Data and Methods in Corpus Linguistics

Corpus linguistics continues to be a vibrant methodology applied across highly diverse fields of research in the language sciences. With the current steep rise in corpus sizes, computational power, statistical literacy and multipurpose software tools, and inspired by neighbouring disciplines, approaches have diversified to an extent that calls for an intensification of the accompanying critical debate. Bringing together a team of leading experts, this book follows a unique design, comparing advanced methods and approaches current in corpus linguistics, to stimulate reflective evaluation and discussion. Each chapter explores the strengths and weaknesses of different datasets and techniques, presenting a case study and allowing readers to gauge methodological options in practice. Contributions also provide suggestions for further reading, and data and analysis scripts are included in an online appendix. This is an important and timely volume, and will be essential reading for any linguist interested in corpus-linguistic approaches to variation and change.

OLE SCHÜTZLER is Professor for Varieties of English at Leipzig University. Mostly working within the frameworks of quantitative sociolinguistics/sociophonetics and corpus linguistics, he takes a general interest in synchronic and diachronic variation and change in English with a special focus on Scottish Englishes.

JULIA SCHLÜTER is Associate Professor for English Linguistics at the University of Bamberg. Her research interests lie in the areas of phonological and grammatical variation in British and American English past and present, empirical – especially corpus-based – methodologies, and applications of linguistic insights and techniques to the teaching of English.

Data and Methods in Corpus Linguistics

Comparative Approaches

Edited by

Ole Schützler
Leipzig University

Julia Schlüter
University of Bamberg

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

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

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/9781108499644
DOI: 10.1017/9781108589314

© Cambridge University Press 2022

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 2022

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

ISBN 978-1-108-49964-4 Hardback

Additional resources for this publication can be found at www.cambridge.org/schuetzler-schlueter

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.

Contents

List of Figures vii
List of Tables xii
List of Contributors xv
Acknowledgements xvi

Introduction: Comparative Approaches to Data and Methods in Corpus Linguistics 1
JULIA SCHLÜTER AND OLE SCHÜTZLER

Part I Corpus Dimensions and the Viability of Methodological Approaches 15

1 Comparing Standard Reference Corpora and Google Books Ngrams: Strengths, Limitations and Synergies in the Contrastive Study of Variable *h*- in British and American English 17
LUKAS SÖNNING AND JULIA SCHLÜTER

2 Comparing Approaches to Phonological and Orthographic Corpus Formats: Revisiting the Principle of Rhythmic Alternation 46
SABINE ARNDT-LAPPE AND SEBASTIAN HOFFMANN

Part II Selection, Calibration and Preparation of Corpus Data 73

3 Comparing Approaches to (Sub-)Register Variation: The 'Press Editorials' Sections in the British, Canadian and Jamaican Components of ICE 75
FABIAN VETTER

4 Comparing Baselines for Corpus Analysis: Research into the *Get*-Passive in Speech and Writing 101
SEAN WALLIS AND SETH MEHL

© in this web service Cambridge University Press & Assessment www.cambridge.org

CAMBRIDGE

Cambridge University Press & Assessment
978-1-108-49964-4 — Data and Methods in Corpus Linguistics
Edited by Ole Schützler , Julia Schlüter
Frontmatter
More Information

5 Comparing Study Designs and Down-Sampling Strategies in Corpus Analysis: The Importance of Speaker Metadata in the BNCs of 1994 and 2014 127
LUKAS SÖNNING AND MANFRED KRUG

Part III Perspectives on Multifactorial Methods 161

6 Comparing Generalised Linear Mixed-Effects Models, Generalised Linear Mixed-Effects Model Trees and Random Forests: Filled and Unfilled Pauses in Varieties of English 163
TOBIAS BERNAISCH

7 Comparing Logistic Regression, Multinomial Regression, Classification Trees and Random Forests Applied to Ternary Variables: Three-Way Genitive Variation in English 194
MATTHEW FAHY, JESSE EGBERT, BENEDIKT SZMRECSANYI AND DOUGLAS BIBER

8 Comparing Bayesian and Frequentist Models of Language Variation: The Case of *Help* + (*to*-)Infinitive 224
NATALIA LEVSHINA

9 Comparing Methods for the Evaluation of Cluster Structures in Multidimensional Analyses: Concessive Constructions in Varieties of English 259
OLE SCHÜTZLER

Part IV Applications of Classification-Based Approaches 289

10 Comparing Corpus-Driven and Corpus-Based Approaches to Diachronic Variation: Grammatical Changes in Late Modern and Present-Day English 291
GEROLD SCHNEIDER

11 Comparing Annotation Types and *n*-Gram Sizes: German Discourse Particles and Their English Reflexes in a Translation Corpus 323
VOLKER GAST

Index 353

© in this web service Cambridge University Press & Assessment www.cambridge.org

Figures

1.1 The distribution of lexemes in terms of their co-occurrence rate with *a*/*an* in the corpus data (*n* = 150) and the GBN data (*n* = 827). 23
1.2 The estimated share of *a* for *historic* in the corpora and the GBN data set for BrE and AmE. 25
1.3 A sketch of the distribution of the 150 types in COCA. 27
1.4 The share of *a* for each of the 150 types in the COCA data: (a) estimates based on a naïve analysis, ignoring the clustering by text file; (b) estimates from a hierarchical analysis. 28
1.5 The estimated share of *a* for *historic* in COCA and the BNC, broken down by text category. 31
1.6 Spot checks on five further items: Estimated share of *a* in COCA and the BNC, broken down by text category. 32
1.7 The proportion of *a* for different items in the corpora (150 types) and GBN (827 types). 34
1.8 Percentage point difference in the share of *a* for each *h*-lexeme. 35
1.9 Comparison of corpus and GBN estimates for a subset of 116 items. 37
2.1 Overall distribution of stress configurations in the Audio BNC data (*n* = 2,097). 60
2.2 Token frequency distribution of w1 in our dataset, labelled with example types. 60
2.3 Proportions of stress shift and stress clash realisations by genre. 62
2.4 Partial effects of prenominal frequency and text type. 63
3.1 Text types in newspapers. 78
3.2 Left: MDS plot of POS profiles of only the texts originally included in ICE (W2E-001-010). Right: MDS plot of POS profiles, distinguishing between sub-registers and including the newly sampled editorials (W2E-001-015). 89

© in this web service Cambridge University Press & Assessment www.cambridge.org

CAMBRIDGE

Cambridge University Press & Assessment
978-1-108-49964-4 — Data and Methods in Corpus Linguistics
Edited by Ole Schützler , Julia Schlüter
Frontmatter
More Information

3.3 Frequencies of six POS tags with the highest importance measures as obtained from the random forest model. 90
3.4 Left: MDS plot of MDA profiles of only the texts originally included in ICE (W2E-001-010). Right: MDS plot of MDA profiles, distinguishing between sub-registers and including the newly sampled editorials (W2E-001-015). 91
3.5 Frequencies of six MDA tags with the highest importance measures as obtained from the random forest model. 92
3.6 Normalized frequencies of modals per sub-register. 94
4.1 A methodological progression: potential baselines for the *get*-passive, from normalised word frequencies to verified alternation. 105
4.2 Picturing dependent variables and their baselines. 107
4.3 Plotting the rate of *get*- and *be*-passive constructions per main verb, DV1a and b, ICE-GB speech and writing and sub-genres of speech and writing, with 95% Wilson score intervals. 110
4.4 An FTF for retrieving forms of *be* followed by a participial adjective and tagged as copular. 112
4.5 Best estimates of *get*-passives out of *get*- and *be*-passive constructions in speech and writing, using 1990s data from ICE-GB. 117
4.6 Examining sub-genres: rates of selecting the *get*-passive against a baseline of alternating passives (DV2a) and unambiguous acts (DV2b) including dialogues, monologues and mixed situations; printed and non-printed material. 118
4.7 Proportion of *get*- and *be*-passives that have *by*-phrases, out of cases that are unambiguous acts, in ICE-GB. 120
5.1 A sketch of the diachronic development of *actually* from an adverb of manner to an epistemic adverb and a discourse marker: frequency and positional distribution (initial, medial, final) at different stages. 129
5.2 Distribution of speakers in the corpora by age and gender. 131
5.3 Distribution of word counts across speakers. 132
5.4 Distribution of word counts across speakers in the demographic sub-groups. 132
5.5 The CQPweb interface to the Spoken BNC2014. 134
5.6 Variation in the usage rate of *actually* across speakers: the rate of *actually* plotted against the total word count for each speaker. 136
5.7 Estimates for the usage rate of *actually* in different sub-groups, graphed against year of birth. 139

© in this web service Cambridge University Press & Assessment www.cambridge.org

CAMBRIDGE

Cambridge University Press & Assessment
978-1-108-49964-4 — Data and Methods in Corpus Linguistics
Edited by Ole Schützler , Julia Schlüter
Frontmatter
More Information

5.8 Illustration of simple random down-sampling. 143
5.9 Illustration of stratified random down-sampling. 144
5.10 Distribution of token counts for *actually* across speakers in the Spoken BNC1994DS. 145
5.11 Results of our simulation study. 147
5.12 Structured random down-sampling: We (randomly) sample a maximum of three tokens per speaker. 149
5.13 The percentage of non-medial (i.e. peripheral) occurrences of *actually* in the Spoken BNC1994DS, by age and gender. 151
5.14 Efficiency of down-sampling designs: precision of estimates for each demographic sub-group in the Spoken BNC1994DS. 152
5.15 Down-sampling design effects for different schemes. 153
6.1 The effects of the independent variables on PAUSE in monofactorial tests. 169
6.2 Absolute frequencies of pauses per speaker. 171
6.3 Effect plots for interactions with VARIETY. 178
6.4 The glmertree for pause choice. 180
6.5 Out-of-bag errors with different numbers of split candidates. 183
6.6 Variable importance ranking for PAUSE in a random forest model. 185
6.7 Partial dependence plots for unfilled pauses. 186
6.8 Variable importance ranking for PAUSE in variety-specific random forest models. 187
7.1 Possessum length effect plots. 204
7.2 Possessum length effect plot, multinomial regression. 209
7.3 Standard classification tree. 212
7.4 Alternative display of classification tree. 215
8.1 Different types of priors and their effects on posteriors, given the same data. 228
8.2 Effects of big data on posteriors for different types of priors. 229
8.3 Log-odds of the effect of time (per year) on the chances of the *to*-infinitive (versus the bare infinitive). 230
8.4 Partial effects of Year_new and Helper. 239
8.5 Interaction between log-transformed Distance and the presence or absence of *to* before *help* (*horror aequi*) in the maximum likelihood model based on the larger sample. 239
8.6 Interaction between the morphological form of *help* and the presence or absence of the Helpee in the maximum likelihood model based on the larger sample. 240
8.7 The coefficients of the maximum likelihood and Bayesian models. 242

© in this web service Cambridge University Press & Assessment www.cambridge.org

Cambridge University Press & Assessment
978-1-108-49964-4 — Data and Methods in Corpus Linguistics
Edited by Ole Schützler , Julia Schlüter
Frontmatter
More Information

8.8 Coefficients of fixed effects of the large-sample maximum likelihood model, large-sample Bayesian model with weakly informative Cauchy priors, small-sample maximum likelihood model and small-sample Bayesian model with informative priors. 245
8.A1 The weakly informative prior distributions used in this study. 255
8.A2 Effect of priors on the models: regression estimates and their 95% credible intervals. 256
8.A3 Examples of converging and non-converging Markov chains. 258
9.1 Schematic representation of ‘naïve’ versus ‘informed’ clusters. 260
9.2 Illustration of plot types: (a) MDS-plot, (b) NeighborNet and (c) dendrogram. 265
9.3 Schematic differences in *gravity*, *density* and *aloofness*. 271
9.4 Group parameters for anticausal and dialogic concessives predicted by the model. 277
9.5 MDS-plots: single group variables (*n* = 24 units). 279
9.6 NeighborNet: single group variables (*n* = 24 units). 280
9.7 MDS-plot: combined group variables (*n* = 24 units). 282
9.8 NeighborNet: combined group variables (*n* = 24 units). 283
10.1 Increase of the progressive passive and the *get*-passive in COHA. 301
10.2 Increase of progressive forms, DO+verb and an approximation to non-finite subordinate clauses with -*ing* in COHA. 302
10.3 Relative frequency of BE *gone* and HAVE *gone* in the ARCHER corpus. 303
10.4 Frequency of full subordinate clauses, *to*-infinitives and -*ing* clauses, from ARCHER. 304
10.5 Frequency of full subordinate clauses introduced by *that*, *to*-infinitives and -*ing* clauses, from COHA. 305
10.6 (Corrected) frequency of finite subordinate clauses introduced by *that* compared to *to*-infinitives (and -*ing* clauses), from COHA. 309
10.7 Cascade from finite subordinate clause to *to*-infinitive and finally -*ing* form. 310
10.8 Dependency tree for ‘Within this half hour will he be asleep’. 314
11.1 Length of the English sentences. 328
11.2 Dependency parse of (10). 331
11.3 Balanced accuracy scores for the particle task with unigrams: initial training/test-split. 335

© in this web service Cambridge University Press & Assessment www.cambridge.org

Cambridge University Press & Assessment
978-1-108-49964-4 — Data and Methods in Corpus Linguistics
Edited by Ole Schützler , Julia Schlüter
Frontmatter
More Information

11.4 Balanced accuracy scores for the particle task with unigrams: mean values and 95% confidence intervals based on six runs of 5-fold cross-validation for $d \in \{640 \dots 1{,}152\}$. 336
11.5 Most important items for the particle task with lemma_DEPREL unigrams, at $d = 896$ (only unigrams that occur with a frequency of at least 1/1,000 words are shown). 337
11.6 Balanced accuracy scores for the particle task with bigrams: Mean values and 95% confidence intervals based on six runs of 5-fold cross-validation for $d \in \{896 \dots 1{,}536\}$. 338
11.7 Balanced accuracy scores for the *doch*-task with unigrams: mean values and 95% confidence intervals based on six runs of 5-fold cross-validation for $d \in \{512 \dots 1{,}664\}$. 340
11.8 Density plots for variable importance (unigram models for the *doch*-task, $d = 512$). 341
11.9 Balanced accuracy scores for the *doch*-task with bigrams: initial training/test-split. 341
11.10 Balanced accuracy scores for the *doch*-task with trigrams: mean values and 95% confidence intervals based on six runs of 5-fold cross-validation for $d \in \{128 \dots 768\}$. 342
11.11 Highest accuracy scores for all *n*-grams: particle task. 343
11.12 Highest accuracy scores for all *n*-grams: *ja*-task. 344
11.13 Highest accuracy scores for all *n*-grams: *doch*-task. 345

© in this web service Cambridge University Press & Assessment

CAMBRIDGE
Cambridge University Press & Assessment
978-1-108-49964-4 — Data and Methods in Corpus Linguistics
Edited by Ole Schützler , Julia Schlüter
Frontmatter
More Information

Tables

1.1 Overview of the data sources: number of *h*-types and *h*-tokens in the corpora and GBN. 22
1.2 Observed counts for *a/an historic* in the four data sets. 24
1.3 Text categories in the BNC and COCA: word count and proportional share. 29
1.4 Observed counts for *a/an historic* in the BNC and COCA, broken down by text category. 30
1.5 Etymological and phonological sub-groups in the corpus and GBN data: type and token frequencies. 39
1.6 Comparative overview of characteristics of the GBN database and the corpora (BNC and COCA). 40
2.1 Comparison of the different approaches to the study of PRA effects. 65
3.1 Differences in situational characteristics of institutional and personal opinion pieces. 84
3.2 Register classification of press editorials in ICE-GB, ICE-JA and ICE-CAN. 85
3.3 Strengths and weaknesses of the two profile types used in the analysis. 97
4.1 Rescaling estimates to obtain the best estimate of the true rate of *get*-passives out of alternating forms (DV2a). 116
4.2 Estimating the true rate and confidence intervals for *get*-passives out of cases that are unambiguously acts with resultant states and out of cases containing *by*-phrases. 117
4.3 Contingency tables for DV4a (left) and DV4b (right), evaluating the tendency to employ *by*-phrases for *get*- and *be*-passives separately (refined sub-sample for *be*-passives, unambiguous acts). 119
4.4 Comparing the strengths and weaknesses along the methodological progression from normalised word frequencies to verified alternation. 123

© in this web service Cambridge University Press & Assessment www.cambridge.org

CAMBRIDGE

Cambridge University Press & Assessment
978-1-108-49964-4 — Data and Methods in Corpus Linguistics
Edited by Ole Schützler , Julia Schlüter
Frontmatter
More Information

5.1 Corpus size before and after applying our exclusionary criteria. 130
5.2 Concordance lines exported from the Spoken BNC2014. 134
5.3 The rate of *actually* across sociolinguistic sub-groups in the Spoken BNC1994DS: results returned by an interface-based query. 137
5.4 Comparative overview of the advantages and limitations of using speaker metadata at the design and analysis stage of a corpus study. 155
5.A1 Distribution of speakers (after application of exclusionary criteria), overall word counts and *actually* tokens across socio-demographic sub-groups. 159
6.1 The corpus design. 165
6.2 Frequencies of filled and unfilled pauses in ICE-GB, ICE-IND and ICE-SL. 166
6.3 Monofactorial test statistics for the independent variables with PAUSE. 168
6.4 Differences across GLMMs, glmertrees and random forests. 188
7.1 Examples of (non-)interchangeable genitive constructions. 197
7.2 Properties considered in data coding. 198
7.3 Coefficient table from series of logistic regressions. 200
7.4 Predicted probabilities for varying values of possessum length. 203
7.5 Full multinomial regression model coefficients. 206
7.6 Predicted probabilities for varying values of possessum length, multinomial regression. 208
7.7 Random forest importance values. 218
7.8 Comparison of methodologies. 220
8.1 Posterior probabilities of positive effects in the model with the Cauchy priors. 243
8.2 Direct comparison of frequentist and Bayesian modelling. 247
9.1 Constructed example of units, parameters and (Euclidean) distances. 264
9.2 Schematic differences in *gravity*, *density* and *aloofness*. 271
9.3 Units and parameters (percentages of formal variants) in the example study. 277
9.4 Gravity, density and aloofness: single group variables (ranked by gravity within each comparison). 281
9.5 Gravity, density and aloofness: combined group variables (ranked by gravity). 283
9.6 Summaries of visual and distance-based quantification methods. 285

© in this web service Cambridge University Press & Assessment www.cambridge.org

10.1 Strongest POS tag mono-, bi- and trigrams, for early period (top) and for late period (bottom) of the ARCHER corpus. 307
10.2 Most overused verb-group tag sequences, for the early period (top) and late period (bottom) of ARCHER. 312
10.3 Most overused dependency labels for the early period (top) and late period (bottom) of ARCHER. 313
10.4 Most overused label+direction tuples in the early ARCHER period. 315
10.5 Brief summary of pros and cons of data-based and data-driven paradigms. 318
11.1 32 types of *n*-grams under analysis. 327
11.2 English tokens most frequently aligned with *ja* and *doch*. 330
11.3 Summary: Evaluation of *n*-gram types (interpretability). 347

Contributors

SABINE ARNDT-LAPPE Department of English Studies, University of Trier, Germany

TOBIAS BERNAISCH Department of English, Justus Liebig University Giessen, Germany

DOUGLAS BIBER Department of English, Northern Arizona University, Flagstaff, USA

JESSE EGBERT Department of English, Northern Arizona University, Flagstaff, USA

MATTHEW FAHY Department of Mathematics and Statistics, Northern Arizona University, Flagstaff, USA

VOLKER GAST Department of English and American Studies, University of Jena, Germany

SEBASTIAN HOFFMANN Department of English Studies, University of Trier, Germany

MANFRED KRUG Department of English and American Studies, University of Bamberg, Germany

NATALIA LEVSHINA Neurobiology of Language Department, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

SETH MEHL Digital Humanities Institute, University of Sheffield, United Kingdom

JULIA SCHLÜTER Department of English and American Studies, University of Bamberg, Germany

GEROLD SCHNEIDER Department of Computational Linguistics, University of Zurich, Switzerland

OLE SCHÜTZLER Institute of British Studies, Leipzig University, Germany

LUKAS SÖNNING Department of English and American Studies, University of Bamberg, Germany

BENEDIKT SZMRECSANYI Department of Linguistics, University of Leuven, Belgium

FABIAN VETTER Department of English and American Studies, University of Bamberg, Germany

SEAN WALLIS Survey of English Usage, University College London, United Kingdom

© in this web service Cambridge University Press & Assessment www.cambridge.org

Acknowledgements

The editors would like to thank the organizers of the Seventh Biennial International Conference on the Linguistics of Contemporary English (7BICLCE) at the University of Vigo in September 2017, in particular Javier Pérez-Guerra and Elena Seoane, for accepting our conference workshop. This brought together several of our later authors, and it inspired intense discussions and ultimately the design of this book. We would also like to thank our contributors for their patience and engagement with our many comments and multiple rounds of revisions. At Cambridge University Press, we are grateful to Helen Barton for her support in the proposal phase, two anonymous readers, whose comments considerably improved the original structure of the book (although we later had to adapt it somewhat to the turn individual contributions had taken), as well as Isabel Collins, Ruth Boyes and their team, who helped us through the final stages of the publication process. We also thank our student assistants Franziska Schuhmann and Przemysław Żuk, without whom we do not know what we would have done.

© in this web service Cambridge University Press & Assessment www.cambridge.org