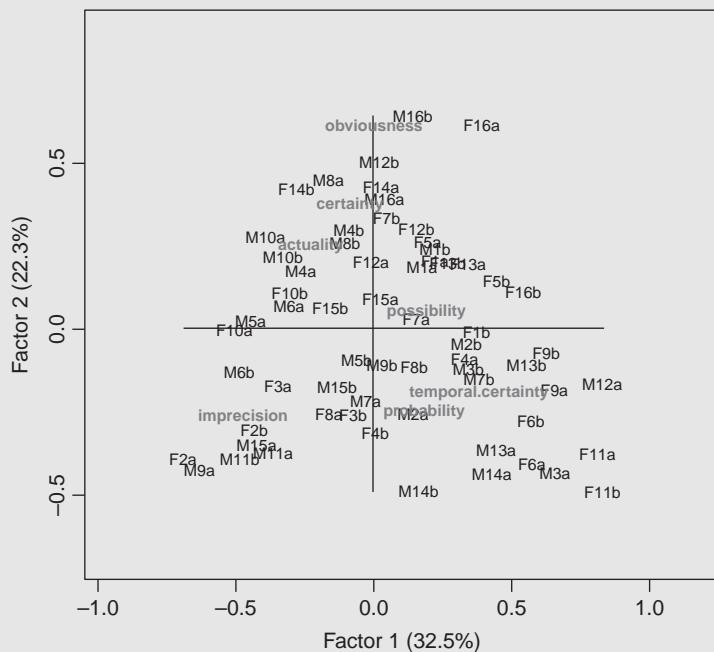


Figure 6.12 Correspondence analysis: use of epistemic markers in BNC64

8. Interpret the following output from the mixed-effects model technique; the linguistic variable is the use of *must* competing with other strong modal terms such as *have to* and *need to* in different genres of British and American English.

```

AIC      BIC    logLik deviance df.resid
2412.3   2456.9  -1198.1   2396.3     1943

Scaled residuals:
    Min     1Q Median     3Q    Max
-1.9037 -0.6202 -0.3895  0.7452  2.4673

Random effects:
Groups Name        Variance Std.Dev.
Text   (Intercept) 1.52      1.233
Number of obs: 1951, groups: Text, 743

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.1783    0.2364   0.754  0.450635
VarietyB_BR 0.1567    0.1515   1.034  0.300982
GenreB_Fiction -0.9825   0.2717  -3.616  0.000299 *** 
GenreC_General -0.2894   0.2569  -1.126  0.259977
GenreD_Press   -0.7682   0.2920  -2.631  0.008511 ** 
SubjectB_I    -1.1864   0.2153  -5.511  3.58e-08 *** 
SubjectC_you   -0.8857   0.2030  -4.363  1.29e-05 *** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

9. *Do men swear more than women?* Use the BNC64 Search & Compare tool (corpora.lancs.ac.uk/bnc64) to test different sociolinguistic hypotheses about swearing in informal British speech. Pay attention to the different statistical measures and their interpretation. Fill in Table 6.5 and pay special attention to the different statistical measures and their interpretation.

Table 6.5 Swearing and gender: BNC64

Swearword	Statistically significant result?	Meaningful difference?
1.		
2.		
3.		
4.		

THINGS TO REMEMBER

- Sociolinguistic variation can be operationalized in different ways: by employing the Labovian meaning-preserving sociolinguistic variable (formal approach), or by following the functional approach and looking at the distribution of linguistic features in groups of speakers.

- The t-test and ANOVA (as well as their non-parametric counterparts: Mann–Whitney U and Kruskal–Wallis) are used to investigate the effect of explanatory social variables (gender, social class) on the use of different linguistic features.
 - Correspondence analysis is an exploratory analytical technique, which compares the use of multiple variables in different speakers reducing them to two factors and producing a powerful visualization – a correspondence plot.
 - Mixed-effects models is a group of sophisticated statistical techniques which can account for multiple variables at the same time and include the effect of individual variation between speakers.
-

Advanced Reading

- Brezina, V. & Meyerhoff, M. (2014). Significant or random? A critical review of sociolinguistic generalisations based on large corpora. *International Journal of Corpus Linguistics*, 19(1), 1–28.
- Cardinal, R. N. & Aitken, M. R. (2013). *ANOVA for the behavioral sciences researcher*. Hove: Psychology Press.
- Corder, G. W. & Foreman, D. I. (2009). *Nonparametric statistics for non-statisticians: a step-by-step approach*. Hoboken, NJ: Wiley.
- Coupland, N. (2007). *Style: language variation and identity*. Cambridge University Press.
- Greenacre, M. (2007). *Correspondence analysis in practice*. Boca Raton, FL: Chapman & Hall/CRC.
- Johnson, D. E. (2009). Getting off the GoldVarb standard: introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass*, 3(1), 359–83.
- Labov, W. (2010). *Principles of linguistic change*, vol. 3: *Cognitive and cultural factors*. Oxford: Wiley-Blackwell.

Companion Website: Lancaster Stats Tools Online

1. All analyses shown in this chapter can be reproduced using the statistical tools available from the companion website. In particular, the available tools include:
 - Group Comparison tool
 - Correspondence analysis tool
 - Mixed-effects logistic regression tool
 - BNC64
2. The website also offers additional materials for students and teachers.

7 Change over Time

Working Diachronic Data

7.1 What Is This Chapter About?

This chapter discusses statistical procedures that can be used to explore historical or diachronic data, that is data that shows the development of a linguistic variable over time. First, specific features of diachronic studies are outlined and techniques that provide effective visualizations of diachronic change are introduced. Second, the chapter focuses on the statistical comparison of two time periods using a procedure called bootstrapping. Next, the diachronic application of cluster analysis (introduced in Chapter 5) is discussed. A specific type of cluster technique (neighbouring cluster analysis), which takes into account the diachronic ordering of data, is introduced. Finally, the chapter presents a method for statistical identification of peaks and troughs in diachronic data and an extension called Usage Fluctuation Analysis (UFA). The peaks and troughs technique applies a non-linear regression model to the data to identify extreme points in time (outliers), where a dramatic change in discourse occurred. UFA traces the development of collocates of a word of interest over time; using the peaks and troughs technique it identifies points where a major change in the use of a word took place.

We'll be exploring answers to five questions:

- How can we measure and visualize language development over time (Section 7.2)?
- How can we evaluate if a change has occurred over time (Section 7.3)?
- How can we statistically group time periods (Section 7.4)?
- How can we model and visualize change in discourse (Section 7.5)?
- How can the techniques discussed in this chapter be used in research (Section 7.6)?

7.2 Time as a Variable: Measuring and Visualizing Time

Think about ...

Before reading this section, look at Figure 7.1. What does it suggest about the development of modal verbs in twentieth-century English?

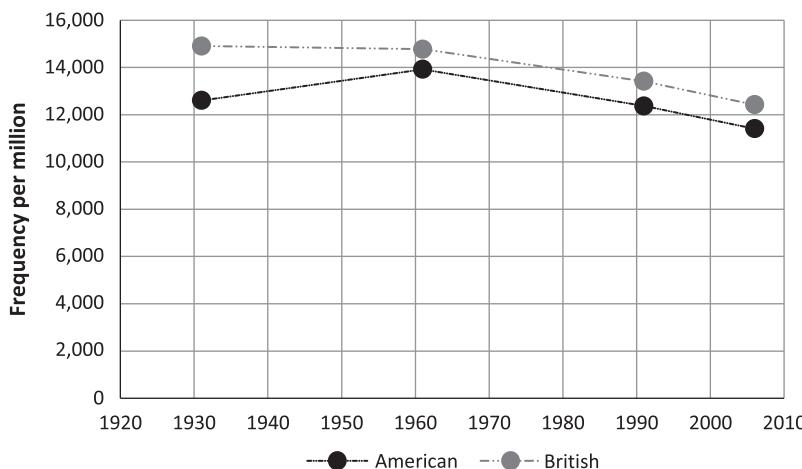


Figure 7.1 *Modals in the Brown family corpora*

The notion of **change over time** is central to the analyses described in this chapter. From a statistical perspective, time is a continuous (scale) variable; this means that we can measure time on a continuum of centuries, decades, years, months, weeks, days, hours, minutes, seconds, milliseconds etc. A study that involves time as a variable is called a **diachronic** or **longitudinal study**. We sometimes also measure time relative to the stages of the human life, that is **age**, which is an important variable in sociolinguistic (see Chapter 6) and language acquisition studies. While the majority of diachronic studies focus on change in language, we should also not forget the flipside of language change, **stability over time**, which is equally interesting. Language is like a large brick house, with many levels, which is undergoing constant renovation on different levels; for some bricks to be replaced, others need to remain in the structure to keep the house standing. Baker (2011) in his study of change in British English over the course of the twentieth century focuses both on words that have changed their frequency and/or meaning as well as on words that remained stable. Baker (2011) calls these words that remain stable in their usage **lockwords** and provides *times* and *money* as examples. From a statistical perspective, the main challenge here is how to identify stability in word frequencies in different corpora, in other words, how to identify lockwords. This is because the whole scientific and statistical paradigm focuses on identification of differences and change, while stability is often overlooked. Moreover, the inferences made in statistics are framed in such a way that we can never accept the null hypothesis, which typically states that there is no difference or change; we simply say that we don't have enough evidence to reject the null hypothesis, because the lack of evidence for rejecting it doesn't imply that it is true – as a parallel, think of a court case, where the verdict 'not guilty' does not necessarily imply that a person did not commit the crime; it

means that the prosecution did not succeed in proving it (see Section 1.3). To identify stability, we therefore need to draw on resources of descriptive statistics and a good understanding of the data to identify stable elements in language use (see the discussion of lockwords in Section 3.4).

To carry out diachronic studies, we need appropriate corpora, which we call historical or diachronic corpora. **Diachronic corpora** sample different stages of language or discourse development across time. Examples of historical corpora for the English language include the Brown family,¹ the *Helsinki Corpus of English Texts*, the *Corpus of Early English Correspondence* (CEEC), the *Corpus of Historical American English* (COHA) and the *Early English Books Online* (EEBO) corpus. These corpora differ in size from several million tokens (e.g. the Brown family, the *Helsinki Corpus*, CEEC) to billions of words (EEBO). In historical corpora, as with any corpus research, we need to critically evaluate both the quality and quantity of linguistic evidence available. Historical data represent an extra challenge due to the diachronic dimension which brings additional sources of variation. Building on Leech (2011), who discusses the development of modal verbs (see the ‘Think about’ task’),² we should consider three important aspects of corpus-based historical linguistics that need to be critically evaluated: (i) the diachronic representativeness of corpora, (ii) alternative interpretations of linguistic development and (iii) fluctuation of the meaning of linguistic forms. Let us consider these three areas one by one. First, **diachronic representativeness** is a characteristic of a historical corpus which allows it to systematically reflect the population (language use) over time. A corpus is a sample of language typically designed to show characteristics similar to those of the larger population, which, for a diachronic corpus, is in a process of change (see Section 1.4). This change involves the emergence of new words, grammatical features or even genres. Diachronic representativeness may be more important than the sheer size of the corpus. For example, Leech (2011) shows that for tracing the development of a grammatical feature such as the modal verbs, the Brown family of corpora, which represents a range of genres of written language, provides a more accurate picture of language change than a much larger corpus based on a single source such as the American *TIME* magazine;³ the *TIME Magazine Corpus*, due to its limited representativeness, is much more likely to reflect local changes involving different editorial practices over time. The results based on this corpus thus cannot be, without further evidence, generalized to

¹ The Brown family corpora were originally designed as synchronic corpora. Because each follows the same sampling frame (see Section 1.4) and all were sampled at different points in time, we can use the whole Brown family as a diachronic corpus – see the examples in this section and Baker (2011, 2017).

² This article is part of a very interesting debate between Leech (2003, 2011) and Millar (2009), where a number of interesting questions related to the analysis of corpora and language change are raised.

³ The *TIME Magazine Corpus*, which has been used for diachronic exploration of English, consists of 100 million words from *TIME* magazine from 1923 to 2006 and is available e.g. at <http://corpus.byu.edu/time>

permit discussion of the overall development of the English language in the twentieth century.

In addition, when considering diachronic representativeness, we need to deal with certain limitations inherent in historical data. We have to realize that only a small fraction of the language from the past has been recorded and preserved. A historical corpus is usually not a balanced sample of the language from a given historical period; it is, rather, a narrow lens that provides an insight into the language that has been preserved (McEnery & Baker 2017). Also, historical corpora are almost entirely corpora of written language. The reason for this is obvious: unless recorded in writing (which is rare), historical spoken language is inaccessible to us. The first audio recordings of spoken language date from the late nineteenth century. Nevertheless, some written genres such as personal letters reflect semi-formal and informal use of language and can even be subjected to variationist sociolinguistic analysis, which is more typically performed with spoken data (Nevalainen & Raumolin-Brunberg 2003: 26). Furthermore, the writing that gets preserved typically represents books, pamphlets and official documents written by a small minority of the people living at the period – those who were literate and educated. Our linguistic samples of historical data are thus biased both in terms of the genres represented and the social composition of the authors (Nevalainen & Raumolin-Brunberg 2003). Drawing attention to these issues with historical corpora, Nevalainen (1999) suggests that any historical linguistic pursuit is ultimately about ‘making the best use of bad data’.

Second, we need to consider **alternative interpretations of linguistic development** based on the evidence available. If there is not enough evidence, or the evidence is biased towards, for example, a certain genre or certain types of speakers/writers (see ‘diachronic representativeness’ above), the patterns we observe are ambiguous. Generally, the further back in history we go the less evidence has been preserved. For example, all preserved texts from the Old English period (450–1100) consist of no more than three million running words,⁴ while from the seventeenth century onwards the corpora include billions of words (EEBO v. 3, as used by McEnery & Baker (2017), has almost one billion words for the seventeenth century alone). In this situation, we need to resist the temptation to supply our interpretation in place of the missing evidence; good practice dictates that we should separate evidence (data) from its interpretation. Returning to Leech’s (2011) discussion, let’s consider the issue of two historical sampling points (1961 and 1991) in relation to our ability to observe and interpret a language change between these points. In Figure 7.1, these two historical points are linked with a straight line and further sampling points (1931 and 2006) confirm this trend. However, we need to realize that the straight line between two data points is already part of an interpretation of the data, which may or may not be accurate. The fewer historical sampling points we have the more akin to conjecture this

⁴ This word count is based on *The Complete Corpus of Old English* (Healey 2004).

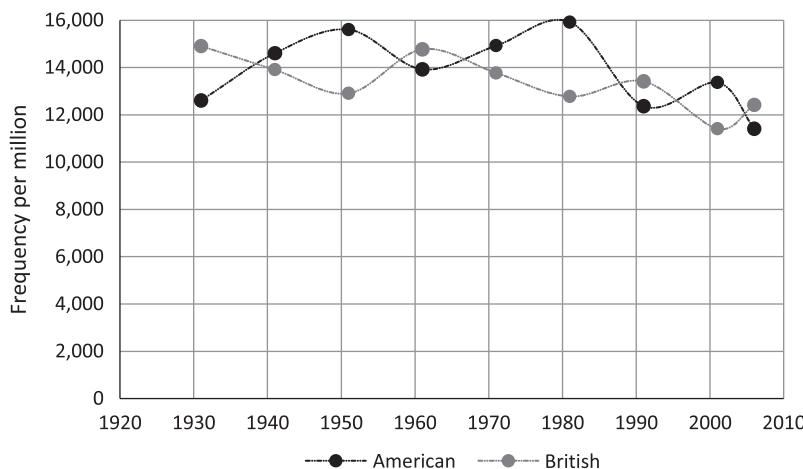


Figure 7.2 *Modals in the Brown family corpora: an alternative interpretation*

interpretation is. Consider the alternative interpretation (one of many) in Figure 7.2 of the same data points as in Figure 7.1.

The issue can be mitigated by adding more data points which allow us to track the trajectory of language change more precisely. However, where acquiring more data points like this is not possible, we need to be open to alternative interpretations. To capture some language changes a straight line (the simplest model) is adequate. Yet for other diachronic processes, more complex models (complex curves) are more appropriate. For example, phonological and grammatical changes often follow an S-shape curve (Nevalainen & Raumolin-Brunberg 2003: 53–5; Blythe & Croft 2012), while many lexical and discourse changes fluctuate as peaks and troughs (see Section 7.5).

Third, the **fluctuation of the meaning of linguistic forms (diachronic polysemy)** is a phenomenon that needs to be carefully considered when looking at the development of language; the same linguistic form often changes meaning (or its set of meanings) over time, so in diachronic analyses we also need to provide an account of the semantic development, not only an overview of changing frequencies of linguistic forms. For example, Leech (2011) explains that in both British and American English the same modal form *may* is increasingly more often used with the epistemic meaning of possibility as in *She may be right*, and increasingly less with the deontic meaning of permission as in *Please may I finish* (with *can* typically used instead of *may*). When looking at the historical development of modals, we therefore also need to report on this semantic development to provide the full picture.

The sensitivity to the three areas discussed above, that is the diachronic representativeness of corpora, alternative interpretations of linguistic development and fluctuation of the meaning of linguistic forms, distinguishes corpus linguistics from linguistically naïve quantitative methods of the ‘big data’

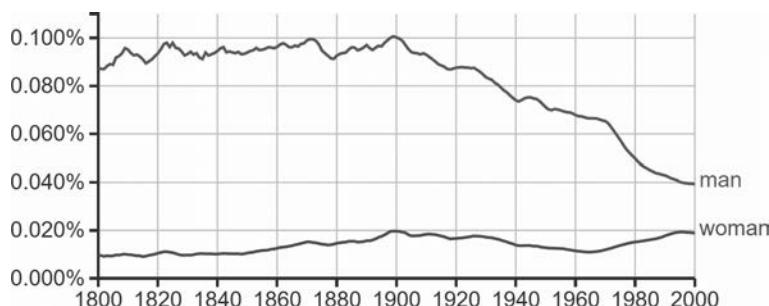


Figure 7.3 *Google n-gram viewer: ‘man’ and ‘woman’*

approaches to language, such as culturomics (Michel et al. 2011); these approaches often present sweeping accounts of the historical development of words and concepts with little or no attention paid to important linguistic clues and principles of good science. For example, it is very easy to produce a graph of an apparent linguistic/social change of the use of the words *man* and *woman*. However, it is much more problematic to interpret this evidence if we don’t have access to the actual examples of use (concordance lines) of these linguistic forms and the information about the changing composition of the corpus. In Figure 7.3 we can see the changing frequencies of the form *man* (top line) contrasted with the frequencies of the form *woman* (bottom line) in the period of 1800–2000. We can see that *man* decreases over time while *woman* increases, which could be interpreted as a closing of the gender-inequality gap in the discourse. This development may, however, be related to the changing composition of the corpus and cultural and semantic shifts or a combination of various other factors – we simply don’t know. From this, we can draw a general principle: an analysis that does not engage with the diachronic data in a way that ensures transparency and accountability as well as comparability across time is meaningless.

We can see how culturomics and similar approaches trade the representativeness and sensitivity to meaning fluctuation for a large amount of evidence (hence the ‘big data’ approaches); this, however, is never a good deal (Leech 2011; Pechenick et al. 2015; McEnery & Baker 2017: section 1.3).

Finally, let us look at some visualization options with diachronic data. We have already seen three figures (7.1, 7.2 and 7.3) that visualize language change in the form of a line graph. A **line graph** is a simple display, which plots the time variable on the x-axis and the frequencies of linguistic variables on the y-axis. Line graphs help us interpret the patterns of change in corpora. Note that the actual line in the graph is produced by connecting the data points and already represents an interpretation of the data. For this reason, in Figures 7.1 and 7.2 (but not in Figure 7.3) this line is made

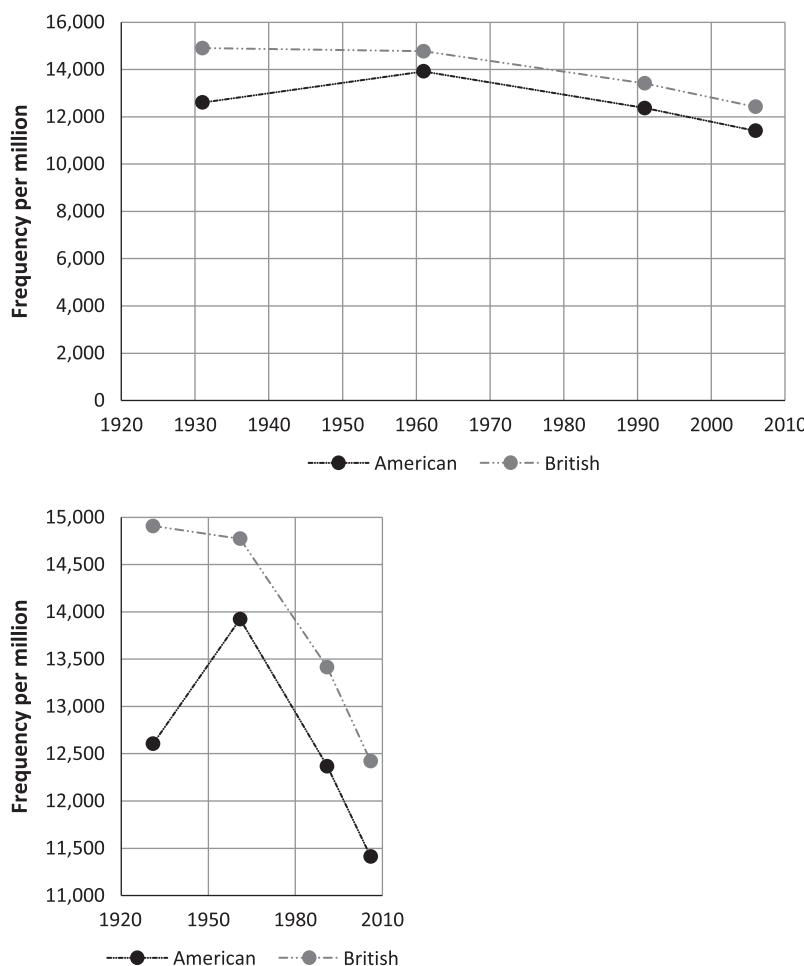


Figure 7.4 *Modals in the Brown family corpora: original (top) and rescaled (bottom)*

distinctly tentative using the dash-dotted line type. In this way, the data points stand out as the solid anchors of evidence, while the overall tendency is only tentatively suggested by the thin dash-dotted line. In addition, when dealing with line graphs, we have to bear in mind that a change or fluctuation in historical data can appear more or less dramatic depending on the scale of the axes and the starting point (origin) of the graph. To illustrate this, the graphs in the top panel and the bottom panel of Figure 7.4 represent exactly the same data (circles) with exactly the same interpretation (dash-dotted line). The only difference is the scale of the axes. Note also that in the bottom panel the y-axis does not start at 0 but at 11,000 and goes up to 15,000 rather than 16,000.

The message from the comparison in Figure 7.4 is this: it is imperative to be mindful of the scale we use in the graph and use exactly the same scale when comparing multiple graphs. Although the scaling of the axes and a graph's aspect ratio (height: length) might seem somewhat arbitrary from the data-plotting perspective (both panels in Figure 7.4 show accurately the same data), these considerations matter for visual decoding of the information – that is for the understanding and correct interpretation of the graph. We can imagine the graph in the bottom panel as a zoomed-in and condensed version of the graph in the top panel; it can help us notice more subtle patterns in the data – Cleveland (1994: 66ff.) demonstrates that graphs which are designed not to be too flat or too spiky with slopes averaging 45° like the graph in the bottom panel are optimal for interpretation. However, we also need to avoid the temptation to artificially ‘blow up’ the effects observable in the graph that are not really present in the data.

Other options for visualizing diachronic data include boxplots and error bars, sparklines and the candlestick plot. **Boxplots and error bars** were discussed in detail in Section 1.5. They offer an opportunity to look inside each diachronic sampling point and analyse variation between individual texts in the historical period. The error bars can be used to display 95% confidence intervals around the mean values for each historical period. Figure 7.5 displays the same data as the top (lighter) line in the graphs in Figure 7.4, that is data for British English. Instead of each historical period being represented by one value (mean), internal variation among the texts in each period is shown.

When dealing with a large number of variables that develop over time, we can use sparklines as efficient summaries of multiple individual trends. A **sparkline** is a small graph the size of a single word that can be seamlessly incorporated into text (Tufte 2006: 46–63). For diachronic data, line sparklines are generally most useful. The following example provides a sparkline that shows the development of the modal *must* in the seventeenth century:

The use of *must* in the seventeenth century is marked by a large amount of fluctuation 

In addition to the overall trend, the sparkline can provide information about the minimum and the maximum value (solid points above and below the line). It is thus an informationally rich form of display that takes incomparably smaller space than if we were to describe the data in words. The sparkline in the example above displays 100 different sampling points.

An alternative to multiple sparklines is a candlestick plot. A **candlestick plot** is a type of data visualization, where the development of a linguistic

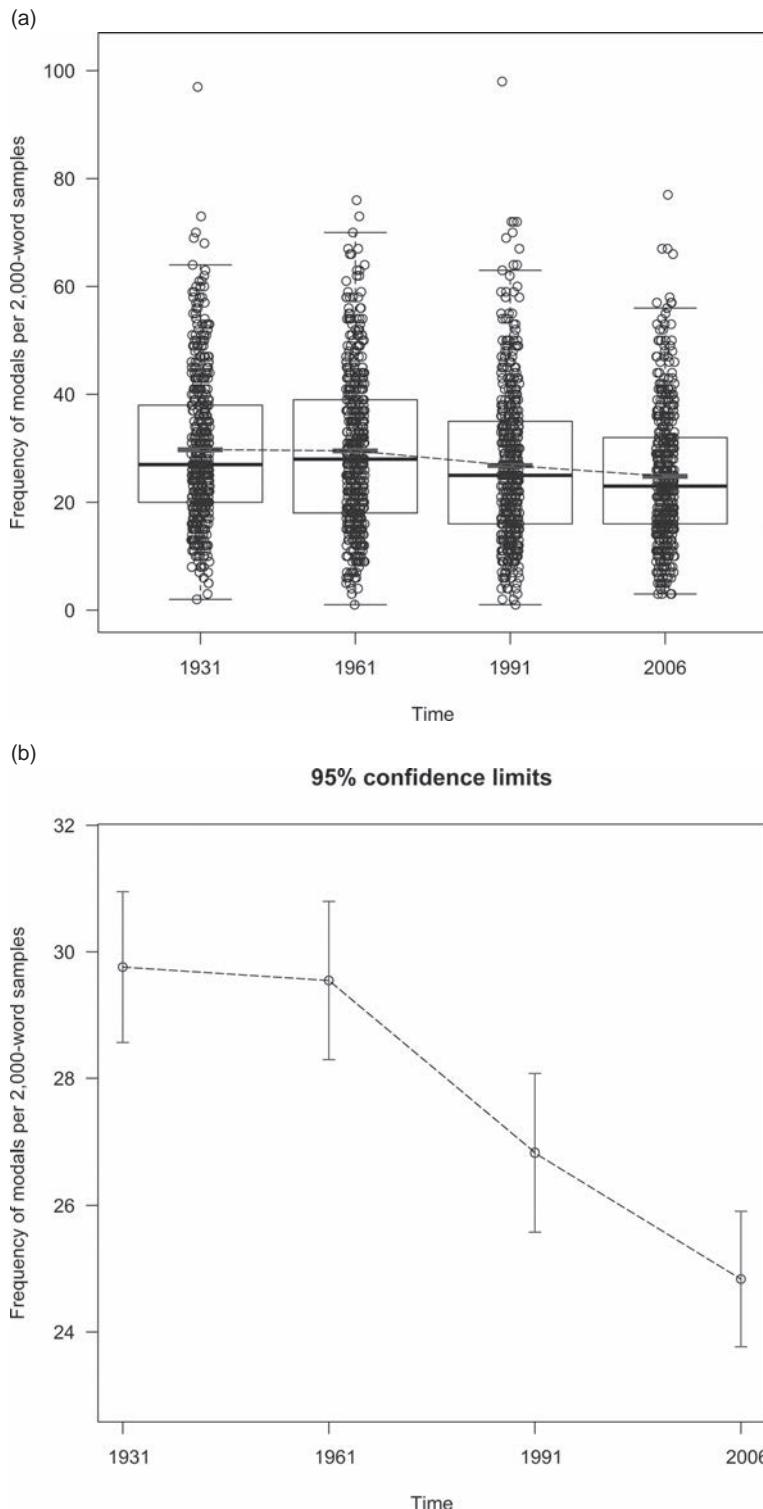


Figure 7.5 *Modals in British English: (a) boxplots; (b) 95% CI error bars*

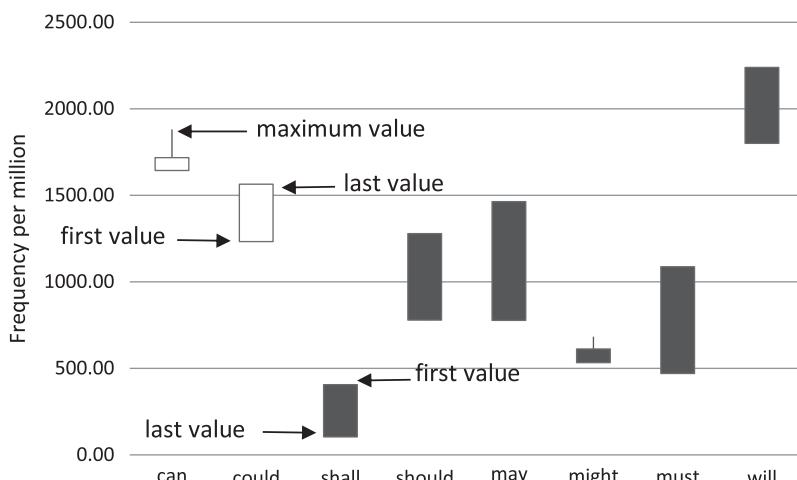


Figure 7.6 *Candlestick plot: the development of individual modals 1931–2006*

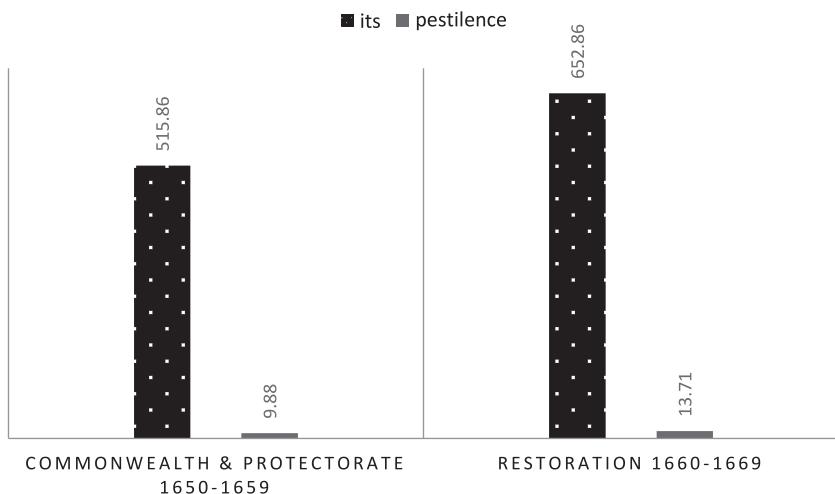
variable is summarized as a box and ‘wicks’ resembling a candlestick (hence the name of the graph); although superficially similar to a boxplot, the candlestick plot works on different principles. It is often used in financial reports, where the development of stock prices is traced over time. Figure 7.6 shows an example of a candlestick plot, which indicates the development of different modal verbs in the period of 1931–2006. Each ‘candlestick’ shows the initial point in the time series, the final point and the minimum and the maximum value. The initial and the final values are indicated by the box; if different from the initial and the final values, the maximum and minimum values are indicated by the wicks projected outside the box. In addition, the colour of the box (filled vs unfilled) shows the direction of the change: a filled box shows decrease, unfilled increase. In Figure 7.6, we can see immediately that in this period, two out of the eight modal verbs increased their frequency (this is indicated by the unfilled box), while the rest have decreased. With the exception of two cases where the maximum value is outside the box, the increase and decrease was ‘smooth’, i.e. without frequency fluctuation.

In sum, the analysis of historical language data presents a challenge, which, in addition to the usual methodological considerations in corpus research, has special requirements connected with the diachronic dimension. Effective visualization using line charts, boxplots, error bars, sparklines, candlestick plots etc. is a useful starting point for any diachronic investigation.

7.3 Finding and Interpreting Differences: Percentage Change and the Bootstrap Test

Think about ...

Before reading this section, look at the following bar chart and compare the use of the possessive pronoun *its* and the use of the term *pestilence* in two historical periods. What can you conclude?



Broadly speaking, the most important question that we have to deal with when comparing two corpora is whether the frequencies of the relevant variables are significantly larger or smaller in corpus 1 than in corpus 2. In addition, when dealing with diachronic comparisons, we need to assess whether the observed differences in frequencies are related to change in the discourse/language over time or whether these are related to other sources of variation. Gablasova et al. (2017a) show that even corpora sampling the same type of language often differ significantly and we therefore need to be very careful about how we interpret the differences. Even more caution needs to be exercised when dealing with historical corpora with typically opportunistic sampling (see the ‘bad data’ problem discussed in Section 7.2). Let us look at some examples from the EEBO v. 3 corpus and possible ways of analysing and interpreting the data. Table 7.1 lists four linguistic variables (*its*, *must*, *time(s)* and *pestilence*) and provides a comparison between two sampling points in the EEBO corpus that reflect two historical periods in the seventeenth century: (i) Commonwealth & Protectorate and (ii) Restoration. In Britain, the second half of the seventeenth century was marked by great political turmoil and

Table 7.1 *Comparison of two periods in the EEBO corpus: Commonwealth & Protectorate and Restoration*

Linguistic feature	Corpus 1 – Commonwealth & Protectorate (1650– 9) corpus size: 168,912,439	Corpus 2 – Restoration (1660– 9) corpus size: 111,998,646	Percentage increase/decrease
<i>its</i>	515.86 ^a	652.86	+27
<i>must</i>	1,173.02	1,135.67	-3
<i>time(s)</i>	1,445.57	1,355.84	-6
<i>pestilence</i>	9.88	13.71	+39

^a relative frequencies per million

many social changes, which, we can hypothesize, affected discourse and the language itself.

To be able to see the difference between the two sampling points a simple percentage increase/decrease was calculated. **Percentage increase/decrease** is a statistic that indicates by how many percentage points the value of a particular linguistic variable increased or decreased between two time periods. It is calculated using the equation below.

$$\% \text{ increase/decrease} = \frac{\text{relative frequency in corpus 2} - \text{relative frequency in corpus 1}}{\text{relative frequency in corpus 1}} \times 100 \quad (7.1)$$

where corpus 1 is a corpus from an earlier period and corpus 2 is a corpus from a later period.

In Table 7.1, percentage increase/decrease for the four linguistic variables is indicated in the last column. The following calculation demonstrates how the percentage increase/decrease was calculated for the first item in the table, possessive pronoun *its*.

$$\text{percentage increase/decrease of } \textit{its} = \frac{652.86 - 515.86}{515.86} \times 100 = 26.6\% \quad (7.2)$$

Overall, Table 7.1 shows two cases of increase and two cases of decrease. In two cases (*its* and *pestilence*), the increase seems to be fairly large (27% and 39% respectively). To be able to say which of the cases of increase and decrease are due to chance (sampling error) and which are statistically significant, we can use a statistical test which evaluates this question. Although often a chi-squared test or the log-likelihood test are used in this situation, these tests are not appropriate for this type of comparison (Lijffijt et al. 2012, 2016; Brezina & Meyerhoff 2014). Instead, we can use a test which takes into account the distribution of the linguistic variables in

corpus	Text A Freq: 10	Text B Freq: 5	Text C Freq: 15	Text D Freq: 1	Text E Freq: 20	mean = 10.2
resampling cycle 1	Text A Freq: 10	Text A Freq: 10	Text B Freq: 5	Text D Freq: 1	Text E Freq: 20	mean ₁ = 9.2
resampling cycle 2	Text A Freq: 10	Text A Freq: 10	Text A Freq: 10	Text B Freq: 5	Text B Freq: 5	mean ₂ = 8.0
[...]						
resampling cycle n	Text C Freq: 15	Text D Freq: 1	Text D Freq: 1	Text E Freq: 20	Text E Freq: 20	mean _n = 11.4

Figure 7.7 Bootstrapping: demonstration of the concept

the individual texts in the two corpora (see ‘Individual text/speaker design’ in Section 1.4). The options that we have for the test include the t-test, the Mann–Whitney *U* test (both discussed in Section 6.3) and the bootstrap test (see below). Bootstrapping methods are a relatively new addition to the statistician’s toolbox (because they require sufficient computational power which wasn’t previously available (Efron 1979)) and offer powerful ways of assessing variation and estimating statistical significance (Efron & Tibshirani 1994; Chernick & LaBudde 2014). **Bootstrapping** is a process of multiple resampling, which often happens thousands of times, with replacement of the data – this means we take a random sample of texts from a corpus in such a way that each text can occur multiple times in the sample because we ‘replace’ it (i.e. place it to the pool again) once it has been taken. In each resampling cycle, we note down the value of the statistic (e.g. mean frequency of a linguistic variable) we are interested in; this gives an insight into the amount of variation in the data and gives us the confidence to generalize from this sample. As an example, take a corpus that consists of five texts A, B, C, D and E, each text having a particular frequency of the linguistic variable of interest (see Figure 7.7).

The name ‘bootstrapping’ refers to the idiom ‘pull oneself up by one’s bootstraps’ referencing ‘the self-help nature of the bootstrap algorithm’ (Efron 1979: 465), which uses the data present in the same sample multiple times to make an estimate about the value of interest in the population.

In this section, we will focus on the bootstrapping test proposed in Lijffijt et al. (2016). The **bootstrap test** is a non-parametric test of statistical significance (that means that it does not make an assumption about the

distribution of the linguistic variable in the population and can be used even with extremely skewed distributions), which compares two corpora and computes the p-value associated with the comparison. The test works on the principle of bootstrapping – multiple resampling with replacement – of two corpora: corpus 1 and corpus 2. Lijffijt et al. (2016) provide the following equation for the test:

$$p = \frac{1 + 2 \times \text{number of bootstrapping cycles} \times (p_1 \text{ or } 1 - p_1, \text{ whichever is smaller})}{1 + \text{number of bootstrapping cycles}} \quad (7.3)$$

where

$$p_1 = \frac{\text{For all bootstrapping cycles sum of value H}}{\text{number of bootstrapping cycles}}$$

and where H can be 1, 0.5 or 0 depending on the following conditions:

- H = 1 if value of interest in resampled corpus 1 > value of interest in resampled corpus 2
- H = 0.5 if value of interest in resampled corpus 2 = value of interest in resampled corpus 1
- H = 0 if value of interest in resampled corpus 2 > value of interest in resampled corpus 1

The idea behind the test is very simple: we compare across a large number of bootstrapping cycles the resampled corpus 1 and the resampled corpus 2 and look for a consistent difference between the resampled corpora, which would produce a low p-value (statistical significance). A low p-value is returned if in all or most cases resampled corpus 1 is either larger (we add 1 in the equation above) or smaller than corpus 2 (we add 0). For the test to return reliable results, enough bootstrapping cycles need to be allowed; 1,000 and above is sufficient.

In practice, when performing the bootstrap test, we need to trace the distribution of the linguistic variable in individual texts and normalize the frequencies – that is, we need to get relative frequencies (see Section 2.3). An example of such a dataset is provided in Figure 7.8 (only the first ten cases out of thousands are displayed). This data is first inserted into the equation (7.3). The output is a p-value, which provides an indication of the statistical significance of the comparison.

Table 7.2 displays the results (p-values) of the bootstrap test for the comparisons from Table 7.1, where we traced the development of four linguistic variables in the seventeenth century. We can now conclude that three of the comparisons are statistically significant; that means that the p-value is smaller than 0.05. In addition to the p-value, we can list the size of the change (effect size). This can be done by computing a standardized effect size measure such as

ID	1650_59	1660_69
1	0	0
2	662.01	0
3	191.36	0
4	1625.28	0
5	475.62	0
6	338.29	0
7	326.26	377.42
8	110.05	3062.48
9	0	1059.2
10	0	1236.4

Figure 7.8 Example of a dataset for the bootstrap test: *its* in EEBO

Table 7.2 Comparison of two periods in the EEBO corpus: results of the bootstrap test

Linguistic feature	p-value based on the bootstrap test (10 k samples)	Significant? (<0.05)	Effect size: Cohen's <i>d</i> and 95% CI
<i>its</i>	0.0001	YES	0.11 [0.15, 0.07]; MIN. EFFECT
<i>must</i>	0.0061	YES	0.06 [0.1, 0.02]; MIN. EFFECT
<i>time(s)</i>	0.8070	NO	0.01 [0.04, -0.05]; MIN. EFFECT, CI includes zero
<i>pestilence</i>	0.0001	YES	0.11 [0.15, 0.07]; MIN. EFFECT

^a raw frequency (relative frequency per million)

Cohen's *d* or robust Cohen's *d* (see Sections 6.3 and 8.4) and provide 95% confidence intervals (CIs) for this metric. Cohen's *d*, or other effect sizes, can also be estimated using the bootstrapping technique.⁵

Finally, we need to decide which of the statistically significant differences between the two corpora representing the periods of 1650–9 and 1660–9 are due to a diachronic change in language and which are due to other factors. This can be done by critically evaluating (i) the corpora in terms of their representativeness and comparability, (ii) the results in relation to other data sources and/or the previous research and (iii) the results in relation to a larger diachronic picture, if available. As an example, Table 7.3 briefly evaluates these three aspects for the results of statistical analyses from Table 7.2.

⁵ Note that the effects reported in Table 7.2 are much smaller than the effects based on overall percentage change reported in Table 7.1; this is because the bootstrapping technique (unlike the overall percentage change) takes into consideration individual variation.

Table 7.3 *Final evaluation of the results: its, must and pestilence*

	(i) Corpus representativeness and comparability	(ii) Other data sources	(iii) Larger diachronic picture
<i>its</i>			
<i>must</i>	The EEBO corpus samples printed books. It is therefore reflective of the publication practice at a particular historical period (e.g. some texts were created earlier than they were printed or were printed at later stages with modifications); the fluctuation in the genre composition also needs to be considered.	Increase in the use of <i>its</i> in the seventeenth century has been reported in the literature based on other corpora (e.g. Nevalainen & Raumolin-Brunberg 2003). The change in the frequency of modals in the seventeenth century has not been reported.	
<i>pestilence</i>			The temporal increase in the 1660s possibly reflects the Great Plague of 1665–6. A competition with other naming terms e.g. <i>plague</i> and <i>black death</i> needs to be investigated.

Reporting Statistics: Bootstrap Test

1. What to Report

The only output of the bootstrap test is the approximated p-value, which depends on the number of samples. Both the p-value and the number of samples should be reported. According to convention, a p-value larger than 0.001 should be reported as an exact value; p-values smaller than 0.001 are reported as $p < .001$. In addition, the effect size (Cohen's d) and the 95% confidence interval for the effect size should be reported.

2. How to Report: An Example

- The bootstrap test (Lijffijt et al. 2016) was used to compare the corpora representing two diachronic sampling points (1650–9 and 1660–9). The difference in the use of the pronoun *its* was statistically significant: $p < .001$, 10 k samples. The effect size (Cohen's d) was very small: $d = .11$, 95% CI [.15, .07];

7.4 Grouping Time Periods: Neighbouring Cluster Analysis

Think about ...

Before reading this section, think about the following question: how would you group the data points in Figure 7.9 according to their similarity?

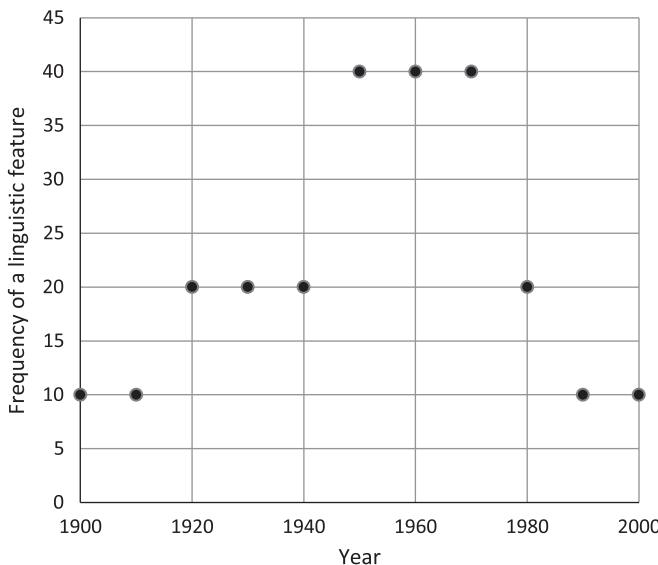


Figure 7.9 Data points over time: an invented example

The general principles of cluster analysis were explained in Section 5.3. In brief, cluster analysis compares distances between individual data points and groups the data points that are closest together creating a new unit (a new data point); this process is repeated until every data point is subsumed under one large cluster. This process helps us identify structure in the data based on similarity of e.g. texts, speakers or genres. With historical analyses, we bring an additional dimension to the cluster analysis, which is time. Time provides an extra layer of structure because it orders our data according to the temporal sequence, one event following another. Depending on our research question we can decide if we want to include the temporal sequence as the organizing principle in the cluster analysis. If the research question is very general (e.g. *Which historical periods are similar in the use of the target linguistic variable(s)?*), we can apply the simple hierarchical agglomerative cluster analysis (see Section 5.3). If, on the other hand, we are interested in identifying larger continuous periods of time based on the similarity of use of the target linguistic variable, we need to consider a cluster procedure that respects the time ordering of individual data points. In this case, we can use a procedure suggested by Gries & Hilpert (2008, 2010) called the variability-based neighbour clustering. **Variability-based neighbour clustering (VNC)** computes similarity only between temporally adjacent data points (e.g. individual years that follow each other) and then merges those that are most similar. This is repeated until all data points are merged into one cluster. The difference between hierarchical agglomerative clustering and VNC is illustrated in Figure 7.10. The data points from the dataset introduced in the ‘Think about’ task can either be grouped together purely on their similarity disregarding the temporal ordering (left panel) or merged as adjacent data points (right). Thus, for example, the hierarchical agglomerative clustering procedure groups together the years 1900, 1910, 1990 and 2000 because they have the same frequency of the target linguistic variable (10). VNC, on the other hand, takes these four years and merges only 1900 and 1910 in one group and 1990 and 2000 in another group, following the principle of adjacent mergers. The resulting dendograms for both procedures can be seen in Figure 7.11. These are alternative graphical displays of the clusters from Figure 7.10.

In VNC, as in the hierarchical agglomerative cluster analysis, two main decisions need to be made about the settings used in the clustering procedure. These are (i) the distance (or similarity) measure and (ii) the method of merging two data points (amalgamation rule). There is a range of distance measures that can be used. Distance measures such as Euclidean, Manhattan or Canberra distance were discussed in Section 5.3. In addition, we can measure similarity in a broader sense by using correlation (Pearson’s or Spearman’s; see Section 5.2); this option is useful if we are not interested in the exact differences between the data points but wish to find similar trends in the dataset. When using correlation, the distance is measured

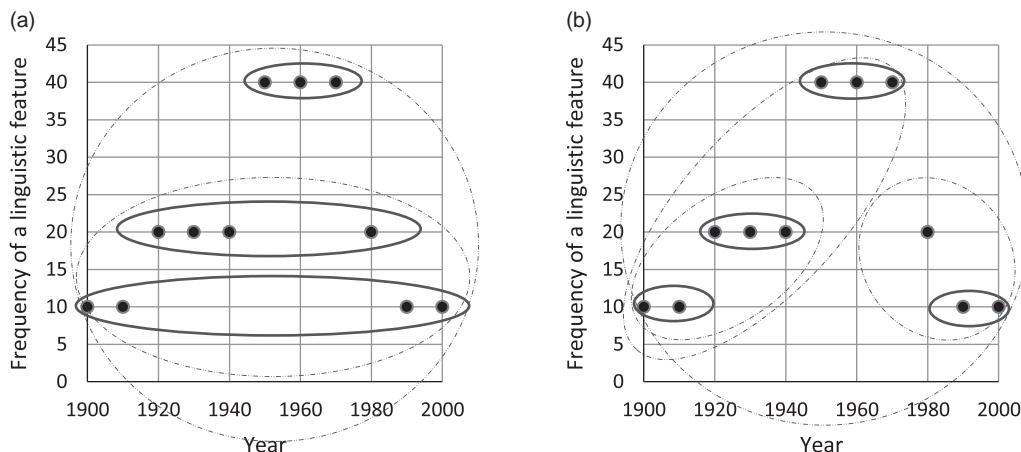


Figure 7.10 *Two clustering principles: (a) hierarchical agglomerative clustering; (b) variability-based neighbour clustering*

as 1 minus the correlation coefficient ($1-r$ or $1-r_s$) because correlation is high (close to 1) when the similarity is high. We can also use the standard deviation or the coefficient of variation (see Section 2.4) as a measure of distance; these measures express the distance of individual values from the mean. Coefficient of variation, in addition, expresses this distance in standardized units, thus making the values displayed on the y-axis of the dendrogram comparable across different linguistic variables. Next, the amalgamation rule depends on what we consider to be the new value of the merged data points. The options include single linkage (SLINK), complete linkage (CLINK), average linkage and Ward's method (see Section 5.3). The decision about these settings should be motivated by the aim of the analysis (research question) because it directly influences the result of the cluster analysis – the shape of the dendrogram. In corpus linguistics, more validation work is needed to empirically evaluate the effect of different settings in the procedure. With VNC, Gries & Hilpert (2008, 2010), Hilpert & Gries (2009) and Hilpert (2011) have used the following options:

- (i) Standard deviation, coefficient of variation, Pearson's correlation, corrected means ratio⁶
- (ii) Average linkage

VNC can be used for categorizing individual temporal data points (e.g. years, decades) into larger historical periods based on the similarity in language use. In this way, we can look at grammatical or lexical change and how it spread throughout time. For instance, we can observe the rise of the possessive pronoun *its* in the

⁶ The details of corrected means ratio, including the equation, can be obtained from Gries & Hilpert (2008: 73).

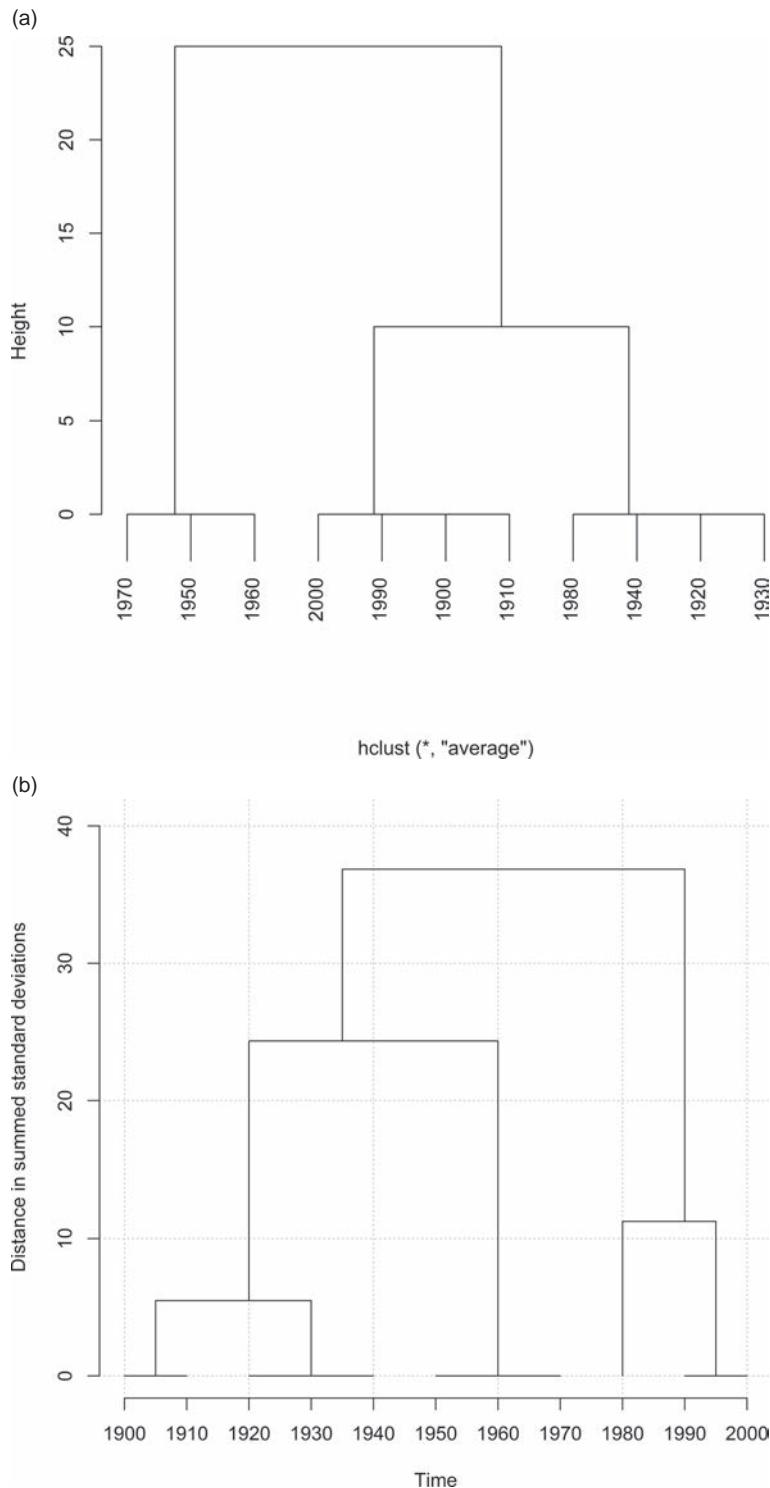


Figure 7.11 Dendograms: (a) hierarchical agglomerative clustering; (b) variability-based neighbour clustering

Table 7.4 *Relative frequency (per million) of the possessive pronoun its in the seventeenth century*

Decade	1600s	1610s	1620s	1630s	1640s	1650s	1660s	1670s	1680s	1690s
Relative freq.	9.39	23.4	77.71	161.93	309.79	515.86	652.86	694.93	673.6	751.37

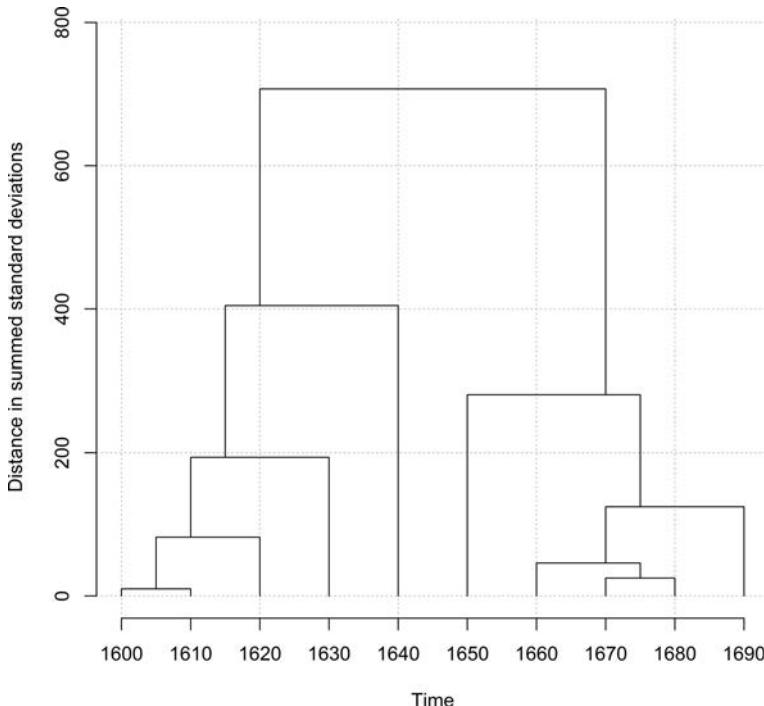


Figure 7.12 *Dendrogram: use of the possessive pronoun its in the seventeenth century*

seventeenth century and categorize the stages in the introduction of this form. This pronoun started to be used as a neuter version of *his*, which was previously used for both the masculine and neuter and competed with paraphrases such as *of it* and *thereof* (Nevalainen & Raumolin-Brunberg 2003: 62). Table 7.4 provides relative frequencies of *its* in the seventeenth century measured per decade.

The dendrogram based on Table 7.4 shows a clear split between two large periods: 1600s–40s and 1650s–90s (see Figure 7.12). If we want to subdivide these large periods further, we need to decide where the major differences lie. In the dendrogram, a difference between clusters is expressed as the height displayed by a vertical line in the plot. Because the comparison of the heights in Figure 7.12 might be difficult (especially with multiple data points), a scree plot somewhat similar to one used in factor analysis (Section 5.4) can help us decide on the major

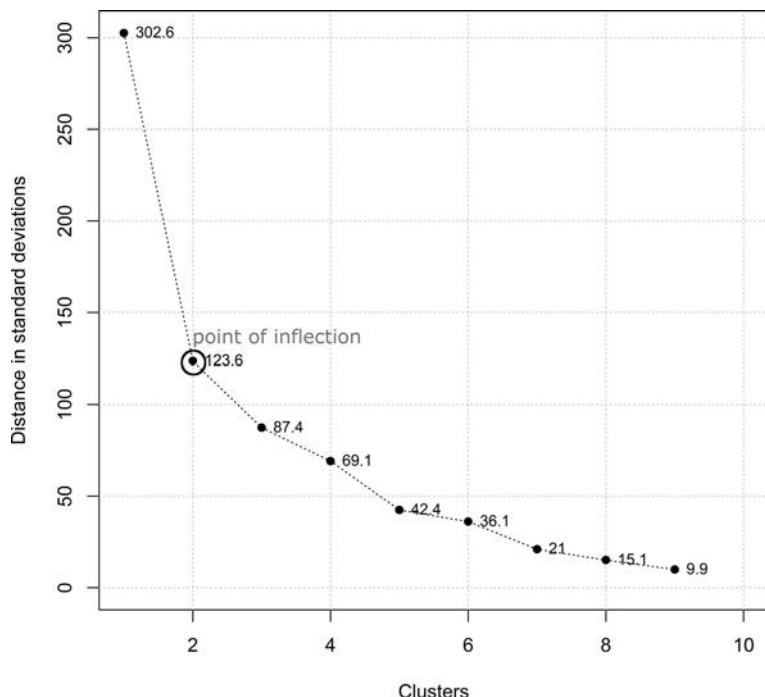


Figure 7.13 *Scree plot: use of the possessive pronoun its in the seventeenth century*

cluster subdivisions. The scree plot displays the number of clusters on the x-axis and the height (distance between the clusters) on the y-axis. The scree plot resembles a shape similar to the slope of a mountain with small stones (data points) covering the slope, hence the term ‘scree plot’. A steep slope indicates large differences between cluster groups, while a flat shape indicates small differences between cluster groups. In Figure 7.13, which presents the scree plot related to the dendrogram in Figure 7.12, we can see that after two large clusters the scree plot starts levelling off; we call this place the point of inflection. Using the scree plot, we can thus confirm that the major subdivision in the dendrogram is between two clusters, the first spanning between the 1600s and 1640s and the second between the 1650s and 1690s.

Reporting Statistics: Variability-Based Neighbour Clustering (VNC)

1. What to Report

VNC is largely an exploratory visual method to show diachronic segmentation of the data. Two parameters are used for cluster identification: (i) the distance measure and (ii) the amalgamation rule as well as the results of the analysis,

the dendrogram (tree plot), need to be reported. The main question to be addressed is: how many meaningful clusters representing continuous historical periods can be observed in the plot? A scree plot can be used to help determine the answer to this question.

2. How to Report: An Example

- The data was analysed using the variability-based neighbour clustering technique (Gries & Hilpert 2008) with standard deviation (SD) as the distance measure and average linkage as the amalgamation rule. The resulting dendrogram is available in Figure 7.12. A scree plot indicated two major cluster groups (1600s–40s and 1650s–90s).

7.5 Modelling Changes in Discourse: Peaks and Troughs and UFA

Think about ...

Before reading this section, look at the development of the frequencies of the word *war* in *The Times* newspaper between 1940 and 2009; *The Times* is a British daily newspaper, part of the ‘quality’ press. Can you link the observed peaks to historical events in the twentieth and twenty-first centuries?



Discourse develops dynamically. There are periods in which certain concepts and/or words come into prominence (or are even invented) and periods in

which these are in the background with other words and concepts foregrounded. Discourse is a linguistic marketplace where words and ideas expressed by these words compete for recognition and use. How should such a dynamic system be analysed when we consider its development over a period of time? Gabrielatos & Marchi (2012) introduced a technique of diachronic discourse analysis which they call the ‘wave, peak and trough (WPT) method’, but which is better known as simply ‘peaks and troughs’. **Peaks and troughs** is a method of analysing and visualizing diachronic data, which applies a non-linear regression model (specifically, Generalized Additive Model or GAM) to data points that show the development of a linguistic variable over time in order to identify statistically significant outliers – points of departure from the general trend. Unlike the more traditional type of regression analysis, which uses a straight line to show the main tendency in the data (see Section 1.2), peaks and troughs fit a curve to the data, which better reflects the rising and falling tendency (hence ‘peaks and troughs’) of diachronic development of discourse. The process of producing the peaks and troughs analysis consists of two obligatory steps and two optional ones:

1. **Obligatory:** Obtaining (relative) frequencies from the corpus of the linguistic variable of interest for each of the periods (e.g. years, decades etc.) covered by the analysis.
2. **Optional:** Computing differences between two consecutive values by taking value 2 and subtracting value 1; this is done to highlight high values preceded by low values (or vice versa) which indicate a more dramatic change.
3. **Optional:** Transformation of the values using binary logarithm (\log_2) to reduce extremes. This step is possible only if all transformed values are positive numbers because logarithm is not defined for negative numbers. Since step 2 typically produces also negative values, logarithmic transformation is possible with data from step 1.
4. **Obligatory:** Fitting a non-linear regression model (displayed as a curve in the graph), computing 95% and 99% confidence intervals (displayed as shaded areas around the curve) and identification of significant outliers – data points outside the confidence interval area.

As an example, let us look at the data from the ‘Think about’ task, which trace the occurrence of the term *war* used in *The Times* newspaper. Both optional and obligatory steps of the peaks and troughs procedure are shown in what follows.

1. **Obligatory:** For illustration, Table 7.5 displays the ten initial data points from the whole dataset, which covers the period 1940–2009; relative frequencies (see Section 2.3.) of the word *war* are used to account for the fact of unequal sample sizes over the period of time.
2. **Optional:** The differences between each two consecutive temporal points from Table 7.5 are displayed in Table 7.6. We can see that this step ensures that high

Table 7.5 *Relative frequencies (per million) of war*

Period	1940	1941	1942	1943	1944	1945	1946	1947	1948	1949
Rel. freq. <i>war</i>	1,473.69	1,609.75	1,623.78	1,505.46	1,283.21	1,299.81	851.03	590.37	479.7	423.14

values (e.g. 1,609.75) preceded by lower values (e.g. 1,473.69) obtain a larger score (136.6) than high values (e.g. 1,623.78) preceded by other high values (e.g. 1,609.75); the resulting score for the latter comparison is only 14.03.

Table 7.6 *Differences between relative frequencies of war*

Period	1940/1	1941/2	1942/3	1943/4	1944/5	1945/6	1946/7	1947/8	1948/9
value1 -value 2	136.06	14.03	-118.32	-222.25	16.6	-448.78	-260.66	-110.67	-56.56

3. **Optional:** The transformation of the values using binary logarithm (log2) is possible only for values from Table 7.5, which are all positive values. The log2 transformed values are shown in Table 7.7.

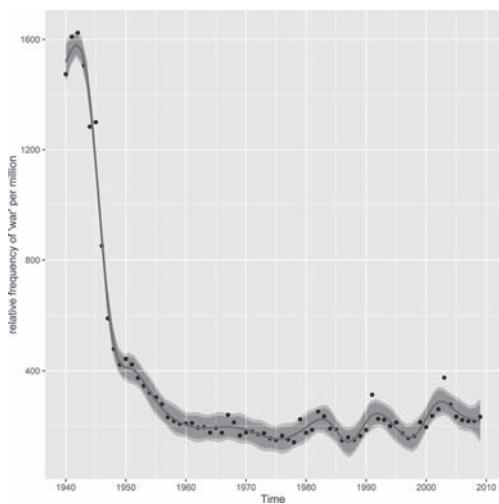
Table 7.7 *Log transformed relative frequency (per million) of war*

Period	1940	1941	1942	1943	1944	1945	1946	1947	1948	1949
Log2 of rel. freq. <i>war</i>	10.53	10.65	10.67	10.56	10.33	10.34	9.73	9.21	8.91	8.72

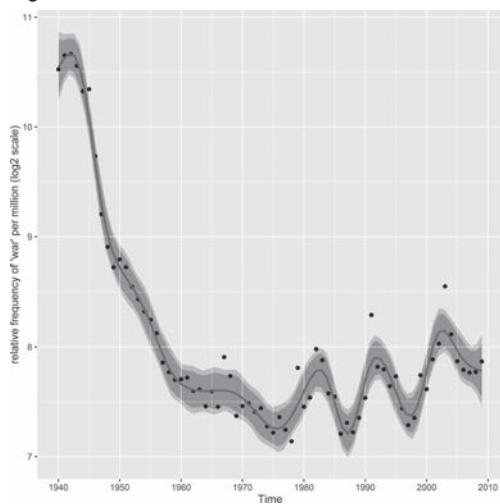
4. **Obligatory:** Finally, we can fit a non-linear regression model to the different versions of the datasets to demonstrate some of the options discussed above. For the model to work properly, we need a sufficient number of sampling points; the peaks and troughs graphs therefore display the whole dataset 1940–2009.

When we look at solutions (a) and (b) in Figure 7.14, we can see that they resemble the simple line graph in the ‘Think about’ task. In addition, the graphs in Figure 7.14 show a curve that best fits the data, 95% (darker area) and 99% (darker area plus the lighter rim) confidence intervals; these help us identify the outliers – points in time where the frequency of *war* is significantly higher or lower than predicted by the model (the line). With the use of the log transformation (solution (b)) some of the later contrasts become more distinct because the graph is not skewed so dramatically by the very high frequencies of the term *war* during the World War II period (1940–5 in the period covered by the graph). Solution (b) thus helps us identify four significant peaks and one trough (that is, points with significant outliers). These are 1967, 1982, 1991, 2003 (peaks) and 1978 (trough). When we look at the concordance lines we can see that the

(a) Based on relative frequencies of *war*; no transformation



(b) Based on relative frequencies of *war*; log transformation



(c) Based on subtracted frequencies of *war* in two consecutive sampling points; no transformation

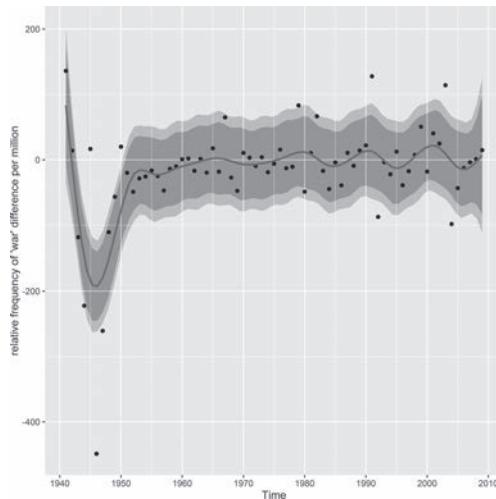


Figure 7.14 Resulting peaks and troughs graphs: settings as indicated

highlighted peaks are connected to the Third Arab-Israeli War (1967), the Falklands War (1982), the Gulf War (1991) and the war in Iraq (2003). On the other hand, the trough in 1978 is more difficult to explain; it can be understood indirectly with reference to the neighbouring years, especially 1979–82 where the mentions of *war* are much more prominent than in 1978. Let's now shift our attention to the final graph in Figure 7.14 (solution (c)), which highlights the differences between consecutive points in time. Here we can observe four peaks (1945, 1950, 1991 and 2003) and three troughs (1946, 1992 and 2004). The peaks

mark the end of World War II (1945), the Korean War and the Cold War (1950), the Gulf War (1991), and the war in Iraq (2003). The troughs are points with a significantly lower number of mentions of *war* relative to the preceding year and can thus be interpreted with reference to the immediately preceding peaks: the end of World War II (1946), Gulf War (1992) and war in Iraq (2004).

Finally, let us discuss an extension of the peaks and troughs technique called Usage Fluctuation Analysis (UFA) (Brezina et al. in prep.; Baker et al. 2017). **Usage Fluctuation Analysis (UFA)** is a method which investigates the development of the meaning of a particular word through analysing a changing profile of collocates around the word. UFA builds on the fact that without context the development of frequencies of a particular linguistic form is hard to interpret (see discussion of culturomics and ‘big data’ approaches in Section 7.2). UFA therefore systematically analyses the development in the most immediate contexts through collocation analysis (see Section 3.2 for the explanation of collocation) and identifies the points where a significant change in meaning occurs. It consists of the following steps:

1. Identification of collocates of a word of interest (node) across the time-series data.
2. Recursive estimation of the difference between collocates at any two consecutive points in time.
3. Use of the peaks and troughs technique (see above) to trace the points where major changes take place; these are identified as significant troughs because these are the largest points of meaning dissimilarity.

As an example, let’s use the data from the ‘Think about’ task again. For UFA, instead of looking at the frequencies of *war*, we’ll analyse the collocates of *war* in *The Times* for each period.

1. Table 7.8 provides a simplified picture of how the collocate profiles develop; in reality, we usually trace a much larger number of collocates. In Table 7.8, collocate A is a consistent collocate because it occurs throughout the period; B is a terminating (stops occurring) and F is an initiating collocate (starts occurring), while C and G are transient collocates, because they appear only for a short period of time (McEnery & Baker 2017: 1.6.1–1.6.4)
2. The best way to estimate the difference between two consecutive periods in terms of their collocate profile is to use a statistic for inter-rater agreement such as Gwet’s AC₁ (see Section 3.5). In this case, instead of asking how consistent two ratings are, we investigate how consistent two consecutive historical

Table 7.8 *Collocate profiles of war*

Period	1940	1941	1942	1943	1944	1945	1946	1947	1948	...	2009
Collocates	A	A	A	A	A	A	A	A	A	A	A
	B	B	B	F	F	F	F	F	F	F	F
	C	C	G	G	G	G	G	G	G	H	

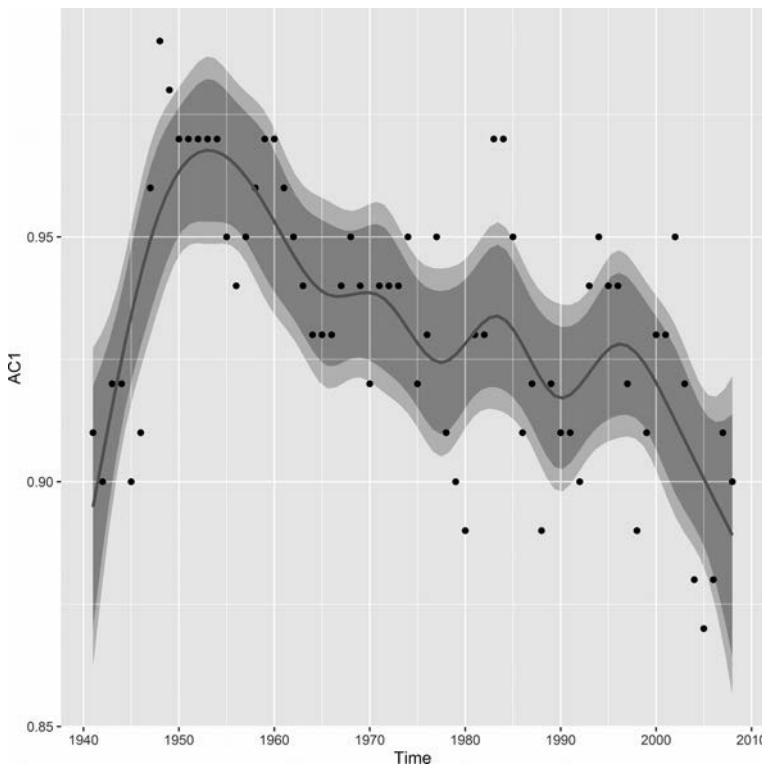


Figure 7.15 Results of UFA for war 1940–2009 (3a–MI(3), L5–R5, C10relative–NC10relative; AC₁)

periods are with respect to the collocates used. For example, over the period 1940–2009, 171 collocates of the node *war* were identified. In 1940, 19 of these collocates co-occurred with *war* (and 152 didn't), while in 1941, 23 of these collocates co-occurred with *war*, 15 of which were the same as in 1940 and 8 of which were newly introduced. This means that the collocate profiles in 1940 and 1941 overlapped in 92.98% of cases (out of 171) with AC₁ = 0.91.

3. Finally, the peaks and troughs technique is applied. Instead of frequencies, we use the inter-rater agreement index AC₁ and look for points with the lowest AC₁ scores as the points of meaning divergence. Figure 7.15 shows the resulting analysis.

In Figure 7.15, we can identify a dozen significant troughs (points of meaning divergence) by looking at data points outside the 99% confidence interval (that is, outside the shaded area). These points of divergence are 1945, 1946, 1975, 1979, 1980, 1986, 1988, 1992, 1998, 2004 and 2005. For example, in 2003, 16 collocates of *war* were identified, among which *bush*, *gulf*, *iraq*, *weapons* and *terror* are related to the war in Iraq; out of the 16 collocates 14 (including the ones listed above) are carried through to 2004 because the war in Iraq is still ongoing.

However, 12 new collocates are introduced largely connected with TV programmes with a war theme; these are *hitler*, *secrets* [of world war II], *documentary*, *starring*, *UKTV* etc. In a similar way, we can investigate other points of meaning divergence. The overall aim of UFA is to reduce the scope of analysis to points where a usage shift occurs; this can then be further investigated qualitatively using concordances.

Reporting Statistics: Peaks and Troughs and UFA

1. What to Report

For peaks and troughs, which has two obligatory and two optional steps, we need to report all the steps which were followed. For UFA, we need to report the details of the procedure for the identification of collocates (using collocate parameter notation described in Section 3.2) as well as the choice of the inter-rater agreement statistic.

2. How to Report: An Example

- The peaks and troughs technique was used (Gabrielatos & Marchi 2012); the difference between the relative frequencies of the word *war* between consecutive points in time during the period of 1940–2009 was measured. The non-linear regression model (GAM) helped to identify four peaks (1945, 1950, 1991 and 2003) and three troughs (1946, 1992 and 2004).
- The development of meaning associations of the word *war* was investigated using the UFA procedure. In the period of 1940–2009, 171 collocates of *war* were identified (3a–MI(3), L5–R5, C10_{relative}–NC10_{relative}). The collocational profiles of consecutive years were compared using the AC₁ agreement statistic.

7.6 Application: Colours in the Seventeenth Century

I'm sitting at my desk, the weather outside is miserable – grey and rainy. Through the window, I can see the mediaeval Lancaster castle on a hill, the Priory and the valley of the river Lune, all in a foggy mist; buildings, trees and people have the same indistinct colour. Only a flag with the typical Lancaster shade of red is flying from the castle like a drop of paint which an artist has left unintentionally in the picture. My thoughts take me back in history and I keep wondering about how colour was perceived in the past. Was it connected with the same associations, sensitivities and cultural frames? On a rainy Lancaster afternoon, I start searching the EEBO corpus. One billion words of early writing provide a unique insight into the use of colour words in the seventeenth century. I scribble notes, produce graphs and p-values. Here's a special kind of a diary of my journey into the past with the 'corpus time machine':

My explorations, 12 November 2016

Question 1: Which colours were the most popular in the seventeenth century?

The line graph in Figure 7.16 provides the answer. It was consistently *red*, *green* and *yellow*.

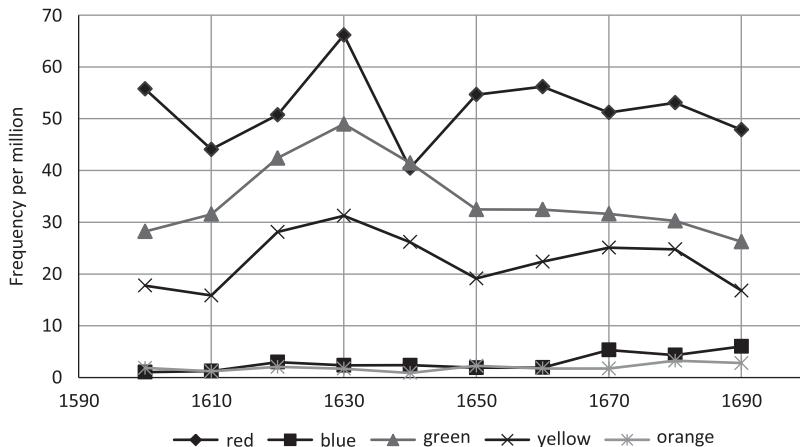


Figure 7.16 *Frequency of colour terms in the seventeenth century*

Question 2: What is the story behind colour terms in the seventeenth century?

A summary picture can be obtained from a candlestick plot (Figure 7.17). The frequency of the three colour terms (*red*, *green* and *yellow*) declined slightly, while the frequency of *blue* and *orange* increased. The long wicks in *red*, *green* and *yellow* show large fluctuations during the period.

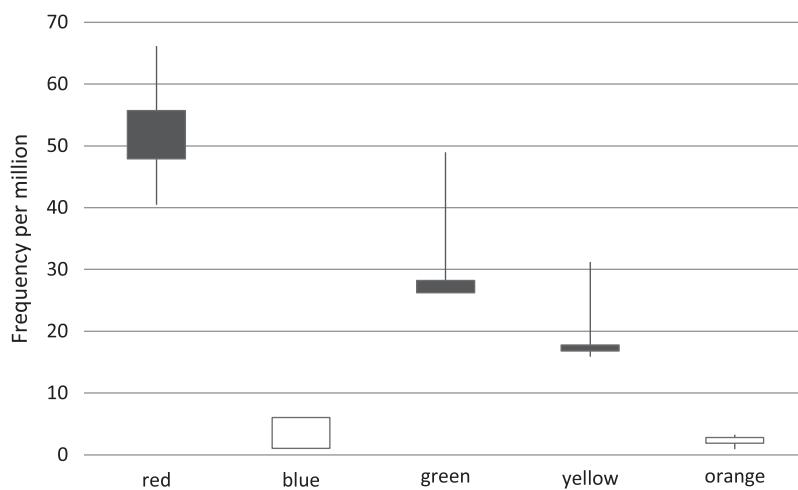


Figure 7.17 Candlestick plot: colours in the seventeenth century

Question 3: The most dramatic drop in the use of the colour term *red* was in 1640s. Is this frequency change statistically significant?

The bootstrap test showed that indeed this change is significant ($p < .001$).

Question 4: Does red, the most popular colour, have the same associations throughout the century?

The stable associations (consistent collocates) include nouns such as *coral, dragon, flowers, iron, rose, roses, sea* and *wine* as well as adjectives (mostly other colour terms): *black, green, hot, scarlet, white* and *yellow*.

UFA (see Figure 7.18) shows that the major shifts in the use of words occurred in decades (ten years were taken as each data point) starting 1606, 1607, 1620, 1621, 1624, 1625, 1631, 1633, 1639, 1640, 1672, 1682, 1683, 1684. For example, the trough 1639/40 can perhaps be traced to the social and political changes in England around this point in time. The literature in this period accentuated Biblical associations of the colour red: *cup* [of fornications], *pottage, garments, pharaoh* and his *host, tail* [of the red dragon = devil]

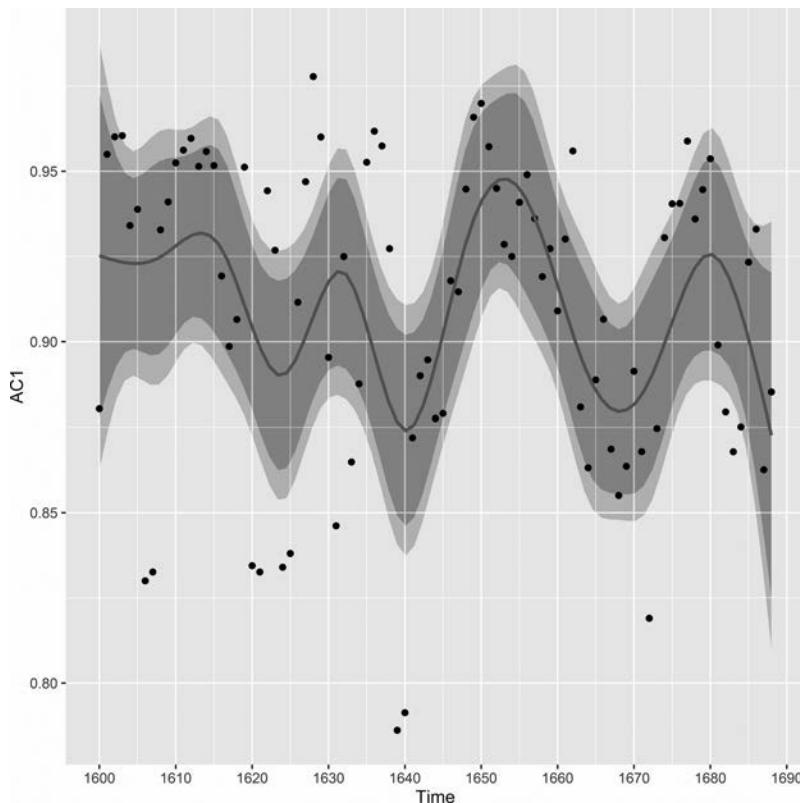


Figure 7.18 Results of UFA for red 1600–99 (3a–MI(3), L5–R5, C10relative–NC10relative; AC1)

Question 5: Looking at the term *red* throughout the seventeenth century, which periods can be identified based on the frequency of use?

VNC can be used to group individual years according to the frequency with which *red* was used. The resulting dendrogram is shown in Figure 7.19, which shows a main split between the 1600s–20s and the 1630s–90s. Otherwise, no other more compact grouping is supported by the data.

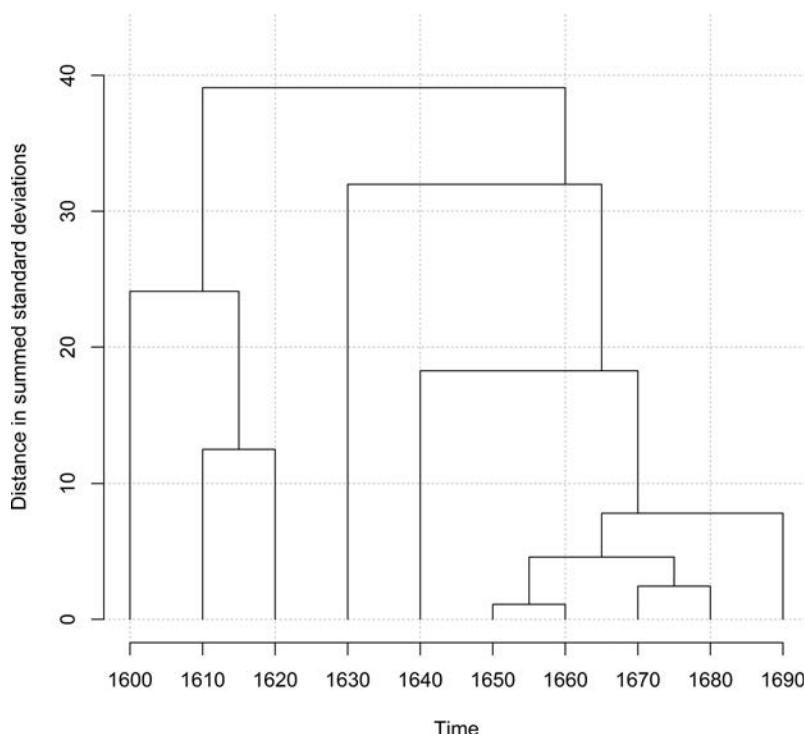


Figure 7.19 VNC: red in the seventeenth century

7.7 Exercises

1. Interpret the three graphs in Figures 7.20–7.22.

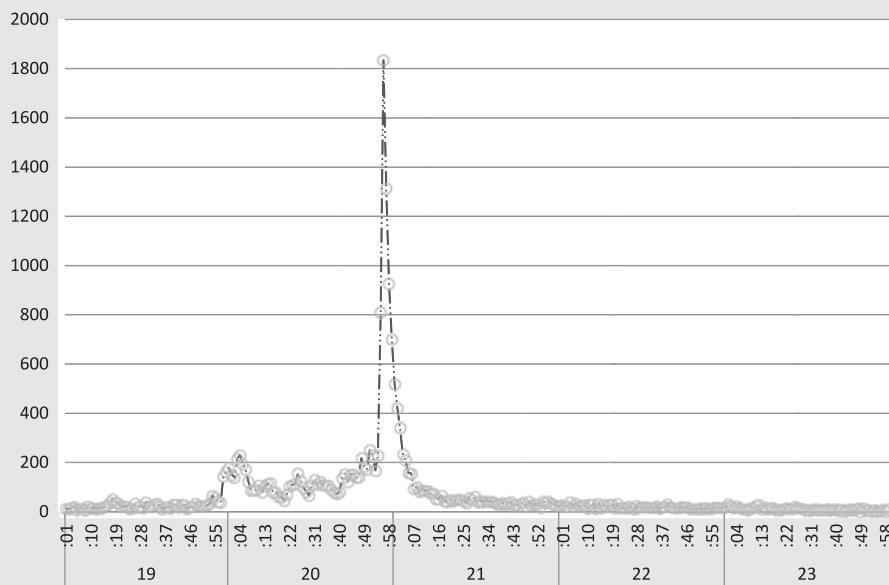


Figure 7.20 *Number of tweets related to an episode of the UK X-Factor (16/11/2014, 7–11pm)*

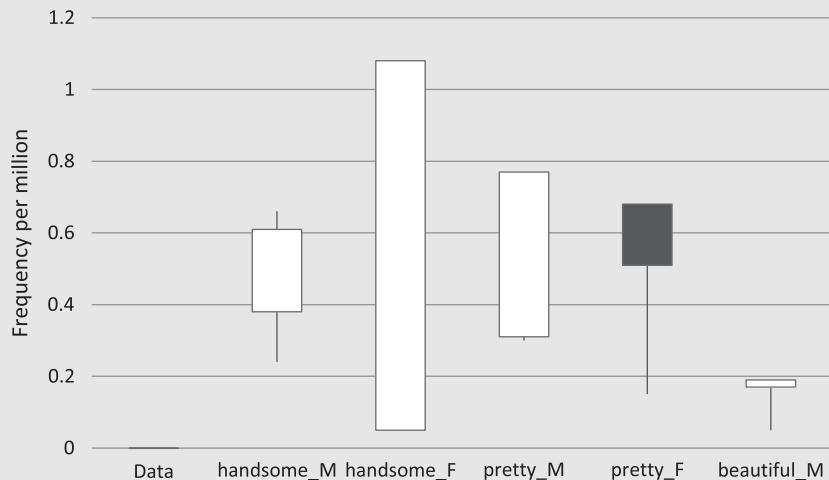


Figure 7.21 *Development of frequencies of handsome, pretty and beautiful followed by a male (M) or female (F) person in the seventeenth century*

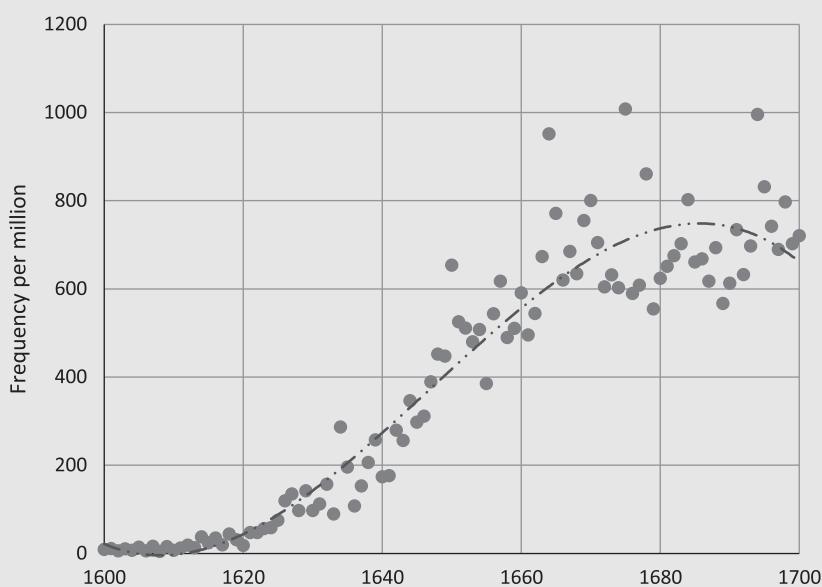


Figure 7.22 Development of frequencies of the possessive pronoun *its* in the seventeenth century

2. Fill in the blanks in the descriptions below.

Over the course of the 20th century, the frequencies of the modal *shall* (↗), *should* (↗), *may* (↗), *might* (↗), *must* (↗), and *will* (↗) , while the frequencies of *can* (↗) and *could* (↗)

In the 17th century, the adjective *handsome* used with a female person (↗) , while *pretty* (↗) in this context; *beautiful* (↗) used with a female person

3. Look critically at the trends in Figure 7.23. Which of these represents the largest change?

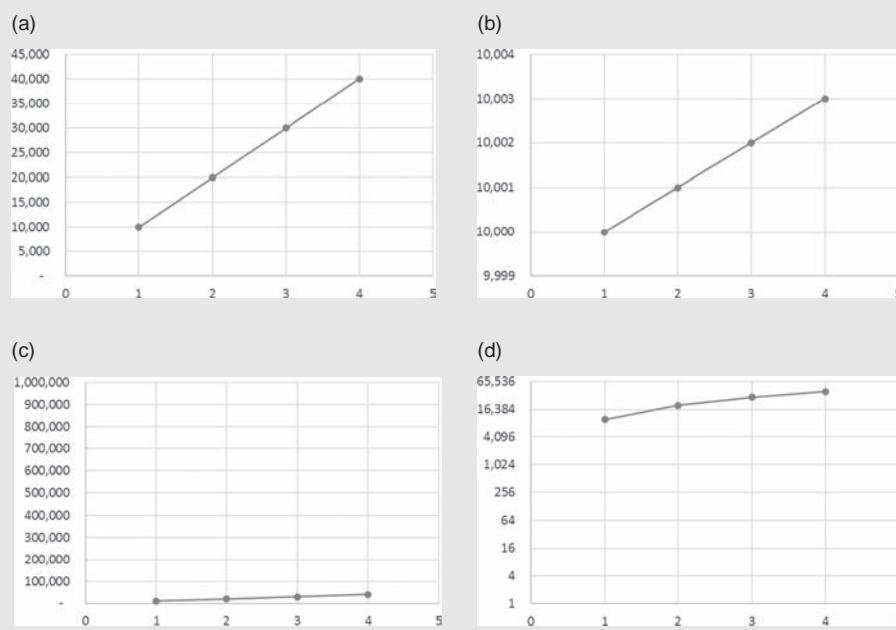


Figure 7.23 *Four frequency change scenarios*

4. Interpret the peaks and troughs graphs in Figures 7.24 and 7.25 showing the development of *handsome* and *pretty* in the seventeenth century.

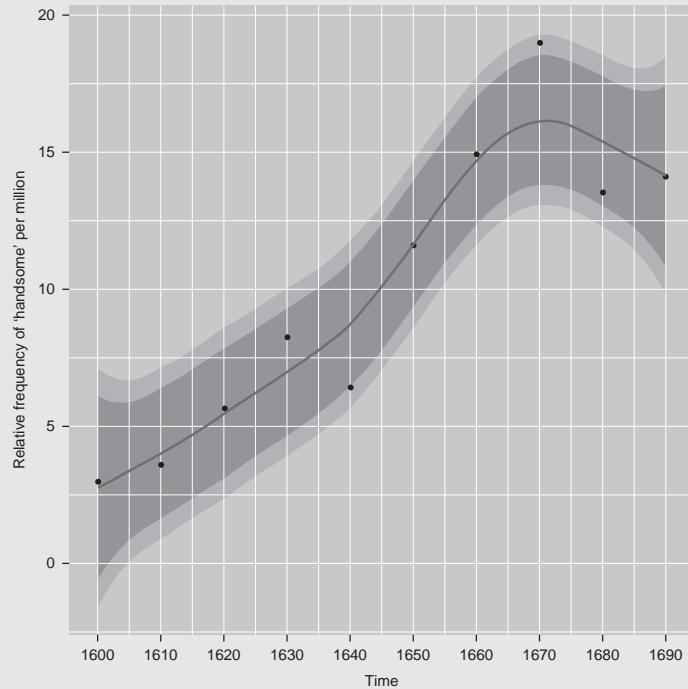


Figure 7.24 *Handsome in the seventeenth century*



Figure 7.25 *Pretty in the seventeenth century*

THINGS TO REMEMBER

- Historical analyses, because they use available and imperfect data, require critical consideration of (i) diachronic representativeness of corpora, (ii) alternative interpretations of linguistic development and (iii) fluctuation of the meaning of linguistic forms.
- Visualization options include line graphs, boxplots and error bars, sparklines and candlestick plots.
- The bootstrapping test is used to compare two corpora (representing different points in time); it makes use of a technique of multiple resampling of corpus data.
- Peaks and troughs is a technique which fits a non-linear regression to historical data, producing a graph which highlights significant outliers in the process of historical development of language and discourse.
- UFA (Usage Fluctuation Analysis) is a complex procedure combining automatic collocation comparison in a given historical period and the peaks and troughs technique.

Advanced Reading

- Baker, P. (2011). Times may change, but we will always have money: diachronic variation in recent British English. *Journal of English Linguistics*, 39(1), 65–88.
- Brezina, V., McEnery, T. & Baker, H. (in prep.). *Usage fluctuation analysis: a new way of analysing shifts in historical discourse*.
- Hilpert, M. & Gries, S. Th. (2009). Assessing frequency changes in multistage diachronic corpora: applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing*, 24(4), 385–401.
- McEnery, T. & Baker, H. (2017). *Corpus linguistics and 17th-century prostitution*. London: Bloomsbury.
- Säily, T. (2014). Sociolinguistic variation in English derivational productivity: studies and methods in diachronic corpus linguistics. *Mémoires de la Société Néophilologique de Helsinki XCIV*, Helsinki. Available at: <https://helda.helsinki.fi/handle/10138/136128>.

Companion Website: Lancaster Stats Tools Online

1. All analyses shown in this chapter can be reproduced using the statistical tools available from the companion website. The tools available for this chapter include:
 - Bootstrapping test
 - Neighbour clusters
 - Peaks and troughs
 - UFA
2. The website also offers additional materials for students and teachers.

8 Bringing Everything Together

Ten Principles of Statistical Thinking, Meta-analysis and Effect Sizes

8.1 What Is This Chapter About?

This is the final chapter of the book; it is about bringing things together on different levels. First, it brings together the statistical knowledge discussed in this book and highlights ten key principles of statistical thinking applied to corpora. Next, the chapter introduces a statistical technique called meta-analysis. Meta-analysis is a way of bringing together results of multiple studies and combining them systematically. In this way, meta-analysis contributes to a better understanding of research results in our field. Unlike a standard narrative-form literature review, which typically considers individual studies in isolation, meta-analysis can combine results from multiple studies into a single mathematical synthesis. Although formal meta-analysis is now fairly common in a number of disciplines such as psychology, second language acquisition, medical science etc., its application in corpus linguistics has been problematic due to the general lack of reporting of effect size measures. This chapter argues in favour of standardized reporting of effect sizes in corpus research and shows how meta-analysis can be carried out. Finally, the chapter reviews common effect size measures and provides a guide for their interpretation.

In this chapter, we'll be exploring the answers to three questions:

- What are the key principles of statistical thinking in corpus linguistics? (Section 8.2)
- How can we synthesize findings from multiple studies? (Section 8.3)
- How should effect sizes be interpreted? (Section 8.4)

8.2 Ten Principles of Statistical Thinking

Think about ...

Compare Tables 8.1 and 8.2. How many differences in the values can you find? Table 8.1 shows the values from the original dataset stored in a spreadsheet; Table 8.2, which appeared in the final research report, was created by manually copying and rounding the numbers from Table 8.1.

Table 8.1 *The use of the past and the present tense in different registers (original dataset)*

Register	PAST_TENSE	PRESENT_TENSE
News – reportage	4.490227273	4.424090909
News – editorial	2.452692308	5.735
Biography	4.97025974	3.502987013
Academic writing	2.249625	3.776375
Fiction – general	6.704482759	4.190689655
Fiction – mystery	8.12	3.141666667
Romance	7.690689655	3.076896552
Humour	4.601111111	4.916666667

Table 8.2 *The use of the past and the present tense in different registers (research report)*

Register	News – reportage	News – editorial	Academic Biography	Fiction – writing	Fiction – general	Fiction – mystery	Romance	Humour
PAST TENSE	4.4	2.5	4.9	2.2	6.7	8.1	7.7	4.6
PRESENT TENSE	4.4	5.7	3.5	3.7	4.2	3.1	7.1	4.9

Statistics is a discipline which, among other things, helps us express quantitative information with precision and rigour; this is required to produce research findings, which can form the basis of well-grounded scientific knowledge and can lead to advancement in the field. To summarize the most important aspects of statistical thinking applied to corpus linguistics, the following ten principles, conveniently labelled with words starting with A–J, are put forward.

1. ATTENTION TO DETAIL: pay attention when looking at corpus tool outputs, entering data into spreadsheets, copying data to research reports and during other types of low-level data processing.

The devil, as the saying goes, is always in the detail. While the focus in statistical textbooks and in the field in general is on statistical techniques, interpretations of p-values etc., low-level operations such as getting data from a corpus tool into a spreadsheet and then into a statistical package often remain in the background. Nevertheless, these operations are equally important for obtaining reliable results: statistics, as we already know, is a discipline concerned not only with data analysis but also with systematic and reliable collection of data (see Section 1.2). If carefully compared, it is apparent that Tables 8.1 and 8.2 from the ‘Think

about' task differ in four entries, demonstrating typos and typical errors from incorrect rounding. Such errors can easily make their way to final research reports. In this case, the errors occur in 25% of the values (cells in the table)! We have to remember that if we make a measurement error or make a mistake while copying data from one tool into another, the final analysis will suffer from this elementary inaccuracy, no matter how sophisticated the statistical technique we use is. As Leek and Peng (2015: 612) point out, 'p-values are just the tip of the iceberg'; attention should therefore be paid to the whole 'data pipeline', which involves data collection, data cleaning, exploratory data analysis, statistical modelling and statistical inference. Corpus tools can help this process by integrating statistical analyses to include as much of the data pipeline as possible. By doing this, the chance of an inadvertent data-copying error can be considerably reduced. For example, a new-generation corpus package #LancsBox (Brezina et al. in prep.) follows this philosophy by offering powerful 'Search & Analyse' functionalities with customizable statistical analyses. Nevertheless, the simple advice holds: pay attention to detail, identify points in the analysis where mistakes can occur and always double and triple check the accuracy of the values used in the analysis.

2. BASICS FIRST: start by familiarizing yourself with the corpus and performing descriptive statistics.

One of the most important tasks for an analyst is to achieve a very good general understanding of the data. This often involves careful reading through the corpus manual to familiarize oneself with the corpus composition, inspecting the concordance lines to see the actual examples of language use behind the numbers we have obtained and producing overviews and simple graphs that reveal the main tendencies in the dataset. All this is subsumed under descriptive statistics. Without a reliable description of the data, generalizing from the sample (corpus) to the population (language as such) using sophisticated statistical tests and p-values lacks grounding. This makes meaningful interpretation of the results difficult, if not impossible. For example, the dataset from Table 8.1 comes from BE06, a one-million-word corpus of written English compiled according to the Brown family sampling frame (Baker 2009). Figure 8.1, taken from an article describing the corpus and its design, provides details about the representativeness of the corpus and the genre/register distribution. Note that the corpus includes traditional written genres/registers but does not sample new online genres/registers such as blog posts, tweets, online discussion forums etc. This type of information is essential for any generalizations based on the corpus. Focusing on the use of the past tense in different genres/registers from Table 8.1, we can see that academic writing has the smallest frequency of the past tense, while mystery fiction contains the highest frequency of past tenses. This pattern becomes immediately obvious from the stacked bar chart in Figure 8.2. The examples of use of the past tense in these two genres/registers (see Table 8.3) helps us contextualize the finding and notice the

Text categories		Number of texts in each category
A	Press: reportage	44
B	Press: editorial	27
C	Press: reviews	17
D	Religion	17
E	Skills, trades and hobbies	36
F	Popular lore	48
G	Belles lettres, biography, essays	75
H	Miscellaneous (government documents, foundation reports, industry reports, college catalogue, industry house organ)	30
J	Learned and scientific writings	80
K	General fiction	29
L	Mystery and detective fiction	24
M	Science fiction	6
N	Adventure and western fiction	29
P	Romance and love story	29
R	Humour	9

Figure 8.1 *Overview of genres in BE06 (Baker 2009)*

pattern of frequent use of the passive voice in academic writing. In sum, the basic prerequisite of successful corpus analysis is having thorough knowledge of the dataset on which the analysis is based.

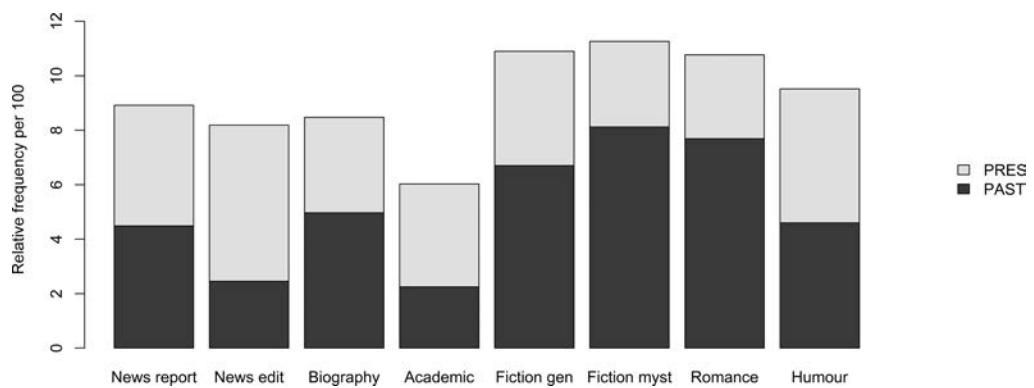


Figure 8.2 *Past tense in different written genres of BE06*

Table 8.3 Examples of use of the past tense in academic writing and mystery fiction

Academic writing	
to measure anxiety and depression, and we control during feeding. Next, a Pearson's correlation analyses how practices of attention and description = 0.806). Secondary outcomes Statistically significant differences initial purities in excess of 99% and the nurse should further problems develop. This and more younger responders than older responders the formulation of a working group that to 100 (best health). Physical function restriction a real need for significantly more State	used self-reported weight and height to calculate body was used to investigate the relationship between weight were re-created for various audiences in the second were also found in dominant hand key grip were obtained from Sigma–Aldrich, Fisher Scientific or Shell was the model that was practiced in this reported multiple pains in the knee-pain group (both included scientists generating transcriptomic data for a variety was worse in the knee-pain group than in supplied social housing and first time buyer homes.
Fiction – mystery	
<p>the guys give a start and I stared straight ahead and for several minutes the kid wanting to bury me,” Mike</p> <p>to make his case strongly enough. He head around the door and saw him as he went. A loud klaxon immediately his own heartbeat, and wondering if he Sharples had been a model prisoner, Tony from just behind the door, and Slider</p> <p>who were deranged but smart. Recently, Allen</p>	
	turned and June Carter came in, that bitch, her eyes, willing the image of their after a long silence. “There’s nothing creepy closed said retrieved his glasses and marked a couple of in an armchair, a book on his slumped blasted out. Christ, if you were crazy already, about to die. His mouth was so was thought bitterly. It was easy to behave when he was being examined through the peephole. guessed had devised a method of avoiding taking his

3. CLARITY: the use of statistical procedures should be clear, transparent and well motivated.

Clarity in corpus statistics has a number of aspects. First, when using statistical techniques, we need to make clear what the reason is for choosing a particular statistical procedure. The choice of statistical procedures is usually motivated by a particular type of corpus data and a particular research design (see Section 1.4). For

example, the differences between individual genres/registers in Table 8.1 can be investigated using the one-way ANOVA test (see Section 6.3) where we take into account one linguistic variable (e.g. past tense) at a time. Alternatively, the data can be explored using multi-dimensional analysis (see Section 5.4), where a large number of linguistic variables are considered at the same time. This book discusses different areas of corpus-based analysis, which are matched with the most suitable statistical techniques. However, in many instances other approaches are equally valid – in each individual case, the motivation for the choice of one technique over another needs to be clearly spelled out. Second, statistics need to be clearly reported. This involves clear presentation of data in graphs and tables but also standard reporting of statistical tests and procedures, so that the reader can easily interpret the results and reproduce the study. Throughout this book, ‘Reporting statistics’ boxes gave practical examples of standardized statistical reporting.

4. DATA: pay special attention to the quality of the corpus data and search procedures.

The success of any research depends on the quality of the data and the effectiveness of the analytical procedure; yet, especially in corpus research, the quality of corpus data is rarely scrutinized. ‘I found this in a corpus and therefore it must be true’ is a sentiment which ascribes corpora a certain kind of magical power that they don’t have. Even well-established corpora such as the BNC include errors, inconsistencies and possible bias. All corpora thus need to be approached critically. For example, Gablasova et al. (2017a) show a large amount of variation in corpora designed to represent the same type of language and argue for the need for empirical validation of corpus data. In addition, the same corpus can provide considerably different answers when searched using different tools. Brezina & Timperley (2017) point out that the answers to the question of ‘How large is the *British National Corpus*?’ differ by up to 17% depending on the tool which we use. In computation and data processing, a well-known acronym, GIGO (‘Garbage in, garbage out’), is used to remind us that the data needs to be approached critically, otherwise the reliability of the analyses cannot be guaranteed.

5. EFFECT SIZE: calculate, report and interpret the size of the effect observed in the data.

It is always important to think about the practical effect of the observations about language we make using corpora. To help us express this aspect of the findings, effect size measures should be used. In broad terms, effect size can be defined as the ‘amount of anything that might be of interest’ (Cumming 2012: 38) to researchers. In corpus research, the focus is on measuring and quantifying linguistic effects such as frequencies of words or phrases in corpora or differences between these frequencies in various subcorpora. Effect size measures include a wide range of descriptive measures (Cumming 2012: 39; Kirk 2005: 2; see Section 8.4 for more discussion).

Table 8.4 *Examples of texts: academic writing and mystery fiction*

Academic writing (BE06-J61)	Mystery (BE06-L10)
Immigrants contribute significantly to the overall economic performance of their host economies. It is therefore not surprising that a large literature is concerned with the earnings mobility of the foreign born population, both in isolation, as well as in comparison with those who are native born. But immigrants have not only an immediate effect on wealth accumulation and earnings and skill composition. They transmit their earnings status, as well as socio-economic and cultural characteristics to the next generation. The economic adjustment process within the immigrant's own generation has long been recognised as an important step in understanding the economic effects of immigration. For understanding the long term consequences of immigration, assessment of intergenerational mobility in immigrant communities is perhaps equally important.	"I <u>couldn't</u> believe it, I never seen anyone famous, not, like, in real life. I <u>gave</u> 'em my best service, and in those days, I <u>was</u> hot, <u>had</u> some moves." Foley nearly <u>said</u> , "You still do." But bit down and <u>wondered</u> where the hell his partner <u>had</u> got to. Probably gone for a bourbon, Shiner back. He'd return, smelling of mints, like that <u>was</u> a disguise. He <u>asked</u> , "You talk to him, to Mr. Cash?" "Not at first. I <u>was</u> getting them vittles, drinks, making sure they <u>were</u> comfortable and after, I dunno, an hour, Johnny said, "Take a pew little lady, get a load off." She <u>rubbed</u> here eyes, then. "He <u>had</u> these amazing boots, all scuffed but, like, real expensive, snakeskin or something, and he <u>used</u> his boot to hook a chair, pull it up beside him." She <u>touched</u> her face, self conscious, <u>said</u> , "I <u>didn't</u> have the scar then, still <u>had</u> some dreams. Jesus".

Let's quickly review a brief example. When comparing two groups such as two subcorpora, the effect size measure Cohen's d is often used (see Section 6.3). Cohen's d for the difference between the use of the past tense in academic writing and mystery fiction (see Table 8.1 for the genres in BE06) is -3.26 , 95% CI $[-3.9, -2.61]$, which is a very large effect; the negative number signifies that the frequency in the second subcorpus (mystery) is larger than the frequency in the first subcorpus (academic writing). This effect is clearly observable when looking at the texts belonging to these two genres/register (see Table 8.4, where past tenses are underlined). The interpretation of effect sizes is discussed in detail in Section 8.4.

6. FOLLOWING THE BEST PRACTICES IN THE FIELD: critically review the statistical practice in the field and follow good examples.

In addition to being aware of more general statistical principles (see Chapter 1), it is important to know how to apply these principles in the analysis of language corpora. For this, following the best practice in the field is essential. This books

contributes to the general review of best practice and offers specific suggestions for the use of statistics in corpus linguistic research. The general advice is this: if you are a beginner in the field, follow the established practice of statistical analysis. Find an article published in a reputable journal which deals with a similar research question to yours and try to apply the same procedures and statistical measures to your data. Then consider if this method was useful for answering your research question. However, you should not stop there. As you become more proficient in statistical techniques, you should critically evaluate the current practice in the field and look for innovation, perhaps inspired by the use of statistics in other disciplines. This requires you to read research outside your area. Here, it is important to realize that not all uses of statistics, even in high-ranking journals, follow best practice. For example, it has been an unchallenged practice in corpus linguistics to use the chi-squared test or the log-likelihood test (likelihood ratio test) for almost any comparison of two or more corpora. Recently, however, this practice has been criticized and alternative measures for more meaningful comparisons have been suggested (e.g. Brezina & Meyerhoff 2014; Lijffijt et al. 2016).

7. GRAPHICS: visualize data to identify patterns.

A picture, as is often said, is worth a thousand words. Effective visualization can help us discover important patterns and relationships in data. The basic principle of good visualization is simple: focus on displaying the data rather than simply adding ‘visual frills’. Make the display informationally rich to allow relationships between variables to be observed (Tufte 2006). For example, Figure 8.2 can be re-rendered as two boxplots, which are informationally richer: one for the past tense, another for the present tense (Figure 8.3). These boxplots show both the central tendencies (means and medians) in each subcorpus as well as individual texts (displayed as circles). No data point is missing, which offers us the full picture and allows us to see the distribution of individual values including outliers (exceptions to the general tendency).

However, it is important to bear in mind that visualization is not appropriate in all cases. Sometimes, tables or plain statements describing the data are more effective. As Tufte puts it (quoting Ad Reinhardt), ‘[a]s for a picture, if it isn’t worth a thousand words, the hell with it’ (Tufte 1997: 119).

8. HIGHLIGHTING BOTH SIMILARITIES AND DIFFERENCES: provide a balanced account of language use.

It is no surprise that when we look for differences between corpora or between various uses of a word or phrase, we usually find some. Most of the methodological thinking (not only in corpus linguistics but other disciplines as well) has been biased towards looking for differences and

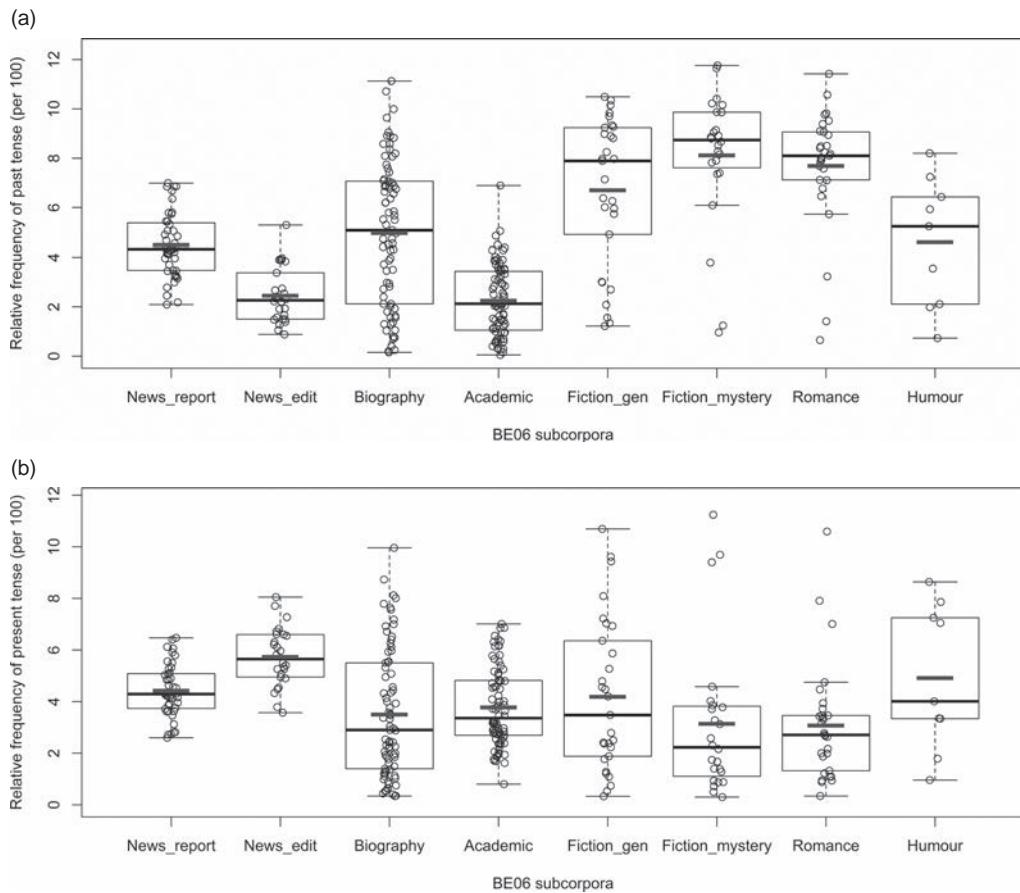


Figure 8.3 *Past tense (a) and present tense (b) in different written genres of BE06: boxplot rendition*

ignoring similarities. Similarities and null effects (effects that are small and/or not statistically significant) are thus often underreported in the literature (see the discussion of publication bias in Section 8.3). In corpus linguistics, the focus on differences, for example, leads us to reporting keywords and disregarding lockwords, words which are stable across corpora (Baker 2011; see Section 3.4) or focusing on sociolinguistic differences (e.g. differences between genders) and forgetting about the profound similarities in language use (Baker 2014; see also Chapter 6). So far, in the example dataset from Table 8.1 we have highlighted mainly differences, for example, between the academic writing and mystery fiction. When looking at Figure 8.3, we can, however, also see a robust pattern of similarities, for example, between news editorials and academic writing when looking at

the use of the past tense or among the majority of genres/registers when looking at the use of the present tense. The methodological message is simple: always strive to provide a balanced account of the data which places the same emphasis on the similarities as on the differences.

9. INTERPLAY BETWEEN STATISTICS AND LINGUISTICS: provide robust statistical analysis that is grounded in linguistic and social theory.

Let's consider a methodological question. Which of these is worse in corpus research: linguistics without statistics or statistics without linguistics? Arguably, each of these hypothetical options fails in one fundamental respect. Linguistics without statistics lacks effective tools for analysing large quantities of language data, while statistics without linguistics can easily turn into a mindless exercise in number crunching without a connection to linguistic and social reality. We have to remember that not all linguistic research is quantitative – qualitative research can bring important insights into the use of language; however, the choice of texts and examples for in-depth analysis raises the question of how typical these are and why we selected them (see Baker et al. 2008). Statistical methods can provide a principled way of selecting examples for more in-depth analysis and help us avoid cherry picking, selecting only examples that fit our pre-conceived ideas. We need to remember that only a mutually informed relationship between statistics and linguistics can provide useful results: statistics offers techniques for reducing the ‘problem space’ and focusing our attention on typical/unusual patterns of language use, which need to be interpreted linguistically (cf. McEnery & Baker 2017).

10. JARGON: use statistical terminology and notation where it helps express things clearly, but try to avoid unnecessary jargon.

For many people, statistics can be opaque. In statistics, we often try to capture complex relationships in data using mathematical expressions, which themselves might put many people off. Throughout this book, where possible, mathematical symbolism has been replaced by glosses in plain English; for instance, $\sum_{i=1}^{10} i$ would be expressed as ‘the sum of all integers from 1 to 10’ (when you work out the maths both ‘renditions’ will give you the number 55). However, when reporting on the results of statistical analyses it is important to provide these in a standardized form; again, this book shows how this can be done in the ‘Reporting statistics’ boxes.

8.3 Meta-analysis: Statistical Synthesis of Research Results

Think about ...

Before reading this section, think about the following situation.

You are in a large city such as London and are looking for a theatre, say the replica of Shakespeare's Globe. You don't have a map so you have to rely on the information from passers-by. Table 8.5 shows five answers to the question 'Where is the Globe theatre?', some of which differ significantly. Where would you go?

Table 8.5 *Overview of answers*

	Direction	Answer	Person profile
Person 1	straight on and right ↑	'I don't know, actually maybe you need to just go straight on and then turn right.'	a person looking like a tourist
Person 2	straight on and left ↓	'That's very easy. Follow this road and then turn left.'	a local shop keeper
Person 3	right and back ↓	'Turn right and then retrace your steps all the way. I'm absolutely sure.'	a man in a Mad Hatter costume
Person 4	straight on and right ↑	'Sorry I'm in a hurry: straight on and right.'	a person exiting an office building
Person 5	straight on and left ↓	'The Globe is not far from here. I've just walked past it. Walk down this road to the bridge, go down the stairs and left alongside the river.'	a person walking a dog

The example situation in the 'Think about' task can be understood as a metaphor for the scientific pursuit. To find out the answer to our research question we go out to collect and evaluate data. In science, this is done repeatedly to ensure the reliability of the answer. The process of repeating a study with the same research question but a different dataset is called **replication**. In our example, replication is demonstrated with asking different passers-by the same question. And as in our example, in science we often get (slightly) different answers from different studies. What is, however, important for the development of our field as a whole is to be able to see the larger picture; we therefore need to put individual findings together and make sense of them globally. The statistical technique that does this is called meta-analysis. **Meta-analysis** is a quantitative procedure of statistical

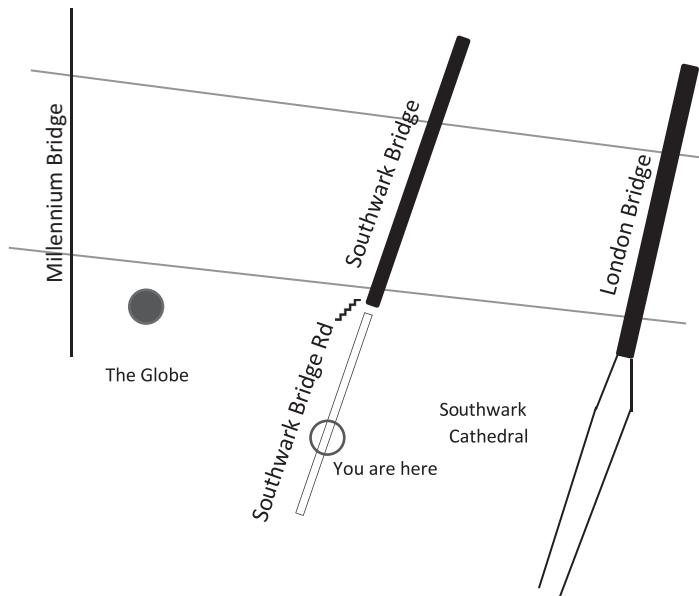


Figure 8.4 *Finding the Globe*

synthesis of research results, which is based on combining the effects reported in multiple individual studies and calculating the summary effect (Borenstein 2009). The whole point of meta-analysis is to bring together the evidence that is scattered across individual studies to increase the confidence about the resulting summary conclusion; this confidence can be quantified using a confidence interval (see below). Returning to our example, to find the Globe theatre, we need to evaluate the individual answers in terms of how certain and reliable these are (note that certainty and reliability are two separate concepts) and then make a decision about the route which should be taken. The route which will take us to the Globe in the situation described in the ‘Think about’ task is ‘straight on and left’ (see Figure 8.4).

So how does meta-analysis work in practice? There are three basic steps that need to be followed: (1) identification of relevant studies, (2) extraction of relevant pieces of information from the studies (coding) and (3) statistical synthesis.

Step 1. Identification of relevant studies

This step defines the area of research which we wish to investigate in the meta-study. The area is given by a research question that is answered repeatedly in multiple studies using different corpora. Of course, the research question can be broader or more specific thus defining the granularity of a meta-study. For

example, a very broad question would be: Is there an effect of gender on the use of language? Guided by this research question, we would be able to identify hundreds of relevant studies but the answer would probably be very broad and non-specific: some linguistic features show gender-based patterns, while others don't; this can be quantified as a very broad summary effect. If, on the other hand, we define the area of research using a specific research question, e.g. Is there an effect of gender on the use of pronouns?, we can tap into something more specific and arguably more interesting; however, the number of relevant studies will be considerably smaller. In practical terms, it is important to explicitly specify inclusion criteria for the studies such as what linguistic and what explanatory variables we are looking for, requirements for research design and relevant time frame for the studies, e.g. studies published since 1990 (Wilson 2009: 161–3).

Once we define the area of research, we need to search all relevant journals and available databases (e.g. Google scholar, Linguistics and Language Behavior Abstracts, ProQuest Dissertations & Theses, EThOS etc.) for studies, both published and unpublished, that address the research question of interest and meet the inclusion criteria. The aim is to collect all available evidence related to the research question. The reason to also consider quality unpublished studies (e.g. PhD theses) is the fact that we want to reduce the effect of a so-called publication bias. **Publication bias** is a well-documented phenomenon (e.g. Kepes et al. 2014) of overreporting stronger and hence statistically significant effects and underreporting so-called **null results** (that is, statistically non-significant results) especially in published work: authors but also journal reviewers and editors often mistakenly view weak effects in methodologically solid studies as not interesting or not important; however, reporting these results is crucial for being able to see the whole picture. What often gets reported are statistically significant results, which represent only the tip of the iceberg with an unaccounted mass ('grey literature') of smaller effects hidden under the water (Rothstein & Hopewell 2009).¹

A problem related to publication bias is the quality of the studies selected for meta-analysis. To avoid publication bias, when selecting the studies, we want to cast the net as widely as possible, yet we want to make sure that what ends up in the net are methodologically sound studies (fish rather than flotsam, to follow up on the fishing metaphor). In meta-analysis as in primary research, the GIGO (garbage in garbage out, see Section 8.2) principle should be a reminder that we need to scrutinize carefully the available evidence (in this case the quality of primary studies) before we use it in our analyses. Valentine (2009) discusses a number of aspects of the

¹ It needs to be noted that the bias towards large and statistically significant effects is present in both the published and unpublished literature to a certain degree. Unpublished reports, however, have potentially more space to report/discuss minor effects.

quality of research studies including internal validity (control for relevant variables, precise measurement), external validity (generalizability of the study), construct validity (measuring what we want to measure) and statistical conclusion validity (statistical rigour). In corpus linguistics, a field where the methodological procedures are still in the process of development and standardization, the quality of the studies that appear in a meta-analysis is of particular concern. We wouldn't suspect authors of deliberately manipulating data or suppressing undesirable results (as might be the case in e.g. commissioned pharmaceutical studies or opinion polls); more often, the authors might unwittingly breach an assumption of a statistical test or use an unsuitable research design. Readers are referred to Section 8.2 for an overview of 'Ten principles of statistical thinking' applied to corpus research as guidelines for rigorous research practice.

Step 2. Extraction of relevant pieces of information from the studies (coding)

The second step involves obtaining relevant data from the studies we identified in step 1. This involves carefully reading through the method and results sections of the research reports, articles and books and noting down information about (i) the (sub)corpora used, (ii) the method and (iii) the effect sizes observed in the study (Wilson 2009). In particular, we need to know how large the compared (sub)corpora were not only in terms of the token count, but, more importantly, in terms of the cases observed (i.e. speakers/texts in the individual text/speaker research design or occurrences of the target linguistic feature in the linguistic feature research design; see Section 1.4). Also, it is important to note down the type of research design, linguistic and explanatory variables included, statistical test(s) used and any other relevant pieces of information about the methodology that may affect the result. Finally, we need to note down either standard (comparable) effect size measures or information (e.g. t -value or F -value) from which we can extrapolate the effect size measure (see Section 8.4). Note that research based on exactly the same corpus/corpora should be included only once, regardless of the number of publications this research appears in. Multiple publications by the same authors or repeated research by different authors based on the same dataset contribute only one piece of information to the meta-analysis.² For example, the following is a possible coding sheet that we can use for recording information from individual studies:

² However, if dealing with a repetition (study by different authors using the same corpus/corpora and the same research question) we might like to compare the results of the original study and the repetition and evaluate the methodologies; based on this we can select the study that is more reliable and/or that provides more complete information for the meta-analysis.

CODING SHEET: PRONOUNS AND GENDER**Header**

RQ: Is there an effect of gender on the use of pronouns?

Coder initials: VB

Study: Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45(3), 211-236.

Quality: high acceptable low

Notes:

Corpus

Name: NA – an opportunistic corpus compiled from multiple psychological studies

Tokens: 45,700,000

Cases: male= 5,971; female= 8,353

Representativeness: Language from different tasks in psychological experiments. 93% written, 7% spoken, participant's age: 2/3 from college-age participants, period of collection 1980-2002

Method

Research design: individual text/speaker linguistic feature whole corpus

Statistical test(s): Welsch's t-test

Relevant linguistic variable(s): all personal pronouns

Other linguistic variables: a range of automatically identified lexical variables

Relevant explanatory variable(s): gender

Other explanatory variables: NA

Results: Effect(s) observed

ES measure reported: YES NO

If yes, which one(s)? Cohen's *d*

Effect1: gender → all pers. pronouns size: *d*=0.36 direction: female +

Effect2: gender → 1st pers. pronouns sg. size: *d*=0.17 direction: female +

...

Attentional statistical information reported (tests, df, means, SDs, SEs):

all pers. pronouns: male: *M* = 12.69, *SD* = 4.63; female: *M* = 14.24, *SD* = 4.06;

...

NOTES

Because in many cases the information about different aspects of the study needs to be inferred or extrapolated from the report and this might involve an element of subjectivity, it is good practice to double code a portion of the data (e.g. 20%) and calculate an inter-coder agreement statistic (see Section 3.5). This is also important if multiple coders are involved.

Let us now have a look at an example of five studies reporting the effect of gender on the use of pronouns (*I, me, my, you, your* etc.). Table 8.6 shows the details of these studies (extracted from the coding sheets), focusing on relevant statistical information reported; some of the information is provided directly, some needs to be inferred (see Section 8.4). Note that two studies (study 4 and study 6) were excluded from the quantitative part of the meta-analysis. Study 4

Table 8.6 *Studies reviewed for the meta-analysis*

Study	Information provided	Information inferred
1. Newman et al. (2008)	Corpus: $n_{male} = 5,971$; $n_{female} = 8,353$ Results (pronouns): Cohen's $d = 0.36$	
2. Argamon et al. (2003) – non-fiction	Corpus: $n_{male} = 179$; $n_{female} = 179$ Results (pronouns): male: $M = 282$; $SE = 12$ female: $M = 390$; $SE = 19$	Cohen's $d = 0.508$
3. Argamon et al. (2003) – fiction	Corpus: $n_{male} = 123$; $n_{female} = 123$ Results (pronouns): male: $M = 860$; $SE = 18$ female: $M = 977$; $SE = 18$	Cohen's $d = 0.586$
4. Argamon et al. (2007)	Corpus: $n_{male} = 25,065$; $n_{female} = 21,682$ Results (pronouns): male: $M = 9.84$; female: $M = 11.97$	Cohen's d cannot be computed because variation (SD) is not reported
5. Colley & Todd (2002)	Corpus: $n_{male} = 24$; $n_{female} = 30$ Results (<i>you, your</i>): male: $M = 0.78$; female: $M = 1.33$; $F(1, 50) = 7.69$	Cohen's $d = 0.759$
6. Rayson et al. (1997)	Corpus: Corpus: $n_{male} = 536$; $n_{female} = 561$ Results (pronouns): male = 13.37%, female = 14.55%, $\chi^2 = 1016.27$	Incompatible research design

had to be excluded because of incomplete reporting of statistics. In such cases, it is good practice to contact the authors of the analysis with a request for additional information. Authors of primary studies should also be encouraged to use reporting standards (see the ‘Reporting statistics’ boxes throughout this book) which provide sufficient information for the evaluation and synthesis (meta-analysis) of the results. Study 6 was excluded because it uses an incompatible research design (whole corpus design).

Step 3. Statistical synthesis

The final step of meta-analysis puts the information from the individual studies together and produces a report of the resulting (combined) effect size and a confidence interval. In its simplest form, the statistical synthesis takes as input the effect size measure (here Cohen's d) and the number of cases (texts/speakers) in each subcorpus (see Table 8.7). The effect sizes for individual studies are often weighted to give more prominence to studies with larger corpora (because these are closer to the true population value), optionally also to those of higher quality (using a quality index). Typically, inverse variance is used for weighting, which takes into account the number of cases – the more cases (texts/speakers) the study includes, the larger the weight (importance).

Table 8.7 *Input data for a simple meta-analysis*

Study	Cohen's <i>d</i>	Cases in sub-corpus 1 (<i>n</i> ₁)	Cases in sub-corpus 2 (<i>n</i> ₂)	Inverse variance
1. Newman et al. (2008)	0.36	5,971	8,353	3,428.0
2. Argamon et al. (2003) – non-fiction	0.508	179	179	86.7
3. Argamon et al. (2003) – fiction	0.586	123	123	59.0
4. Colley & Todd (2002)	0.759	24	30	12.4

Inverse variance for Cohen's *d*, a typical measure in meta-analyses when comparing two text types or groups of speakers using the individual text/speaker design, is calculated as follows:³

$$\text{Inverse variance} = \frac{1}{\frac{n_1 + n_2}{n_1 \times n_2} + \frac{d^2}{2(n_1 + n_2)}} \quad (8.1)$$

For example, the inverse variance for the first study (Newman et al. 2008) would be calculated as:

$$\text{Inverse variance} = \frac{1}{\frac{5,971 + 8,353}{5,971 \times 8,353} + \frac{0.36^2}{2 \times (5,971 + 8,353)}} = 3,428.0 \quad (8.2)$$

The process that combines the effect sizes has two options: the fixed-effects model and the random-effects model. The fixed-effects model assumes that all the studies in the meta-analysis are exact replications of each other (use the same design, methods, variables etc.) and differ only in the corpora used. Because under this assumption the only variation to control for is the variation from one sample to another (sampling error), the fixed-effects model simply combines the weighted effects to produce the summary effect. Random effect, on the other hand, assumes multiple sources of variation when studies investigate slightly different constructs using somewhat different methodological choices, which in reality is often the case. Random effect is therefore recommended as the default option. When the random-effects model is used for the data from Table 8.7, we get the summary effect of .47 with the 95% CI [.32, .62].⁴ A visual summary of the meta-analysis is provided in Figure 8.5, which uses a so-called forest plot. The **forest plot** provides a summary of a meta-analysis by displaying the effect sizes of individual studies (filled squares) and their confidence intervals (whiskers) as well as the overall effect size (diamond); the size of the square represents the weight of an individual study in the analysis.

³ Equations for calculating (inverse) variance for other effect size measures are available in Shadish & Haddock (2009: 264ff). However, you don't have to worry about the details of the equations because the meta-analysis tool on the companion website calculates these automatically.

⁴ The fixed-effects-model value is .37, 95% CI [.34, .40].

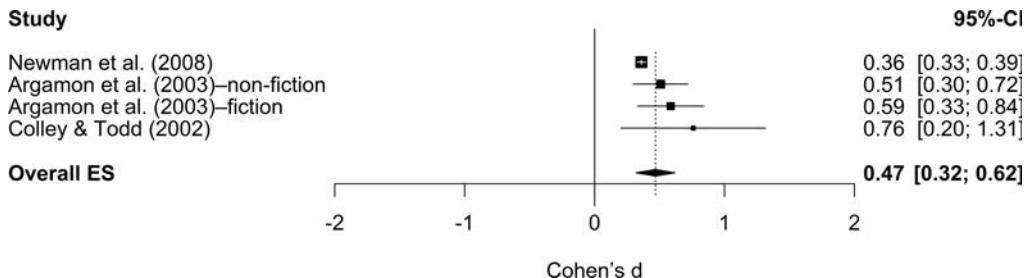


Figure 8.5 *Forest plot: meta-analysis of four studies*

In addition to the simple meta-analysis described above, more complex models can be used, which take into account so-called moderator variables. Moderator variables are additional variables identified in the coding process (step 2) as having impact on the effect sizes of the studies; these variables can be incorporated in the meta-analysis models. For more information on this type of analysis, see Schmidt & Hunter (2015: 381ff.). It is important to note that more complex models of meta-analysis are only possible if there are sufficient studies with the moderator variables reported for us to synthesize. In sum, knowledge in science is cumulative. Meta-analysis provides tools for effective statistical synthesis of research results. However, the success of the technique directly depends on (i) the quality of the input studies and (ii) inclusion of a broad range of research reports.

Reporting Statistics: Meta-analysis

1. What to Report

Meta-analysis is a complex procedure which requires detailed reporting of methodological decisions at each step of the process: (i) identification of relevant studies, (ii) coding and (iii) statistical synthesis. It is also good practice to provide (in an appendix, in additional data online etc.) the coding sheet and the spreadsheet/database with the coded studies on which the meta-analysis is based. For a detailed discussion on reporting meta-analysis, see Rosenthal (1995).

2. How to Report: An Example

- A meta-analysis was performed with studies reporting the effect of gender (explanatory variable) on the use of pronouns (linguistic variable) in English. Both published and unpublished (PhD theses) sources were searched. The studies were coded according to a coding sheet (see the Appendix), with 20% of the entries annotated by two coders (inter-rater agreement: 95%). Initially, six studies were identified, from which only four could be used in the meta-analysis (see the Appendix for coded data). The effect sizes of the individual

studies were weighted by inverse variance and the random-effect model was used. The procedure identified a summary effect $d = .47$ with the 95% CI [.32, .62]. A visual summary of the meta-analysis can be seen from the forest plot in Figure 8.5.

8.4 Effect Sizes: A Guide for Meaningful Use

Think about ...

Before reading this section, think about the following situation:

You are preparing for a half marathon and want to evaluate the effect of the training. Before the training, you run 200 metres in one minute; after the training, you run 220 yards in the same amount of time. Did the training help?

To answer the question from the ‘Think about’ task, we first need to convert the values to a common unit. If you are used to operating in metric units, you need to convert yards to metres (220 yards \approx 201.2 metres); if, on the other hand, you are used to operating in imperial units, you convert metres to yards (200 metres \approx 218.7 yards). The next question is whether we consider a 1.2-metre (1.3-yard) improvement important or not. It clearly is an improvement, but at first sight it seems fairly trivial. Would we be willing to invest a lot of time and energy into improving by just over one metre/yard? However, if we consider the larger picture, in a half marathon (21,098 metres or 23,073 yards) the effect of the training would mean that you improve your time by 38 seconds.

This example demonstrates the basic thinking about effects and effect sizes that we can also apply to linguistic data. ES measures, as we know, quantify the observed linguistic variables and the differences and changes in their frequencies. It is important to gain a good understanding of what exactly the measures mean and how to interpret them. In our half marathon example, we can show easily what we mean by one metre or one yard by drawing a line that represents the unit of measurement. However, we need to realize that the decision about the practical importance of the effect has little to do with statistics: in our example, it is our personal decision about whether a 38-second improvement in a half marathon is worth the training time and energy. To make an informed decision, the 38 seconds need to be contextualized and compared with the times of other runners etc.

This section focuses on the use of effect size (ES) measures in corpus linguistics. The concept of effect size was introduced in Chapter 1 and different effect size measures have been discussed throughout this book. Table 8.8 shows an overview of ES measures used in this book including a brief description of each measure. These range from simple statistics such as the mean or percentage

Table 8.8 *Effect size measures introduced in this book*

Effect size measure	Brief description	Comparable across different linguistic variables?	Book section
Mean	Basic descriptive statistic showing the frequency of a linguistic variable in a corpus.	NO	1.2
% change	An indication of increase/decrease of the frequency of a linguistic variable between two points in time.	YES	7.3
MI-score, t-score, log Dice etc.	Association measures for identifying the colocation relationship between words.	YES; Note that t-score depends on corpus size and hence it is a problematic ES measure	3.2
simple maths parameter, %DIF etc.	Statistics for identifying keywords.	YES	3.4
Cohen's d	A standardized measure for expressing the difference between the frequency of a linguistic variable in two (sub)corpora in standard deviation units.	YES	1.3; 3.2; 6.3; 7.3
robust Cohen's d	A robust version of the d measure (see above) calculated with trimmed means and winsorized SDs .	YES	6.3
r	Best known as the correlation coefficient but also used as a standardized effect size measure in different situations.	YES	1.3; 5.2; 6.3
r_s	A version of r used for ordinal data.	YES	5.2
r_{rb} (rank biserial correlation)	A type of correlation used as an effect size measure for the Mann–Whitney U test.	YES	6.3
r^2 (coefficient of determination)	A measure expressing the amount of variation accounted for by an explanatory variable (predictor).	YES	5.2, 5.4
η^2 (eta squared)	An omnibus effect size measure traditionally reported with the ANOVA test.	YES	1.3; 6.3
PR (probability ratio), also known as risk ratio or relative risk	A ratio of two probabilities with which two variants of a linguistic variable occur in a particular context.	YES	4.3
(log) odds ratio	A measure used to express the effect of a predictor in the output of logistic regression and mixed effects models.	YES	4.3, 4.4
OR/ln(OR)	An effect size measure showing the association between nominal variables (categorical data).	YES	4.3
Cramer's V			

Table 8.9 *Effect size transformation and extrapolation*

Input	Output	Transformation/ Extrapolation	Example
r	Cohen's d	$d = \frac{2r}{\sqrt{1 - r^2}}$	$r = 0.3; d = \frac{2 \times 0.3}{\sqrt{1 - 0.3^2}} = 0.629$
Cohen's d	r	$r = \frac{d}{\sqrt{d^2 + 4}}$	$d = 0.5; r = \frac{0.5}{\sqrt{0.5^2 + 4}} = 0.25$
η^2 [similar group sizes]	Cohen's d	$d = \frac{2 \times \sqrt{\eta^2}}{\sqrt{1 - \eta^2}}$	$\eta^2 = 0.01; d = \frac{2 \times \sqrt{0.01}}{\sqrt{1 - 0.01}} = 0.2$
log odds ratio [$\ln(\text{OR})$]	Cohen's d	$d = \frac{\ln(\text{OR}) \times \sqrt{3}}{\pi}$	$\ln(\text{OR}) = 0.9; d = \frac{0.9 \times \sqrt{3}}{3.14} = 0.5$
t-test $t; n_1; n_2$	Cohen's d	$d = t \times \sqrt{\frac{n_1 + n_2}{n_1 \times n_2}}$	$t = 3; n_1=50; n_2=25;$ $d = 3 \times \sqrt{\frac{50 + 25}{50 \times 25}} = 0.73$
one-way ANOVA $F; n_1; n_2$	Cohen's d	$d = \pm \sqrt{\frac{F(n_1 + n_2)}{(n_1 \times n_2)}}$	$F = 9; n_1=50; n_2=25;$ $d = \pm \sqrt{\frac{9 \times (50 + 25)}{50 \times 25}} = 0.73$

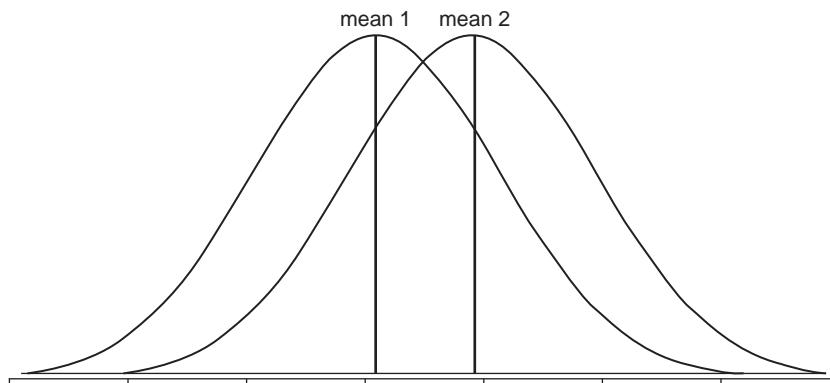
change to more complex measures such as the association measures used for identifying collocation, Cohen's d , η^2 (log)odds etc.

Two questions in particular will be addressed: (1) How should we deal with inconsistent reporting of effect sizes? and (2) How exactly should effect size measures be interpreted? Because different studies may report a variety of ES measures or might not report these measures at all we often need to carry out transformations or extrapolate the effect size from statistical tests (e.g. t-test, ANOVA) and sample sizes ($n_1; n_2$). The transformation and extrapolation are important when performing meta-analysis where we often have to put together incomplete or inconsistent pieces of information. Table 8.9 shows basic extrapolations and transformations that can be carried out in corpus research (Borenstein 2009; Fritz et al. 2012).

To address the second question (how to interpret ES measures?), we need to gain a good understanding of how a particular ES measure works in practice. In textbooks, ES measures are usually ascribed standard interpretations, typically based on Cohen's (1988) recommendations, which divide the size of the observed effect into 'small', 'medium' and 'large'. Table 8.10 provides a summary of the standard interpretation of four common ES measures (Cohen's d , r , η^2 and

Table 8.10 *Effect size: standard interpretation*

Effect	ES measure	<i>r</i>	Cohen's <i>d</i>	η^2	Cramer's <i>V</i> [2 × 2 table]
Small		0.1	0.3	0.01	0.1
Medium		0.3	0.5	0.06	0.3
Large		0.5	0.8	0.14	0.5

Figure 8.6 *Comparison of two subcorpora*

Cramer's *V*). However, although useful as broad guidelines, these standard interpretations, as Cohen (1988) stresses repeatedly, shouldn't be applied mechanically; on the contrary, each discipline should review the range of effects reported in the studies and critically assess their practical implications by asking the question of how the presence/absence of the target linguistic variable(s) affects the grammar, text type, speaker group, discourse etc.

Because effect sizes are relatively new in corpus linguistics, the rest of the discussion will provide a guide to interpreting the practical importance of standardized ES measures; standardized ES measures such as Cohen's *d* or *r* operate on a standardized scale which is different from the scale on which the variables are measured (e.g. frequency per million). One of the ways to interpret standardized effect sizes such as Cohen's *d*, *r*, or η^2 is to relate them to the concept/measure of probability of superiority, which can be understood more intuitively. **Probability of superiority (PS)** is the probability that a speaker/text picked randomly from the subcorpus with a larger mean value of the target variable will have a higher score than a speaker/text picked randomly from the subcorpus with a lower mean value. This can be graphically demonstrated in Figure 8.6.⁵ The two bell-shaped curves represent the distribution of values in subcorpus 1 and

⁵ For a dynamic version of this figure see Kristoffer Magnusson's visualization of Cohen's *d* at <http://rpsychologist.com/d3/cohend/>

subcorpus 2;⁶ means are indicated by the thick vertical lines. We can see that the curves overlap to a large extent (69% in Figure 8.6). Yet, there is a clear difference between these two subcorpora with the PS value of 71%, which means that in 71% of the cases a randomly selected speaker/text from subcorpus 2 will have a higher value of the variable of interest than a randomly selected speaker/text from subcorpus 1. The PS value 71 is associated with 0.8 Cohen's d value (see Table 8.11), which in the standard interpretation is a large effect.

To see the practical impact of corpus linguistic effects expressed as d , r , η^2 (r^2), Table 8.11 provides a comparison between these values and their relevant PS values in percentages,⁷ showing the amount of superiority of one subcorpus over another in the terms described above. In addition, several benchmark points are illustrated with a sparkline and a linguistic example based on the *British National Corpus* (BNC). Readers are encouraged to supply further benchmark points to the table based on their research or research found in the literature, thus creating an even more

Table 8.11 *Effect size measures: BNC validation*

	d	r	r^2 or η^2	PS (%)	Linguistic example from the BNC
	0	0	0	50	<i>the</i> in two randomly selected subcorpora of the BNC
	0.1	0.05	0.002	53	
	0.2	0.1	0.01	56	
	0.3	0.15	0.022	58	
	0.4	0.2	0.038	61	
	0.5	0.24	0.059	64	
	0.6	0.29	0.083	66	
	0.7	0.33	0.11	69	<i>lovely</i> in female vs male speech
	0.8	0.37	0.14	71	
	0.9	0.41	0.17	74	
	1	0.45	0.2	76	
	1.1	0.48	0.23	78	
	1.2	0.51	0.27	80	
	1.3	0.55	0.3	82	
	1.4	0.57	0.33	84	

⁶ This is an idealized picture. In reality, distributions of linguistic variables are typically positively skewed (i.e. with a long right tail as in the following sparkline).

⁷ The overview of the d , r , η^2 (r^2) and PS values is taken from Fritz et al. (2012: 8).

Table 8.11 (*cont.*)

d	r	r^2 or η^2	PS (%)	Linguistic example from the BNC
	1.5	0.6	0.36	86
	1.6	0.63	0.39	87
	1.7	0.65	0.42	89
	1.8	0.67	0.45	90
	1.9	0.69	0.47	91
	2	0.71	0.5	92
	2.2	0.74	0.55	94
	2.4	0.77	0.59	96
	2.6	0.79	0.63	97
	2.8	0.81	0.66	98
	3	0.83	0.69	98
	3.2	0.85	0.72	99
	3.4	0.86	0.74	99
	3.6	0.87	0.76	99
	3.8	0.89	0.78	100
	4	0.89	0.8	100

nuanced scale. This scale of effects can in turn help contextualize and interpret effects observed in individual studies.

8.5 Exercises

- What is the most important thing you have learnt from this book? Write this down in the space below.

- Provide transformations of the effect size measures in Table 8.12.

Table 8.12 Effect size transformations

Input	Output (transformation)
$r = 0.9$	Cohen's $d =$
Cohen's $d = 1.3$	$r =$
$\eta^2 = 0.05$	Cohen's $d =$
$\ln(OR) = 0.2$	Cohen's $d =$
$t = 2; n_1 = 100; n_2 = 100$	Cohen's $d =$
$F = 10; n_1 = 100; n_2 = 100$	Cohen's $d =$

3. Interpret the forest plots in Figures 8.7 and 8.8.

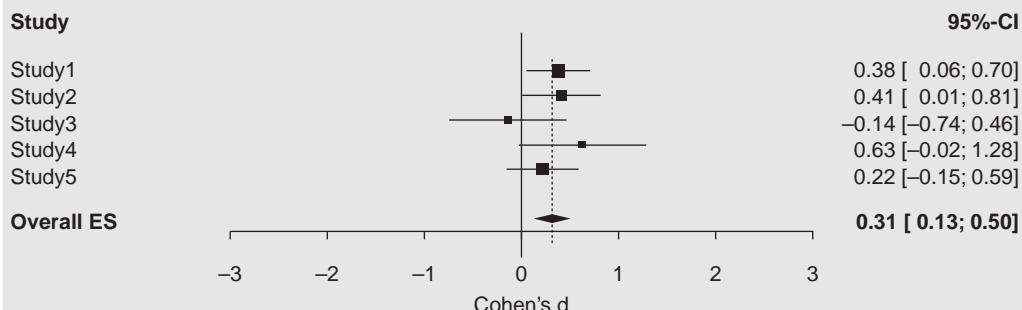


Figure 8.7 Forest plot: example 1

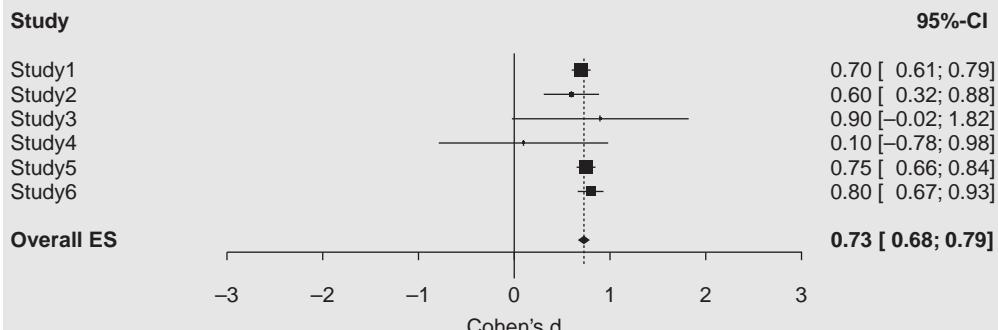


Figure 8.8 Forest plot: example 2

THINGS TO REMEMBER

- Statistics helps us express quantitative information with precision and rigour.
- Meta-analysis provides a statistical summary of multiple studies by combining their effect sizes.
- The results of meta-analysis can be visualized using a forest plot.
- To deal with inconsistent reporting of effect sizes, we can convert one effect size measure into another or extrapolate it.
- Standardized effect size measures can be understood in terms of the probability of superiority.
- Effect size measures can be interpreted with the help of benchmark points, which show examples of easily imaginable linguistic effects and the corresponding values of common effect size measures.

Advanced Reading

- Borenstein, M., Hedges, L. V., Higgins, J. & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester: John Wiley & Sons.
- Brezina, V. & Meyerhoff, M. (2014). Significant or random? A critical review of sociolinguistic generalisations based on large corpora. *International Journal of Corpus Linguistics*, 19(1), 1–28.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
- Cooper, H., Hedges, L. & Valentine, J. (eds.) *The handbook of research synthesis and meta-analysis*. New York: Russell Sage Foundation.
- Cumming, G. (2012). *Understanding the new statistics*. New York: Routledge.

Companion Website: Lancaster Stats Tools Online

1. All analyses shown in this chapter can be reproduced using the statistical tools available from the companion website. The tools available for this chapter include:
 - Effect size calculator
 - Meta-analysis calculator
2. The website also offers additional materials for students and teachers.

Final Remarks

Statistics is a powerful analytical tool. This book has demonstrated a number of different ways in which statistical techniques can be used to explore corpora. The robust evidence found in these electronic collections of language offers countless possibilities for both linguistic and social research providing a unique insight into patterns of language use. Statistics, if applied appropriately, can facilitate the process of analysis by serving as a zoom lens through which we can observe the linguistic reality: the details of individual examples of language use as well as the larger picture of grammar, vocabulary and discourse. We need to remember, however, that the lens should always be a transparent one: what we want to observe is not the tools themselves – a showcase full of sophisticated statistical techniques – but the linguistic data. Our analysis thus should always be primarily focused on the data and should take data seriously; if our beliefs and theories are contradicted by the data we shouldn't simply dismiss the data as ‘inconvenient’ evidence (or hide it behind complex statistical jargon) but, on the contrary, we should engage with it, seeking to genuinely understand and explain the findings. Only in this way can our investigation be meaningful and truly scientific.

Mastery of statistics is empowering. However, as statistical tools and analyses become more complex and sophisticated, they can also become rather daunting for the users. This is because statistical analysis involves many choices. Among other things, we need to select a suitable corpus, an effective analytical technique and an appropriate interpretation of the results. These decisions can often feel challenging especially for novice researchers. The growing demand in corpus linguistics for statistical sophistication and the lack of appropriate resources for beginner and intermediate users of statistics can thus easily lead to frustration. This book (together with Lancaster Stats Tools online) hopes to be a resource addressing this issue by offering readers a guide for making informed choices about statistics in language analysis. The main message of the volume is twofold. First, statistics is not about number crunching or remembering equations (computers are much better at these tasks than humans) but about understanding core, underlying principles of quantitative analysis. Second, I would like to encourage the readers not to let themselves be overwhelmed by the complex statistical techniques or the newest fads on the statistical marketplace. Every summer, a large number

of students from different parts of the world come to Lancaster to learn about corpora and statistics during a week of Lancaster summer schools in corpus linguistics. These students often ask me what the best statistical test is to use with corpora, what the best collocation measure is etc. I usually respond: in many cases, the most powerful statistical technique is common sense.

References

- Argamon, S., Koppel, M., Fine, J. & Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts. *Text*, 23(3), 321–46.
- Argamon, S., Koppel, M., Pennebaker, J. W. & Schler, J. (2007). Mining the blogosphere: age, gender and the varieties of self-expression. *First Monday*, 12(9). http://firstmonday.org/issues/issue12_9/argamon/index.html
- Arppe, A. (2008). Univariate, bivariate, and multivariate methods in corpus-based lexicography: a study of synonymy. Helsinki: University of Helsinki. Available at <https://helda.helsinki.fi/bitstream/handle/10138/19274/univaria.pdf?sequence=2> (accessed 29/12/2015).
- Azen, R. & Walker, C. M. (2011). *Categorical data analysis for the behavioral and social sciences*. London: Routledge.
- Baker, H., Brezina, V. & McEnery, T. (2017). Ireland in British parliamentary debates 1803–2005: plotting changes in discourse in a large volume of time-series corpus data. In T. Säily, A. Nurmi, M. Palander-Collin & A. Auer (eds.), *Exploring future paths for historical sociolinguistics* (Advances in Historical Sociolinguistics), pp. 83–107. Amsterdam: John Benjamins.
- Baker, P. (2009). The BE06 corpus of British English and recent language change. *International Journal of Corpus Linguistics*, 14(3), 312–37.
- (2011). Times may change, but we will always have money: diachronic variation in recent British English. *Journal of English Linguistics*, 39(1), 65–88.
- (2014). *Using corpora to analyze gender*. London: Bloomsbury.
- (2017) *American and British English: divided by a common language*. Cambridge University Press.
- Baker, P., Gabrielatos, C. & McEnery, T. (2013). Sketching Muslims: a corpus driven analysis of representations around the word ‘Muslim’ in the British press 1998–2009. *Applied Linguistics*, 34(3), 255–78.
- Baker, P., Gabrielatos, C., Khosravinik, M., Krzyżanowski, M., McEnery, T. & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, 19(3), 273–306.
- Balakrishnan, N., Voinov, V. & Nikulin, M. S. (2013). *Chi-squared goodness of fit tests with applications*. Waltham, MA: Academic Press.
- Barlow, M. (2013). Individual differences and usage-based grammar. *International Journal of Corpus Linguistics*, 18(4), 443–78.
- Baroni, M., Bernardini, S., Ferraresi, A. & Zanchetta, E. (2009). The WaCky Wide Web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43 (3), 209–26.

- Baroni, M. & Ueyama, M. (2006). Building general-and special-purpose corpora by web crawling. In *Proceedings of the 13th NIJL International Symposium, Language corpora: their compilation and application*, pp. 31–40.
- Benzécri, J. P. (1992). *Correspondence analysis handbook*. New York: Marcel Dekker.
- Bestgen, Y. (2014). Inadequacy of the chi-squared test to examine vocabulary differences between corpora. *Literary and Linguistic Computing*, 29(2), 164–70.
- Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.
- Biber, D. & Conrad, S. (2009). *Register, genre, and style*. Cambridge University Press.
- Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.
- Biber, D. & Jones, K. (2009). Quantitative methods in corpus linguistics. In A. Lüdeling & M. Kytö (eds.), *Corpus linguistics: an international handbook*, vol. 2, pp. 1287–1304. Berlin: Walter de Gruyter.
- Biber, D., Reppen, R., Schnur, E. & Ghanem, R. (2016). On the (non) utility of Julland's D to measure lexical dispersion in large corpora. *International Journal of Corpus Linguistics*, 21(4), 439–64.
- Blythe, R. A. & Croft, W. (2012). S-curves and the mechanisms of propagation in language change. *Language*, 88(2), 269–304.
- Boneau, C. A. (1960). The effects of violations of assumptions underlying the t test. *Psychological Bulletin*, 57(1), 49.
- Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. Hedges & J. Valentine (eds.), *The handbook of research synthesis and meta-analysis*, pp. 221–35. New York: Russell Sage Foundation.
- Brezina, V. (2013). BNC64 Search & Compare. Available at: <http://corpora.lancs.ac.uk/bnc64> (accessed 20/08/2016).
- (2014). Effect sizes in corpus linguistics: keywords, collocations and diachronic comparison. Presented at the ICAME 2014 conference, University of Nottingham.
- Brezina, V. & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the New General Service List. *Applied Linguistics*, 36(1), 1–22.
- Brezina, V., McEnery, T. & Baker, H. (in prep.) Usage fluctuation analysis: a new way of analysing shifts in historical discourse.
- Brezina, V., McEnery, T. & Wattam, S. (2015). Collocations in context. *International Journal of Corpus Linguistics*, 20(2), 139–73.
- Brezina, V. & Meyerhoff, M. (2014). Significant or random? A critical review of socio-linguistic generalisations based on large corpora. *International Journal of Corpus Linguistics*, 19(1), 1–28.
- Brezina, V. & Timperley, M. (2017). How large is the BNC? A proposal for standardized tokenization and word counting. CL2017, Birmingham. Available at: www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2017/general/paper303.pdf (accessed 08/03/18).
- Brezina, V., Timperley, M., Gablasova, D. & McEnery, T. (in prep.). #LancsBox: a new generation corpus tool for researchers, students and teachers.
- Cabin, R. J. & Mitchell, R. J. (2000). To Bonferroni or not to Bonferroni: when and how are the questions. *Bulletin of the Ecological Society of America*, 81(3), 246–8.
- Chernick, M. R. & LaBudde, R. A. (2014). *An introduction to bootstrap methods with applications to R*. Hoboken, NJ: John Wiley & Sons.

- Chomsky, N. (2000). *New horizons in the study of language and mind*. Cambridge University Press.
- Clausen, S. E. (1998). *Applied correspondence analysis: an introduction*. Thousand Oaks, CA: Sage.
- Cleveland, W. S. (1994). *The elements of graphing data*. Summit, NJ: Hobart Press.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
- Cohen, M. P. (2000). Note on the odds ratio and the probability ratio. *Journal of Educational and Behavioral Statistics*, 25(2), 249–52.
- Colley, A. & Todd, Z. (2002). Gender-linked differences in the style and content of e-mails to friends. *Journal of Language and Social Psychology*, 21(4), 380–92.
- Conrad, S. & Biber, D. (2001). Multidimensional methodology and the dimensions of register variation in English. In S. Conrad & D. Biber (eds.), *Variation in English: multidimensional studies*, pp. 18–19. Harlow: Pearson Education.
- Coupland, N. (2007). *Style: language variation and identity*. Cambridge University Press.
- Covington, M. A. & McFall, J. D. (2010). Cutting the Gordian knot: the moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2), 94–100.
- Crystal, D. (2003). *English as a global language*. Cambridge University Press.
- Cumming, G. (2012). *Understanding the new statistics*. New York: Routledge.
- Cumming, G., Fidler, F. & Vaux, D. L. (2007). Error bars in experimental biology. *Journal of Cell Biology*, 177(1), 7–11.
- Davies, H. T. O., Crombie, I. K. & Tavakoli, M. (1998). When can odds ratios mislead? *British Medical Journal*, 316(7136), 989–91.
- Davies, M. & Gardner, D. (2010). *A frequency dictionary of contemporary American English: word sketches, collocates and thematic lists*. London: Routledge.
- de Winter, J. C. (2013). Using the Student's t-test with extremely small sample sizes. *Practical Assessment, Research & Evaluation*, 18(10), 1–12.
- Diggle, P. J. & Chetwynd, A. G. (2011). *Statistics and scientific method: an introduction for students and researchers*. Oxford University Press.
- Divjak, D. & Gries, S. Th. (2006). Ways of trying in Russian: clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory*, 2(1), 23–60.
- Dodge, Y. (2008). *The concise encyclopedia of statistics*. New York: Springer.
- Edgell, S. E. & Noon, S. M. (1984). Effect of violation of normality on the t test of the correlation coefficient. *Psychological Bulletin*, 95(3), 576.
- Efron, B. (1979). Computers and the theory of statistics: thinking the unthinkable. *SIAM Review*, 21(4), 460–80.
- Efron, B. & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Boca Raton, FL: CRC Press.
- Erceg-Hurn, D. M. & Mirosevich, V. M. (2008). Modern robust statistical methods: an easy way to maximize the accuracy and power of your research. *American Psychologist*, 63(7), 591.
- Everitt, B. S., Landau, S., Leese, M. & Stahl, D. (2011). *Cluster analysis*. New York: John Wiley & Sons.
- Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (eds.), *Corpus linguistics: an international handbook*, vol. 1, pp. 223–33. Berlin: Walter de Gruyter.
- Field, A., Miles, J. & Field, Z. (2012). *Discovering statistics using R*. London: Sage.
- Firth, J. (1957). *Papers in linguistics*. Oxford University Press.

- Francis, W. N. & Kučera, H. (1979). *Brown Corpus manual: manual of information to accompany a standard corpus of present-day edited American English for use with digital computers*. Brown University, Providence, RI. Available at <http://clu.uni.no/icame/brown/bcm.html>
- Friendly, M. (2002). A brief history of the mosaic display. *Journal of Computational and Graphical Statistics*, 11(1), 89–107.
- Frigolí, E. & Hardy, J. (2014). Conducting multi-dimensional analysis using SPSS. In T. B. Sardinha & M. V. Pinto (eds.), *Multi-dimensional analysis, 25 years on: a tribute to Douglas Biber*, pp. 297–316. Amsterdam: John Benjamins.
- Fritz, C. O., Morris, P. E. & Richler, J. J. (2012). Effect size estimates: current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, 141(1), 2–18.
- Gablasová, D., Brezina, V. & McEnery, A. M. (2017a). Exploring learner language through corpora: comparing and interpreting corpus frequency information. *Language Learning*, 67(S1), 130–54.
- (2017b). Collocations in corpus-based language learning research: identifying, comparing and interpreting the evidence. *Language Learning*, 67(S1), 155–79.
- Gablasová, D., Brezina, V., McEnery, T. & Boyd, E. (2017). Epistemic stance in spoken L2 English: the effect of task and speaker style. *Applied Linguistics*, 38(5), 613–37.
- Gabrielatos, C. & Marchi, A. (2012) Keyness: appropriate metrics and practical issues. Presented at CADS International Conference 2012, Corpus-assisted Discourse Studies: More than the sum of Discourse Analysis and computing? University of Bologna, Italy.
- Glass, G. V. (1965) A ranking variable analogue of biserial correlation: implications for short-cut item analysis. *Journal of Educational Measurement*, 2(1), 91–5.
- Greenacre, M. (2007). *Correspondence analysis in practice*. Boca Raton: Chapman & Hall/CRC.
- Gries, S. Th. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 13(4), 403–37.
- (2010). Dispersions and adjusted frequencies in corpora: further explorations. In S. Th. Gries, S. Wulff & M. Davies, *Corpus linguistic applications: current studies*, pp. 197–212. Amsterdam: Rodopi.
- (2013a). *Statistics for linguistics with R: a practical introduction*. Berlin: Walter de Gruyter.
- (2013b). 50-something years of work on collocations: what is or should be next ... *International Journal of Corpus Linguistics*, 18(1), 137–66.
- Gries, S. Th. & Hilpert, M. (2008). The identification of stages in diachronic data: variability-based neighbour clustering. *Corpora*, 3(1), 59–81.
- (2010). Modeling diachronic change in the third person singular: a multifactorial, verb- and author-specific exploratory approach. *English Language and Linguistics*, 14(03), 293–320.
- Gries, S. Th., Newman, J., Shaoul, C. & Dilts, P. (2009). N-grams and the clustering of genres. Presented at workshop on Corpus, Colligation, Register Variation at the 31st Annual Meeting of the Deutsche Gesellschaft für Sprachwissenschaft, March.
- Grieve-Smith, A. (2007). The envelope of variation in multidimensional register and genre analyses. In E. Fitzpatrick (ed.), *Corpus linguistics beyond the word: corpus research from phrase to discourse*, pp. 21–42. Amsterdam: Rodopi.

- Gwet, K. (2002). Inter-rater reliability: dependency on trait prevalence and marginal homogeneity. *Statistical Methods for Inter-Rater Reliability Assessment Series*, 2, 1–9.
- Hand, D. J. (2010). Evaluating diagnostic tests: the area under the ROC curve and the balance of errors. *Statistics in Medicine*, 29(14), 1502–10.
- Hardie, A. (2014) Log ratio – an informal introduction. <http://cass.lancs.ac.uk/?p=1133>
- Harrington, J., Palethorpe, S. & Watson, C. I. (2000). Does the Queen speak the Queen's English? *Nature*, 408(6815), 927–8.
- Hayton, J. C., Allen, D. G. & Scarfello, V. (2004). Factor retention decisions in exploratory factor analysis: a tutorial on parallel analysis. *Organizational Research Methods*, 7(2), 191–205.
- Healey, A. diPaolo (ed.) (2004). *The Complete Corpus of Old English in Electronic Form*. Dictionary of Old English Project. Centre for Medieval Studies, University of Toronto.
- Hill, T., Lewicki, P. & Lewicki, P. (2006). *Statistics: methods and applications: a comprehensive reference for science, industry, and data mining*. Tulsa, OK: StatSoft.
- Hilpert, M. (2011). Dynamic visualizations of language change: motion charts on the basis of bivariate and multivariate data from diachronic corpora. *International Journal of Corpus Linguistics*, 16(4), 435–61.
- Hilpert, M. & Gries, S. Th. (2009). Assessing frequency changes in multistage diachronic corpora: applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing*, 24(4), 385–401.
- Hosmer, D. W., Lemeshow, S. & Sturdivant, R. X. (2013). *Applied logistic regression*, 3rd edn. Hoboken, NJ: John Wiley & Sons.
- Hudson, T. (2015). Presenting quantitative data visually. In L. Plonsky (ed.), *Advancing quantitative methods in second language research*, pp. 78–105. London: Routledge.
- Ito, R. & Tagliamonte, S. (2003). Well weird, right dodgy, very strange, really cool: layering and recycling in English intensifiers. *Language in Society*, 32(02), 257–79.
- Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P. & Suchomel, V. (2013). The TenTen corpus family. *Proceedings of the International Conference on Corpus Linguistics 2013*, pp. 125–7. Lancaster University.
- Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning*, 63(s1), 87–106.
- Johnson, D. E. (2009). Getting off the GoldVarb standard: introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass*, 3(1), 359–83.
- Julland, A. G., Brodin, D. R. & Davidovitch, C. (1970). *Frequency dictionary of French words*. The Hague: Mouton.
- Julland, A. G. & Chang-Rodríguez, E. (1964). *Frequency dictionary of Spanish words*. The Hague: Mouton.
- Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20(1), 141–51.
- Kepes, S., Banks, G. C. & Oh, I. S. (2014). Avoiding bias in publication bias research: the value of ‘null’ findings. *Journal of Business and Psychology*, 29(2), 183–203.
- Kerby, D. S. (2014). The simple difference formula: an approach to teaching non-parametric correlation. *Innovative Teaching*, 3, 1–9.
- Kilgarriff, A. (1997). Putting frequencies in the dictionary. *International Journal of Lexicography*, 10(2), 135–55.
- (2005). Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory*, 1(2), 263–76.

- (2009). Simple maths for keywords. In *Proceedings of the Corpus Linguistics Conference*, Liverpool, July.
- (2012). Getting to know your corpus. In *Proceedings of the 15th International Conference on Text, Speech and Dialogue*, pp. 3–15. Berlin: Springer.
- Kirk, R. E. (1996). Practical significance: a concept whose time has come. *Educational and Psychological Measurement*, 56(5), 746–59.
- (2005). Effect size measures. *Wiley StatsRef: Statistics Reference Online*. <http://dx.doi.org/10.1002/9781118445112.stat06242.pub2>
- Krippendorff, K. (2012 [1980]). *Content analysis: an introduction to its methodology*. London: Sage.
- Kruskal, W. H. & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260), 583–621.
- Kučera, H. & Francis, W. N. (1967). *Computational analysis of Present-Day American English*. Providence, RI: Brown University Press.
- Labov, W. (1966). *The social stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.
- (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.
- (2010). *Principles of linguistic change*, vol. 3: *Cognitive and cultural factors*. Oxford: Wiley-Blackwell.
- Lakoff, G. & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.
- Lakoff, R. T (1975). *Language and woman's place*. New York: Harper & Row.
- Lavandera, B. R. (1978). Where does the sociolinguistic variable stop? *Language in Society*, 7(02), 171–82.
- Leedesma, R. D. & Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: an easy-to-use computer program for carrying out parallel analysis. *Practical Assessment, Research & Evaluation*, 12(2), 1–11.
- Leech, G. (1992). Corpora and theories of linguistic performance. In J. Svartvik (ed.), *Directions in corpus linguistics*, pp. 105–22. Berlin: Mouton de Gruyter.
- (2003). Modals on the move: the English modal auxiliaries 1961–1992. In R. Facchinetto, F. R. Palmer & M. Krug (eds.), *Modality in contemporary English*, 223–40. Berlin: Mouton de Gruyter.
- (2011). The modals ARE declining. *International Journal of Corpus Linguistics*, 16(4), 547–64.
- Leech, G., Garside, R. & Bryant, M. (1994). CLAWS4: the tagging of the British National Corpus. In *Proceedings of the 15th Conference on Computational Linguistics*, Kyoto, vol. 1, pp. 622–8.
- Leech, G., Rayson, P. & Wilson, A. (2001). *Word frequencies in written and spoken English: based on the British National Corpus*. London: Routledge.
- Leek, J. T. & Peng, R. D. (2015). Statistics: p values are just the tip of the iceberg. *Nature*, 520(7549), 612.
- Lijffijt, J., Nevalainen, T., Säily, T., Papapetrou, P., Puolamäki, K. & Mannila, H. (2016). Significance testing of word frequencies in corpora. *Literary and Linguistic Computing*, 31(2), 374–97.
- Lijffijt, J., Säily, T. & Nevalainen, T. (2012). CEECing the baseline: lexical stability and significant change in a historical corpus. In *Studies in Variation, Contacts and Change in English*, vol. 10. Helsinki: Research Unit for Variation, Contacts and Change in English (VARIENG).

- Love, R., Dembry, C., Hardie, A., Brezina, V. & McEnery, T. (2017). The Spoken BNC2014: designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22 (3).
- Lumley, T., Diehr, P., Emerson, S. & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, 23(1), 151–69.
- Malvern, D. & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, 19, 85–104.
- Mann, H. B. & Whitney, D. R. (1947) On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18(1), 50–60.
- Manning, C. D. (2011). Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In A. F. Gelbukh (ed.), *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 171–89. Berlin: Springer.
- McEnery, T. (2006). *Swearing in English: bad language, purity and power from 1586 to the present*. Abingdon: Routledge.
- McEnery, T. & Baker, H. (2017). *Corpus linguistics and 17th-century prostitution: computational linguistics and history*. London: Bloomsbury.
- McEnery, T. & Hardie, A. (2011). *Corpus linguistics: method, theory and practice*. Cambridge University Press.
- Mehl, M. R., Vazire, S., Ramirez-Esparza, N., Slatcher, R. B. & Pennebaker, J. W. (2007). Are women really more talkative than men? *Science*, 317(5834), 82.
- Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P. . . & Pinker, S. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014), 176–82.
- Microsoft (2010). *Microsoft Word* [software].
- Millar, N. (2009). Modal verbs in TIME: frequency changes 1923–2006. *International Journal of Corpus Linguistics*, 14 (2), 191–220.
- Nevalainen, T. (1999). Making the best use of ‘bad’ data: evidence for sociolinguistic variation in Early Modern English. *Neophilologische Mitteilungen*, 499–533.
- Nevalainen, T. & Raumolin-Brunberg, H. (2003). *Historical sociolinguistics: language change in Tudor and Stuart England*. London: Routledge.
- Newman, M. L., Groom, C. J., Handelman, L. D. & Pennebaker, J. W. (2008). Gender differences in language use: an analysis of 14,000 text samples. *Discourse Processes*, 45(3), 211–36.
- Nini, A. (2015) *Multidimensional Analysis Tagger (v. 1.3) – manual*. Available at: <https://sites.google.com/site/multidimensionaltagger/> (accessed 26/08/15).
- Osborne, J. W. (2012). *Best practices in data cleaning: a complete guide to everything you need to do before and after collecting your data*. Thousand Oaks, CA: Sage.
- (2015). *Best practices in logistic regression*. Thousand Oaks, CA: Sage.
- Pearson, K. (1920). Notes on the history of correlation. *Biometrika*, 13(1), 25–45.
- Pechenick, E. A., Danforth, C. M. & Dodds, P. S. (2015). Characterizing the Google Books corpus: strong limits to inferences of socio-cultural and linguistic evolution. *PloS One*, 10(10), e0137041.
- Phillips, M. (1985). *Aspects of text structure: an investigation of the lexical organisation of text*. Amsterdam: North-Holland.
- Popper, K. (2005 [1935]). *The logic of scientific discovery*. London: Routledge.

- Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics* 13(4), 519–49.
- Rayson, P., Berridge, D. & Francis, B. (2004). Extending the Cochran rule for the comparison of word frequencies between corpora. *Proceedings from 7th International Conference on Statistical Analysis of Textual Data (JADT 2004)*, pp. 926–36.
- Rayson, P., Leech, G. N. & Hodges, M. (1997). Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics*, 2(1), 133–52.
- Richardson, J. T. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review*, 6(2), 135–47.
- Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin*, 118(2), 183.
- Rothstein, H. & Hopewell, S. (2009). Grey literature. In H. Cooper, L. Hedges & J. Valentine (eds.), *The handbook of research synthesis and meta-analysis*, pp. 103–26. New York: Russell Sage Foundation.
- Savický, P. & Hlaváčová, J. (2002). Measures of word commonness. *Journal of Quantitative Linguistics*, 9(3), 215–31.
- Schmidler, E., Ziegler, M., Danay, E., Beyer, L. & M. Bühner. (2010). ‘Is it really robust?’ Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 6(4), 147–51.
- Schmidt, F. L. & Hunter, J. E. (2015). *Methods of meta-analysis: correcting error and bias in research findings*. Thousand Oaks, CA: Sage Publications.
- Scott, M. (1997). PC analysis of key words – and key key words. *System*, 25(2), 233–45. (2004). *WordSmith tools version 4*. Oxford University Press.
- Shadish, W. R. & Haddock, C. K. (2009). Combining estimates of effect size. In H. Cooper, L. Hedges & J. Valentine (eds.), *The handbook of research synthesis and meta-analysis*, pp. 257–78. New York: Russell Sage Foundation.
- Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, 46, 561–84.
- Shakespeare, W. (1992). *The Poems: Venus and Adonis, The Rape of Lucrece, The Phoenix and the Turtle, The Passionate Pilgrim*. Cambridge University Press.
- Sheskin, D. J. (2007). *Handbook of parametric and nonparametric statistical procedures*. Boca Raton, FL: Chapman & Hall/CRC.
- Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.
- Sprent, P. (2011). Fisher Exact Test. In *International encyclopedia of statistical science*, pp. 524–5. Berlin: Springer.
- Stubbs, M. (2001). *Words and phrases: corpus studies of lexical semantics*. Oxford: Blackwell.
- Tagliamonte, S. A. (2006). *Analysing sociolinguistic variation*. Cambridge University Press.
- Theus, M. & Urbanek, S. (2008). *Interactive graphics for data analysis: principles and examples*. Boca Raton, FL: CRC Press.
- Toothaker, L. E. (1993). *Multiple comparison procedures*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–089. Newbury Park, CA: Sage.

- Trafimow, D. & Marks, M. (2015) Editorial. *Basic and Applied Social Psychology*, 37(1), 1–2.
- Tufte, E. (1997). *Visual explanations*. Cheshire, CT: Graphics Press.
- (2001). *Visual display of quantitative information*. Cheshire, CT: Graphics Press.
- (2006). *Beautiful evidence*. Cheshire, CT: Graphics Press.
- Tweedie, F. & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, 32, 323–52.
- Upton, G. J. (1992). Fisher's exact test. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 395–402.
- Valentine, J. (2009). Judging the quality of primary research. In H. Cooper, L. Hedges & J. Valentine (eds.), *The handbook of research synthesis and meta-analysis*, pp. 129–46. New York: Russell Sage Foundation.
- Verma, J. P. (2016). *Repeated measures design for empirical researchers*. Hoboken, NJ: John Wiley & Sons.
- Vine, B. (1999). *Guide to the New Zealand component of the International Corpus of English (ICE-NZ)*. School of Linguistics and Applied Language Studies, Victoria University of Wellington.
- Vine, E. W. (2011). High frequency multifunctional words: accuracy of word-class tagging. *Te Reo*, 54, 71.
- Williams, G. (1998). Collocational networks: interlocking patterns of lexis in a corpus of plant biology research articles. *International Journal of Corpus Linguistics*, 3(1), 151–71.
- Wilson, D. B. (2009). Systematic coding. In H. Cooper, L. Hedges & J. Valentine (eds.), *The handbook of research synthesis and meta-analysis*, pp. 159–76. New York: Russell Sage Foundation.
- Xiao, R. (2009). Multidimensional analysis and the study of world Englishes. *World Englishes*, 28(4), 421–50.

Index

- AIC (Akaike information criterion), 123, 124, 125
ANOVA, 13, 170, 192, 194, 195, 197, 198, 276,
 277
 multi-way, 197
 one-way, 169, 170, 177, 192, 194, 262, 277
association measure, 67, 69, 70, 71, 72, 76, 79, 84,
 276, 277
assumption
 of a test, 13
- bias
 types of, 16, 17, 262, 265, 269
bootstrap test, 231, 232, 233, 235, 249
bootstrapping, 195, 231, 232, 233
- case, 6
central tendency
 measures of, 10, 13
chi-squared
 test, 13, 113, 114, 117, 121, 126, 131, 202, 203,
 204, 230, 264
cluster analysis, 153, 154, 159, 173, 174, 236, 237
 hierarchical agglomerative, 154, 159, 236
 variability-based neighbour, 236
coefficient of determination (r^2), 147, 276
coefficient of variation, 50, 51, 52, 237
Cohen's d , 72, 86, 190, 191, 195, 263, 272, 273,
 277, 276, 277, 278
Cohen's κ , 90, 91
collinearity, 120, 121
collocation, 66, 67, 68, 69, 70, 71, 74, 75, 76, 77,
 79, 94, 108, 245, 277
 collocation parameters notation (CPN), 74, 75,
 77, 79
 graph, 75
 network, 76, 77, 79, 94
 span, 67
 window, 69, 70
collocation window, *see* collocation: span
complete separation, 120, 124
concordance index (C-index), 125, 126, 129, 133
confidence interval, 13, 20, 24, 30, 31, 116, 117,
 128, 129, 144, 146, 150, 191, 199, 226, 233,
 235, 242, 243, 246, 268, 272, 273
contingency table, 69, 70, 84, 108, 113, 114, 204
- corpus, 6
 of interest, 80
 population-based, 18
 reference, 80, 81
corpus linguistics, 2
correlation, 121, 141, 142, 144, 146, 147, 148, 150,
 151, 164, 165, 172, 191, 196, 197, 236, 276
 coefficient, 144, 145, 147, 150, 197, 237, 276
 interclass, 91
 matrix, 148, 148, 150, 165, 172
 negative, 141, 144, 149, 151, 164, 172, 176
 Pearson's, 142
 positive, 141, 142, 144, 151, 164
 rank biserial, 196
 Spearman's correlation, 146
correspondence analysis, 200, 202, 204, 205, 206,
 214
covariance, 142, 144, 146
Cramer's V , 114
cross-tabulation, 108, 109, 110, 111, 112, 117, 133,
 200, 202, 204, 206
- data
 sparseness, 19
data point, *see* case
dataset, 6
degrees of freedom (df), 50, 114, 115, 117, 187,
 188, 189, 190, 193, 194, 197
Delta P, 70, 72, 74
Dice, 70, 72, 76
directionality, 70, 74
dispersion, 10, 11, 13, 46, 47, 48, 50, 51, 53, 54, 61,
 70, 74, 85
distance
 chi-squared, 200, 203, 204, 206
 Euclidean, 153, 154, 203, 204
 Manhattan, 153, 159, 173
DP (deviation of proportions), 52
- effect
 fixed, 209, 211
 main, 125
 random, 209, 211
effect size, 14, 20, 30, 32, 91, 114, 115, 116, 117,
 125, 128, 144, 145, 150, 190, 191, 195, 196,

- 197, 232, 233, 235, 262, 270, 272, 273, 274, 275, 276, 277, 278, 279
- eigenvalue, 166
- envelope of variation, 106, 185
- eta squared (η^2), 195, 276
- factor analysis, 164, 164, 165, 170, 172, 174, 176, 200, 205, 239
- Fisher exact test, 113
- Fleiss's κ , 91
- frequency
- absolute, 22, 42, 43, 44, 48, 54, 55, 57
 - average reduced (ARF), 54, 55, 57, 61
 - distribution, 8, 13, 25, 60
 - expected, 70, 71, 84, 113, 114
 - marginal, 109
 - mean, 43, 231
 - observed, 67, 68, 69, 70, 84, 113, 142
- graph
- bar chart, 23, 103, 229
 - boxplot, 23, 30, 228
 - candlestick plot, 226, 228, 248
 - correspondence plot, 200, 202, 205, 206
 - dendrogram, 154, 158, 159, 237, 239, 241, 250
 - error bars, 14, 24, 197, 226, 228
 - forest plot, 273, 275
 - geomapping, 27
 - histogram, 8, 25
 - line graph, 224, 243, 248
 - mosaic plot, 109
 - scatterplot matrix, 25, 26
 - scree plot, 166, 167, 174, 239, 241
 - sparkline, 226, 228, 279
 - stacked bar chart, 103, 259
- Gwt's AC₁, 90, 91
- homoscedasticity, 13, 189, 192
- independence of observations, 112, 120, 121, 189, 192
- information/ink ratio, 23
- intercept, 118, 120, 124, 125, 126, 127
- inter-rater agreement, 87, 88, 90, 91, 92, 95, 129, 208, 245, 246, 247, 274
- keywords, 79, 80, 81, 82, 83, 85, 86, 87, 93, 265, 276
- negative, 80, 81, 82, 83
 - positive, 80, 81, 82, 83, 84, 87, 93
- Kruskal–Wallis test, 195, 196, 197, 199
- lemma, 40, 41, 60, 61
- lexeme, 40, 41, 60
- lexico-grammatical frame, 102, 106, 107, 119, 129, 130
- line of the best fit, *see* regression line
- linearity, 13, 120, 121
- lockwords, 80, 81, 83, 220, 265
- log Dice, 70, 72, 76, 276
- log likelihood, 72, 84, 86, 113, 123, 124, 126, 127, 277
- log odds, 121, 124, 125, 126, 127, 277
- log odds ratio, 127, 277
- log ratio, 72, 74, 85, 86
- logistic regression, 106, 117, 118, 119, 120, 121, 122, 124, 125, 126, 127, 128, 129, 132, 134, 209, 210, 211, 276
- Mann–Whitney *U* test, 13, 195, 196, 197, 231, 276
- mean, 3, 4, 10, 11, 12, 13, 19, 24
- 20% trimmed, 10
 - grand, 192, 193
- meaning fluctuation analysis, 245
- median, 10, 11, 23, 199
- meta-analysis, 14, 267, 268, 269, 270, 271, 272, 273, 274, 277
- MI, 72
- MI2, 70, 72
- MI3, 72
- minimum sensitivity, 72
- mixed-effects models, 121, 208, 209, 211
- model
- baseline (or null), 123
 - parsimonious, 123
- Monte Carlo, 166
- MU, 72
- multidimensional analysis, 170, 262
- NHST (null-hypothesis significance testing), 12, 13, 20
- node, 67, 68, 69, 70, 71, 75, 76, 94, 208, 245, 246
- non-parametric test, 13, 196, 199, 231
- normal distribution, 8, 13, 189, 195
- normalized frequency, *see* relative frequency
- normality, 13, 191, 192
- null hypothesis, 12, 13, 20, 30, 84, 91, 111, 113, 114, 147, 202, 220
- odds, 14, 116, 121, 125, 126, 127, 128, 129, 133, 277, 276
- odds ratio, 14, 116, 127, 128, 129, 276
- outlier, 9, 10, 23, 154
- parallel analysis, 166
- parametric test, 13
- peaks and troughs, 242, 243, 245, 246, 247
- percentage increase/decrease, 230
- phi, 114
- population, 6, 12, 13, 14, 15, 16, 18, 19, 20, 24, 49, 52, 82, 116, 128, 144, 145, 147, 150, 188, 189, 191, 192, 195, 199, 221, 231, 232, 259, 272
- post-hoc test, 194, 195, 197, 199

- predictor, 105, 109, 111, 117, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 132, 177, 208, 276
 interactions, 125
 probability of superiority (PS), 197, 278
 probability ratio, 115, 116
 pseudo- R^2 , 126
 p-value, 12, 13

r (effect size), 14, 191, 192, 277
 range₁, 11, 48
 interquartile, 11, 23
 range₂, 11, 48
 raw agreement, 89, 90
 raw frequency, *see* frequency: absolute
 regression line, 4, 26, 141
 relative frequency, 43
 repeated measures test, 189, 197
 replication, 267
 representativeness, 17, 18, 19, 30, 82, 221, 222, 223, 224, 234, 259
 research design
 individual-text/speaker, 21, 22
 linguistic feature, 22, 105, 106, 108, 119, 270
 whole corpus, 21, 104, 272
 rogue value, 9

 sample, 3, 6
 sampling
 random, 15, 16
 sampling frame, 16, 17, 18, 43, 87, 221, 259
 science, 2
 simple maths parameter (SMP), 85, 86, 276
 standard deviation, 11, 48, 49, 50, 142, 144, 152, 153, 168, 190, 191, 195, 237, 241, 276
 population, 52
 sample, 49, 50, 52, 187
 standard error, 124, 128
 statistical significance, 12, 20
 statistics, 3
 descriptive, 4, 14, 18, 50, 221, 259
 inferential, 12, 13, 19, 20, 117

 token, 39, 41, 54, 57, 58, 68, 270
 t-score, 72, 73
 t-test, 12, 13, 30, 31, 187, 189, 190, 191, 192, 194, 195, 197, 213, 231, 277
 type, 39, 40, 41, 60

 type/token ratio (TTR), 57, 58, 59, 163, 171, 175, 176
 moving average, 58
 standardized, 58

 VARBRUL, 210, 211
 variable, 6, 8, 10, 11, 24, 105, 106, 107, 108, 109, 112, 113, 115, 117, 119, 122, 123, 125, 129, 132, 141, 142, 147, 148, 152, 164, 165, 166, 168, 172, 172
 ambient, 185
 categorical, 108, 118, 119, 121, 122, 154, 208
 dummy, 127
 explanatory, 6, 105, 108, 111, 117, 119, 120, 192, 197, 209, 269, 270
 judgement, 87, 89, 91, 92, 120, 208
 level of, 119
 lexico-grammatical, 106, 107, 117, 120
 linguistic, 6, 8, 13, 14, 19, 22, 25, 30, 38, 103, 105, 107, 108, 109, 110, 113, 118, 119, 123, 125, 139, 141, 142, 148, 151, 153, 158, 161, 163, 164, 166, 170, 171, 172, 174, 185, 187, 189, 190, 195, 196, 214, 224, 228, 229, 230, 232, 236, 237, 242, 262, 274, 275, 276, 278
 linguistic ambient, 107
 moderator, 274
 nominal, 7, 90, 105, 276
 ordinal, 7, 108, 142, 147
 outcome, 105, 107, 118, 119, 121, 122, 128, 129, 132, 208
 predictor, 118, 119, 120, 121, 123, 127, 132, 208
 scale, 7, 8, 119, 122, 141, 142, 147
 sociolinguistic, 185, 187, 207
 sociolinguistic Labovian, 185
 variable context, 106, 207
 variance, 13, 187, 188, 189, 190, 192, 193, 194, 272, 273, 275
 variation
 functional, 104, 161, 170, 185
 percentage of, 51, 206
 register, 104, 177

 Wald's *z*, 128
 whelk problem, 46

 Zipf's law, 44, 46
 z-score₁, 72
 z-score₂, 152, 153, 154