

Comparative Variation Analysis

Variation studies is an increasingly popular area in linguistics, becoming embedded in curriculum design, conferences, and research. However, the field is at risk of fragmenting into different research communities with different foci. This pioneering book addresses this by establishing a canon of state-of-the-art quantitative methods to analyze grammatical variation from a comparative perspective. It explains how to use these methods to investigate large datasets in a responsible fashion, providing a blueprint for applying techniques from corpus linguistic, variationist, and dialectometric traditions in novel ways. It specifically explores the scope and limits of syntactic variability in a global language such as English, and investigates three grammatical alternations in nine varieties of English, exploring what we can learn about the grammatical choices that people make based on both observational and experimental data. Comprehensive yet accessible, it will be of interest to academic researchers and students of sociolinguistics, corpus linguistics, and World Englishes.

BENEDIKT SZMRECSANYI is Professor of Linguistics at KU Leuven. His research interests include language variation and its interfaces with typology, geolinguistics, complexity theory, and psycholinguistics.

JASON GRAFMILLER is Assistant Professor of Corpus-Based Sociolinguistics at the University of Birmingham. His research involves the application of various quantitative techniques to examine the forces shaping how language varies across regions, styles, and time.

STUDIES IN LANGUAGE VARIATION AND CHANGE

Series Editor

Sali A. Tagliamonte, University of Toronto

Studies in Language Variation and Change is dedicated to studies of systematic and inherent variation in language, including contemporary or historical sources. It is concerned with the impact of society, geography and culture in so far as they intersect with the structures and processes of language. *Studies in Language Variation and Change* is firmly situated in the variationist sociolinguistic enterprise with its roots in historical linguistics, dialectology, anthropology and importantly in the advancing quantitative methods of the field. The series concentrates on book length syntheses of research that engages with the details of linguistic structure in actual speech production and processing (or writing). It emphasizes replicability of findings, consistent reporting and building critical and substantive explanations out of empirical foundations.

Published so far in the series:

Sociolinguistic Variation in Children's Language: Acquiring Community Norms by Jennifer Smith and Mercedes Durham

Explanations in Sociosyntactic Variation, edited by Tanya Karoli Christensen and Torben Juel Jensen

Meaning, Identity and Interaction: Sociolinguistic Variation and Change in Game-theoretic Pragmatics by Heather Burnett

Forthcoming titles:

Linguistic Variation and Language Change: Synchrony Meets Diachrony by Alexandra D'Arcy

Comparative Variation Analysis

Grammatical Alternations in World Englishes

Benedikt Szmrecsanyi

KU Leuven

Jason Grafmiller

University of Birmingham



CAMBRIDGE
UNIVERSITY PRESS

Cambridge University Press & Assessment
 978-1-108-49156-3 — Comparative Variation Analysis
 Benedikt Szmrecsanyi, Jason Grafmiller
 Frontmatter
[More Information](#)



Shaftesbury Road, Cambridge CB2 8EA, 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 Cambridge University Press & Assessment,
 a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of
 education, learning and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/9781108491563

DOI: 10.1017/9781108863742

© Benedikt Szmrecsanyi and Jason Grafmiller 2023

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 & Assessment.

First published 2023

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

Library of Congress Cataloging-in-Publication Data

Names: Szmrecsanyi, Benedikt, 1976– author. | Grafmiller, Jason, author.

Title: Comparative variation analysis : grammatical alternations in world
 Englishes / Benedikt Szmrecsanyi, Catholic University Leuven ;
 Jason Grafmiller, University of Birmingham.

Description: Cambridge ; New York, NY : Cambridge University Press, 2023. |

Series: Studies in language variation and change | Includes
 bibliographical references and index.

Identifiers: LCCN 2023025075 | ISBN 9781108491563 (hardback) |
 ISBN 9781108798471 (paperback) | ISBN 9781108863742 (ebook)

Subjects: LCSH: English language – Variation. | English language – Syntax. |
 English language – Foreign countries. | English language – Data processing.

Classification: LCC PE1074.7 .S96 2023 | DDC 427–dc23/eng/20230622

LC record available at <https://lccn.loc.gov/2023025075>

ISBN 978-1-108-49156-3 Hardback

Cambridge University Press & Assessment 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.

Cambridge University Press & Assessment
978-1-108-49156-3 — Comparative Variation Analysis
Benedikt Szendrői, Jason Grafmiller
Frontmatter
[More Information](#)

To Angela, Leo, and Alex.

Contents

<i>List of Figures</i>	<i>page</i> ix
<i>List of Tables</i>	xi
<i>Series Editor's Preface</i>	xiii
<i>Acknowledgments</i>	xvi
<i>Data Availability Statement</i>	xvii
1 Introduction	1
1.1 Variationist Sociolinguistics and Corpus-Based Variationist Linguistics	4
1.2 Comparative Linguistics and Comparative Variation Analysis	5
1.3 Dialectology, Dialectometry, and Dialect Typology	6
1.4 Probabilistic Linguistics and Probabilistic Grammar	7
1.5 Psycholinguistics	8
1.6 English as a World Language	9
1.7 Structure	10
2 Grammatical and Syntactic Variation	12
2.1 Does Grammatical Variation Even Exist?	12
2.2 From Traditional Dialectology to Variationist Linguistics	14
2.3 Grammatical Variation in English	15
2.4 Comparative Perspectives on Grammatical Variation in English	19
2.5 Grammatical Alternations Subject to Study in This Book	21
2.6 Summary	32
3 World Englishes and Dialect Typology	34
3.1 Theoretical Models of World Englishes	34
3.2 A View from Dialect Typology	42
	vii

viii	Contents	
	3.3 Varieties in This Study	46
	3.4 Summary	55
4	The Data	56
	4.1 Data Types in Variationist Linguistics	56
	4.2 The Corpora	60
	4.3 Defining the Variable Contexts	65
	4.4 Annotating the Constraints	75
	4.5 Summary	81
5	Alternation-by-Alternation Analysis	82
	5.1 Methods	82
	5.2 A Comparative Analysis of the Genitive Alternation	84
	5.3 A Comparative Analysis of the Dative Alternation	92
	5.4 A Comparative Analysis of the Particle Placement Alternation	98
	5.5 Discussion	106
6	Distances, Similarities, and Coherence	112
	6.1 Background	112
	6.2 VADIS: An introduction	114
	6.3 VADIS: The Empirical Pipeline	115
	6.4 Quantification via Similarity Coefficients	121
	6.5 Evaluating Coherence	123
	6.6 Mapping Out (Dis)similarity Relationships	128
	6.7 Discussion	135
7	Experimental Corroboration	141
	7.1 Methods	143
	7.2 Results	148
	7.3 Discussion	157
8	Where Are We Now, and Where to Next?	166
	8.1 Empirical Summary	167
	8.2 Typological Universals	171
	8.3 Forces Driving Probabilistic Indigenization	173
	8.4 L1 Transfer and L2 Learning	176
	8.5 Implications for Variationist Sociolinguistics	180
	8.6 The Road Ahead	187
	<i>References</i>	191
	<i>Index</i>	217

Figures

3.1	The Three Circles model (inspired by Kachru, 1985)	36
3.2	Multidimensional scaling visualization derived from co-occurrence matrix containing 76 features × 46 varieties (adapted from Szmrecsanyi and Kortmann, 2009, 1651).	45
5.1	Conditional Random Forest permutation variable importance ranking for the pooled genitive dataset.	86
5.2	Conditional Random Forest permutation variable importance ranking of constraints on the genitive alternation by variety of English.	87
5.3	Partial effects plot of the interaction of VARIETY_OF_E with PORANIMACYBIN in the genitive regression model.	91
5.4	Partial effects plot of the interaction of VARIETY_OF_E with PORFINALSIBILANCY in the genitive alternation.	92
5.5	Conditional Random Forest permutation variable importance ranking for the pooled dative dataset.	95
5.6	Conditional Random Forest permutation variable importance ranking of constraints on the dative alternation by variety of English.	96
5.7	Partial effects plot of the interaction of VARIETY_OF_E with RECPRON in the dative alternation model.	99
5.8	Conditional Random Forest permutation variable importance ranking for the pooled particle placement dataset.	101
5.9	Conditional Random Forest permutation variable importance ranking of constraints on the particle placement alternation by variety of English.	102
5.10	Intercept adjustments for the random effect VARIETY_OF_E.	104
5.11	Partial effects plot of the interaction of CIRCLE with DIRECTION-ALPP in the particle placement model.	105
5.12	Partial effects plot of the interaction of CIRCLE with SEMANTICS in the particle placement model.	106
5.13	Partial effects plot of the interaction of CIRCLE with DIROBJLET-TLENGTHBIN in an alternative particle placement model.	107
		ix

x List of Figures

6.1	Variation-Based Distance and Similarity Modeling (VADIS) distance matrix for the third line of evidence in the particle placement alternation (all data included, eight constraints considered).	125
6.2	Multidimensional scaling representation of third line distances for the particle placement alternation (see Figure 6.1).	129
6.3	Multidimensional scaling representation of compromise distances per alternation: a) genitive alternation; b) dative alternation; c) particle placement alternation.	130
6.4	Multidimensional scaling representation of the Γ -matrix (a single compromise distance matrix merged across all lines and alternations).	131
6.5	Dendrogram: clustering the Γ -matrix (a single compromise distance matrix merged across all lines and alternations).	132
6.6	Visualizing aggregate similarities: NeighborNet diagram depicting the Γ -matrix (a single compromise distance matrix merged across all lines and alternations).	134
6.7	Multidimensional scaling representation of the fused distances from all three VADIS lines for seventy-five simulated datasets representing five hypothetical dialect varieties.	138
6.8	Hierarchical clustering of simulated datasets based on VADIS Line 2.	139
7.1	Experimental items versus corpus model predicted probabilities.	146
7.2	Welcome page and instructions for the ratings experiments.	147
7.3	Experimental ratings versus corpus model predicted log odds, with LOESS smooths.	148
7.4	Experimental ratings across varieties, with median rating.	149
7.5	Experimental ratings versus corpus model predictions, averaged by item and variety.	150
7.6	Partial effects plots of interaction of VARIETY on participant ratings.	154
7.7	Partial effects plots of interaction of VARIETY and CORPUS PREDICTION on participant ratings.	155
7.8	Partial effects plots of interaction of VARIETY and DIRECT OBJECT LENGTH on participant ratings.	156
7.9	Predicted probabilities by DIROBJLENGTH (in words), register and VARIETY obtained from random forest model predicting the split variant (adapted from Szmrecsanyi et al., 2016a, figure 3).	162
8.1	Partial effects plot of the interaction of VARIETY_OF_E with RECPRON in the dative alternation model.	178

Tables

3.1	English varieties and their theoretical categorization.	<i>page</i> 55
4.1	Design of the ICE corpora. Values reflect number of texts.	62
4.2	Design of the GloWbE corpus (from Davies and Fuchs 2015, 6).	63
4.3	Summary of genitive variants in ICE and GloWbE corpus data.	75
4.4	Summary of dative variants in ICE and GloWbE corpus data.	75
4.5	Summary of particle placement variants in ICE and GloWbE corpus data.	76
4.6	Annotation schema for the coding of ANIMACY across the three alternations.	77
4.7	Annotation scheme for the coding of NP TYPE across the three alternations.	78
5.1	Variant rates of genitive constructions across varieties of English.	85
5.2	The genitive alternation: fixed effect coefficients in mixed-effects logistic regression analysis.	89
5.3	Variant rates of dative constructions across varieties of English.	93
5.4	The dative alternation: fixed effect coefficients in mixed-effects logistic regression analysis.	97
5.5	Variant rates of particle placement constructions across varieties of English.	100
5.6	The particle placement alternation: fixed effect coefficients in mixed-effects logistic regression analysis.	104
6.1	Predictor sets used for the VADIS analysis.	116
6.2	Model estimates for three fictitious varieties A, B, and C.	118
6.3	Distance matrix for fictitious varieties A, B, and C (Euclidean distance).	118
6.4	Mean distances and mean similarities per variety.	119
6.5	Similarity coefficients across lines of evidence and alternations. Input dataset: all available data.	121
6.6	Core grammar scores (Γ) and hierarchies of stability for subsets of the data.	122

xii	List of Tables	
6.7	DBC _{alternation} measurements: Mantel correlation coefficients between fused distance matrices (combining all lines of evidence and based on all available data).	126
6.8	Mantel correlation coefficients between line-of-evidence-specific distance matrices.	127
7.1	Average ratings (with standard deviations) by item across participants.	151
7.2	Summary statistics, fixed effects coefficients, and random effects standard deviations for linear mixed-model fitting participant rating as function of DIRECT OBJECT LENGTH, CORPUS MODEL PREDICTION, and participant VARIETY, GENDER, and EDUCATION LEVEL.	153
7.3	Agreement scores reflecting the percentage of experimental items in which participants' preferred variant matched the variant predicted by corpus model.	159
7.4	Agreement scores reflecting the percentage to which participants' preferred variant matched the variant predicted by corpus model.	160

Series Editor's Preface

As Language Variation and Change scholarship has expanded beyond variable-by-variable analyses and North America, the scholars of the field have begun to broaden their academic relationships as well. I got to know Benedikt Szmrecsanyi through a collaborative research project funded by the National Science Foundation in the USA:

Bresnan, Joan. 2010–2014. *The Development of Syntactic Alternations*. Research project funded by the National Science Foundation (NSF). BCS-1025602. With collaborators Ford, Marilyn, Hay, Jen, Rosenbach, Anette, Szmrecsanyi, Benedikt, and Tagliamonte, Sali A.

The project brought together multiple data sets with a plan to build parallel corpora in order to conduct a unified analysis of two well-studied syntactic variables, the dative and genitive alternation. What a great plan, right? It was a veritable epic journey, not only in expanding studies of variation and change, but also in setting the stage for combining data sets, advancing statistical modelling, and importantly, establishing best practice for a multiparty collaborative enterprise. While the grandiose aims of the project were to refine the variable contexts in consultation with other members of the team, we spent most of our time in deep discussion about which was the best way to code contexts of variation and analyze them “properly.” Everyone had their own backgrounds, training and predilections and we all had our own vocabulary. Looking back, I wish we had recorded our meetings and discussions. They would have been a gold mine for studying how academic disciplines operate in practice; data, coding, analysis, but oh, so many different ways of doing things and an insightful journey into how different people come to an understanding of each other's point of view. During the same time frame statistical methods were changing quickly as mixed-effects models and new tools for analysis were introduced. This required us to adjust the analytic models and rework the statistical modeling techniques. One of Joan's Ph.D. students at the time was Jason Grafmiller, who was instrumental in conducting the analyses of the big, amalgamated data set that was eventually published in *Glossa*.

Szmrecsanyi, Benedikt, Grafmiller, Jason, Bresnan, Joan, Rosenbach, Anette, Tagliamonte, Sali A. and Todd, Simon. 2017. *Spoken Syntax in a Comparative Perspective: The Dative and Genitive Alternation in Varieties of English. Glossa*. 861–27.

This backstory gives you the history from which developed the authorship team of Benedikt Szmrecsanyi and Jason Grafmiller. They were uniquely poised to take variationist work to a new level, not only across corpora and with cutting-edge methods, but also with established experience in functional collaboration.

In this book, Benedikt and Jason have taken variationist work to the next level, extending the comparative endeavour to a global perspective, analyzing nine varieties of English from two compendia of data, the International Corpus of English (ICE) and the Global Corpus of Web-based English (GloWbE) – parallel, balanced corpora of the standard national varieties of each country, including both spoken and written registers. As Benedikt and Jason started presenting their work at conferences, I was consistently impressed with their methods, techniques and the importance of the findings arising from their studies.

On July 9–10, 2018, some of the original Bresnan-lead team and a few others met for a reunion workshop at Annette Rosenbach's beautiful farm *Tanagra* in McGregor, Western Cape, in South Africa. After yet another superb presentation of results about their research in a talk entitled “Measuring Variable Grammars,” I suggested to Benedikt and Jason that it was time to pull together their joint work in a single monograph to consolidate the new direction of comparative studies and advanced statistical techniques they had pioneered. Among their new techniques they had developed was a Variation Based Distance and Similarity Modelling (VADIS) method which quantifies the similarity between and coherence across datasets as a function of the correspondence in their patterns of choice between competing variants of a variable. In December 2018, Jason came to Toronto to teach me how to use the VADIS method. One cannot have too many tools with which to probe data!

Comparative Variation Analysis: Syntactic Variation in World Englishes explores the stability and plasticity of probabilistic grammar(s) across (standard) varieties of English around the globe in three syntactic alternations: the genitive, dative, and particle placement alternations. In keeping with the mission statement of the series, *Studies in Language Variation and Change*, the book advances the field in several ways. First, it applies rigorous quantitative methods, not only using the standard methods of the field but augmenting the “sociolinguistic toolkit” to include new visualization methods. Second, it brings into the field a new level of cross-variety comparison, extending the comparative sociolinguistic enterprise to incorporate new methods (e.g. VADIS). Third, it advances the techniques brought to bear on cross-linguistic

analysis by using experimental methods. Szmrecsanyi and Grafmiller adhere to quantitative procedures meeting the standards of replicability, consistent reporting, and the embedding of research findings in sociolinguistic theory. Their efforts to develop and employ multiple “lines of evidence,” corpus, dialectal and experimental, to arrive at a fulsome explanation lends greater validity to the eventual explanations they propose.

As the chapters in Benedikt and Jason’s book build from the introduction, to the variables, to the methods, the reader will notice they are entering new and exciting territory for variationist studies. Benedikt and Jason step beyond alternation-by-alternation analyses (Chapter 5), which is itself cutting-edge, into analyses of distance, similarity, and coherence (Chapter 6), where the reader will learn the VADIS method. But that is not all; Chapter 7 demonstrates experimental corroboration of corpus predictions that tap the grammatical knowledge of participants’ preferences. In the final chapter, “Where Are We Now, and Where to Next?” Benedikt and Jason pull together all the evidence and synthesize it insightfully, pointing analysts to “the road ahead.” They propose two explanations for the prevalence of probabilistic grammar universals: 1) shared histories, the “common ancestry” of the varieties under investigation, and 2) shared humanity, “the influence of theoretically universal biases in production and comprehension that influence language structure.”

By the end of the book, Benedikt and Jason have taken readers on an adventure in language variation, deep into the largest corpora in the world, using the most cutting-edge methods and they have sorted out – I would judge – to be among the most complex set of findings ever. Now, there is an apocryphal story, passed down to me from late-night conversations with colleagues and students, that when faced with inscrutable findings, Benedikt’s advice and *modus operandi* is to partake in a glass of wine and deeply ponder the patterns. Whether this is true or not, the fact that Benedikt and Jason have been able to concisely and incisively interpret their complex and multi-faceted results (i.e. see the forest in the trees) is testimony to their scholarship and deep understanding of human language. The intricate design in *World Englishes* has sorted itself into a comprehensive understanding at a new level that will leave readers equipped for advancing their own studies in the future.

Sali A. Tagliamonte

Acknowledgments

First and foremost, we would like to thank Melanie Röthlisberger and Benedikt Stemmler (née Heller) for their extensive and inspired work on the dative and genitive alternations. Melanie and Benedikt were hardworking PhD students in the early stages of the project which this book summarizes, and we cite their work throughout this book. Thank you, Melanie and Benedikt, for the heavy lifting and for your commitment – this project is your project, really. We are likewise extremely grateful to Laura Rosseel, whose comparatively short stint as a project postdoc proved indispensable for getting the rating task experiments on their way, and for codesigning the VADIS method. There are countless other people who contributed directly or indirectly, in one way or another. Here we would like to mention in particular the participants of our 2016 workshop on Probabilistic Variation across Dialects and Varieties (Joan Bresnan, Anette Rosenbach, Marianne Hundt, Sali Tagliamonte, Tobias Bernaisch, Christoph Wolk, Lars Hinrichs, Natalia Levshina, Daniel Ezra Johnson, Magali Paquot, and Dominique Hess [née Bürki]), from whose inspiring feedback we profited enormously. We would also like to express our thanks to Hubert Cuyckens and Dirk Geeraerts (both KU Leuven), who were heavily and passionately involved during the design phase of the project (the first-named author learnt a lot about grant-proposal writing from Hubert and Dirk). On the institutional plane, we are grateful to KU Leuven for hosting the project, to the Quantitative Lexicology and Variational Linguistics (QLVL) research group at KU Leuven's Department of Linguistics for intellectual stimulation, and to the Research Foundation Flanders (FWO, grant no. G.0C59.13N) for generous funding. Finally, our heartfelt thanks go to Matt Hunt Gardner, Thomas Van Hoey, and Yi Li for a thorough reading of some of the chapters. Needless to say, all remaining errors are our own.

Data Availability Statement

The data we report on in this study were compiled from two corpora: the International Corpus of English (<http://ice-corpora.net/ice/index.html>) and the Corpus of Global Web-Based English (<https://www.english-corpora.org/glowbe/>). Restrictions apply to the availability of these corpora, which were used under license for this study. The annotated datasets that were collected from these corpora, along with the dataset of experimental ratings, are freely available on our Open Science Framework (OSF) repository at <https://osf.io/5hvtw/>. All statistical analyses were conducted using R statistical software (<https://www.r-project.org/>), and the corresponding R code for our analyses can also be found in our OSF repository. Tools for applying the Variation-Based Distance and Similarity Method can be found in the `VADIS` R package (<https://github.com/jasongraf1/VADIS>) and accompanying vignettes.