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

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

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

To Angela, Leo, and Alex.

Contents

Figures

Tables

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 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.