Similar Languages, Varieties, and Dialects

Language resources and computational models are becoming increasingly important for the study of language variation. A main challenge of this interdisciplinary field is that linguistics researchers may not be familiar with these helpful computational tools and many natural language processing (NLP) researchers are often not familiar with language variation phenomena. This essential reference introduces researchers to the necessary computational models for processing similar languages, varieties, and dialects. In this book, leading experts tackle the inherent challenges of the field by balancing a thorough discussion of the theoretical background with a meaningful overview of state-of-the-art language technology. The book can be used in a graduate course or as a supplementary text for courses on language variation, dialectology, and sociolinguistics or on computational linguistics and NLP.

Part I covers the linguistic fundamentals of the field, such as the question of status and language variation. Part II discusses data collection and preprocessing methods. Finally, Part III presents NLP applications, such as speech processing, machine translation, and language-specific issues in Arabic and Chinese.

DR. MARCOS ZAMPIERI is an assistant professor at the Rochester Institute of Technology where he leads the Language Technology Group. He obtained his PhD from Saarland University in Germany with a thesis on computational approaches to language variation. He has previously held research and teaching positions in Germany and the UK. Dr. Zampieri published over 80 peer-reviewed papers on topics such as language acquisition and variation, offensive language identification, and machine translation. Since 2014, he is the main organizer of the workshop series on NLP for Similar Languages, Varieties and Dialects (VarDial). He is the lead organizer of the popular OffensEval shared tasks on offensive language identification at SemEval 2019 and 2020.

DR. PRESLAV NAKOV is Principal Scientist at Qatar Computing Research Institute at Hamad Bin Khalifa University. He leads the Tanbih megaproject, developed in collaboration with MIT. He coauthored a book on semantic relations between nominals, two books on computer algorithms, and many research papers in top-tier conferences and journals. He received the Young Researcher Award at RANLP 2011. He was also the first to receive the Bulgarian President’s John Atanasoff Award, named after the inventor of the first automatic electronic digital computer. Dr. Nakov’s research has been featured in more than 100 news outlets, including *Forbes*, the *Boston Globe*, and the *MIT Technology Review*.

Studies in Natural Language Processing

Volumes in the SNLP series provide comprehensive surveys of current research topics and applications in the field of natural language processing (NLP) that shed light on language technology, language cognition, language and society, and linguistics. The increased availability of language corpora and digital media, as well as advances in computer technology and data sciences, has led to important new findings in the field. Widespread applications include voice-activated interfaces, translation, search engine optimization, and affective computing. NLP also has applications in areas such as knowledge engineering, language learning, digital humanities, corpus linguistics, and textual analysis. These volumes will be of interest to researchers and graduate students working in NLP and other fields related to the processing of language and knowledge.

Also in the Series

Douglas E. Appelt, *Planning English Sentences*
Madeleine Bates and Ralph M Weischedel (eds.), *Challenges in Natural Language Processing*
Steven Bird, *Computational Phonology*
Peter Bosch and Rob van der Sandt, *Focus*
Pierette Bouillon and Federica Busa (eds.), *Inheritance, Defaults and the Lexicon*
Ronald Cole, Joseph Mariani, Hans Uszkoreit, Giovanni Varile, Annie Zaenen, Antonio Zampolli, and Victor Zue (eds.), *Survey of the State of the Art in Human Language Technology*
David R. Dowty, Lauri Karttunen, and Arnold M Zwicky (eds.), *Natural Language Parsing*
Ralph Grishman, *Computational Linguistics*

Graeme Hirst, *Semantic Interpretation and the Resolution of Ambiguity*
András Kornai, *Extended Finite State Models of Language*
Kathleen R. McKeown, *Text Generation*
Martha Stone Palmer, *Semantic Processing for Finite Domains*
Terry Patten, *Systemic Text Generation as Problem Solving*
Ehud Reiter and Robert Dale, *Building Natural Language Generation Systems*
Manny Rayner, David Carter, Pierette Bouillon, Vassilis Digalakis, and Matis Wiren (eds.), *The Spoken Language Translator*
Michael Rosner and Roderick Johnson (eds.), *Computational Lexical Semantics*
Richard Sproat, *A Computational Theory of Writing Systems*
George Anton Kiraz, *Computational Nonlinear Morphology*
Nicholas Asher and Alex Lascarides, *Logics of Conversation*
Margaret Masterman (edited by Yorick Wilks), *Language, Cohesion and Form*
Walter Daelemans and Antal van den Bosch, *Memory-Based Language Processing*
Chu-Ren Huang, Nicoletta Calzolari, Aldo Gangemi, Alessandro Lenci, Alessandro Oltramari, and Laurent Prévot (eds.), *Ontology and the Lexicon: A Natural Language Processing Perspective*
Thierry Poibeau and Aline Villavicencio (eds.), *Language, Cognition, and Computational Models*

Similar Languages, Varieties, and Dialects

A Computational Perspective

Edited by

Marcos Zampieri
Rochester Institute of Technology

Preslav Nakov
Qatar Computing Research Institute, HBKU

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/9781108429351
DOI: 10.1017/9781108565080

First published 2021

Printed in the United Kingdom by TJ Books Ltd, Padstow Cornwall

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

Library of Congress Cataloging-in-Publication Data
Names: Zampieri, Marcos, editor. | Nakov, Preslav, editor.
Title: Similar languages, varieties, and dialects : a computational perspective /
Marcos Zampieri, Preslav Nakov.
Description: 1. | New York : Cambridge University Press, 2021. |
Series: Studies in natural language processing |
Includes bibliographical references and index.
Identifiers: LCCN 2020014771 (print) | LCCN 2020014772 (ebook) |
ISBN 9781108429351 (hardback) | ISBN 9781108565080 (epub)
Subjects: LCSH: Language and languages–Variation. | Dialectology. |
Sociolinguistics. | Computational linguistics.
Classification: LCC P120.V37 S556 2020 (print) | LCC P120.V37 (ebook) |
DDC 417–dc23
LC record available at https://lccn.loc.gov/2020014771
LC ebook record available at https://lccn.loc.gov/2020014772

ISBN 978-1-108-42935-1 Hardback

Contents

Part I Fundamentals

Part II Methods and Resources

Cambridge University Press
978-1-108-42935-1 — Similar Languages, Varieties, and Dialects
Edited by Marcos Zampieri, Preslav Nakov
Frontmatter
More Information

Contributors

Željko Agić Unity Technologies

Bengt J. Borgström MIT Lincoln Laboratory

Charlotte Gooskens University of Groningen

Nizar Habash New York University Abu Dhabi

Wilbert Heeringa Fryske Akademy

Chu-Ren Huang The Hong Kong Polytechnic University

Menghan Jiang Peking University/The Hong Kong Polytechnic University

Steffen Klaere University of Auckland

Jingxia Lin Nanyang Technological University Singapore

Nikola Ljubešić Josef Stefan Institute

Miriam Meyerhoff Victoria University of Wellington

Preslav Nakov Qatar Computing Research Institute, HBKU

John Nerbonne University of Freiburg

Dong Nguyen Alan Turing Institute

Jelena Prokić Leiden University

Tanja Samardžić University of Zurich

Yves Scherrer University of Helsinki

Dingxu Shi Guangdong University of Foreign Studies

Rachael Tatman Rasa Technologies

Jörg Tiedemann University of Helsinki

Pedro A. Torres-Carrasquillo MIT Lincoln Laboratory

VINCENT J. VAN HEUVEN University of Pannonia, Leiden University

JAMES A. WALKER La Trobe University

MARTIJN WIELING University of Groningen

HONGZHI XU Shanghai International Studies University

MARCOS ZAMPIERI Rochester Institute of Technology

Foreword

Over the past decade, we have witnessed a revolution in Natural Language Processing (NLP) driven by advances in machine learning, growing amounts of natural language data, and increased computing power. As a result, computational linguists have made major advances across a range of tasks with diverse technological applications. Crucially, these advances are changing the lives of people around the world, who are increasingly interacting with NLP as part of their daily routine. These innovations are also starting to revolutionise the field of linguistics, where very large corpora and NLP methodologies are being used to expand our scientific understanding of language, allowing linguists to study language at scale for the first time, as demanded by the inherent complexity of this most basic of human behaviours. However, although these technological advances are improving our world in many ways, they are also creating new problems.

Perhaps most obvious of these problems is that NLP technologies tend to work best for speakers of the dominant languages of the world – thereby inadvertently increasing the power of speakers of these languages. Modern NLP workflows generally rely on large collections of training data, often carefully annotated by hand, which are usually only available for major languages. As a result, the benefits of the AI revolution in NLP have largely been reserved for speakers of a small number of high-resourced languages, especially English. To address this issue, there have been calls for NLP researchers to clearly acknowledge the languages upon which they work – especially if that language is simply English, as is often the case – and by extension to accept that NLP applications that are trained and evaluated on high-resourced languages will not necessarily perform as well on languages of the world more generally. This is popularly known as the ‘Bender Rule’, named after Emily Bender, the computational linguist who has championed this perspective over the past decade. The rule has even been evoked in linguistics, primarily to encourage linguists who work on the English language to acknowledge how this naturally limits the generalisability of their research.

But speakers of low-resourced languages are not the only ones who are being left behind by the current NLP revolution. In addition to zooming out to acknowledge the typological diversity of human language, computational linguists must zoom in and acknowledge the sociolinguistic diversity of human languages. Even English is not simply English – it is a complex system of mutually and sometimes not-so-mutually intelligible dialects, characterised by systematic patterns of variation, changing over time in ways and for reasons that linguists are only beginning to understand. Treating this variation as noise or assuming that speakers of non-standard dialects will be adequately served by tools trained on standard dialect data amounts to contributing to the entrenchment of social inequities.

Speakers of non-standard dialects are being technologically disenfranchised by NLP. For speakers standard American English or Standard British English, NLP tools for English work well and will inevitably continue to improve over time, but performance necessarily falls for speakers of non-standard dialects, especially those associated with the most marginalised communities. As NLP tools increasingly become embedded into our daily lives, it is crucial that these technologies work for regular people regardless of their social background. Computational linguists have a responsibility to make sure this is the case – to fight against the inequalities they have helped create. The challenge is to build robust NLP technologies that model linguistic variation directly, probably at the expense of maximising economic value.

This situation presents a clear intellectual challenge for computational linguists. It is arguably as complex as any problem currently faced in NLP. It is certainly a far greater technical problem than the divide between high- and low-resourced languages, where the solution at least is clear – the compilation and annotation of large corpora for low-resourced languages. Alternatively, there is no clear approach to building NLP tools that can handle language internal variation, not least because linguists do not have anywhere near a complete understanding of how language internal variation is patterned. For example, building corpora that represent languages, regardless of their size, is often fairly straightforward, as languages are often delimited by relatively hard borders – be they defined politically or based on mutual intelligibility – but dialects generally blend into each other, making even compiling representative dialect corpora a difficult task, and at least part of the solution.

This volume represents an important first step to address the challenge of language-internal variation in NLP. By bringing together linguists who specialise in language variation and change and computational linguists who develop NLP tools capable of working with language-internal variation, this collection provides an indispensable introduction for any researcher in NLP who takes this challenge seriously. This volume also meets the Bender Rule by considering language-internal variation cross-linguistically, highlighting the fact that all language, be they low- or high-resourced, contain internal variation.

Crucially, this volume also opens up a dialogue between NLP researchers and linguists interested in variation. It is sometimes remarked that linguists do not have much to offer NLP these days, at least for high-resourced languages like English, given that the grammars of these languages are fairly well understood. Such an attitude reflects a narrow view of language and linguistics that artificially simplifies the endeavour of NLP, in much the same way that generative linguists once artificially simplified the endeavour of linguistics. When the NLP community takes variation seriously, sociolinguists, dialectologists, historical linguists, and corpus linguists should be the first port of call. Similarly, given the complexity of linguistic variation, these are probably the same linguists that have the most to gain from engaging with computational linguistics, as demonstrated by the emerging interdisciplinary field of computational sociolinguistics, with which a number of the contributors to this volume are affiliated. Ideally linguists and computational linguists will work together to better understand and process language variation, with research in one area informing the other, accelerating progress in both.

Jack Grieve
University of Birmingham
February 18, 2020

Cambridge University Press
978-1-108-42935-1 — Similar Languages, Varieties, and Dialects
Edited by Marcos Zampieri, Preslav Nakov
Frontmatter
More Information

Introduction

Variation is intrinsic to human language. Language differs from speaker to speaker, from community to community, as well as across time, genre, media, etc. Natural language processing (NLP) systems are typically trained on standard contemporary language varieties such as the language found in books and newspapers. Such systems work very well on the kind of language they are trained on, but their performance degrades when faced with variation.

One of the most relevant dimensions of variation from a computational perspective is diatopic language variation, or the variation of language spoken (and written) in different places and/or regions of a linguistic area, e.g., language varieties and dialects. Dialects are per se nonstandard, thus posing challenges to most off-the-shelf NLP tools. On the other hand, the similarity between closely related languages such as Dutch–Flemish, Bulgarian–Macedonian, and Turkish–Kazakh can provide opportunities for researchers and developers.

With these challenges and opportunities in mind, we introduce you to the present book, *Similar Languages, Varieties, and Dialects: A Computational Perspective*. The book consists of fourteen chapters written by well-known researchers in dialectology, language variation, sociolinguistics, computational linguistics, and natural language processing.

The idea for this book came from the success of the series of workshops – Natural Language Processing for Similar Languages, Varieties and Dialects (VarDial) – that have been organized yearly since 2014 and have been co-located with international NLP conferences such as COLING, EACL, NAACL, and RANLP. VarDial has become an important forum for scholars working on topics related to the study of diatopic language variation from a computational perspective and the application of NLP methods to dialects and similar languages. Since the workshop's first edition, there has been an uphill trend in the number of submissions as well as in the number of research papers published on related topics in specialized journals and conferences.

Even though the interest of the research community has seen steady growth, to the best of our knowledge, so far there have been no books approaching diatopic language variation from a computational perspective. While there have been several well-known handbooks and edited volumes published on topics such as dialectology (Chambers and Trudgill, 1998); language variation, and change (Chambers et al., 2002); and sociolinguistics (Meyerhoff, 2015), the computational aspect has remained largely underexplored.

We believe that this book fills an important gap in the existing literature. It is interdisciplinary by nature, and it can be useful for both experienced researchers and (graduate) students in computer science, linguistics, natural language processing, and related areas. The book provides a concise introduction to core topics in language variation and an overview of the computational methods applied to similar languages, varieties, and dialects.

The book is divided into three parts. Part I covers the fundamentals of language variation and the study of dialects and similar languages. Chapter 1 discusses different dimensions of language variation and how they are manifested. Chapter 2 focuses on the phonetic variation in dialects. The question of status, i.e., dialect versus language, is discussed in Chapter 3. Mutual intelligibility between similar languages and dialects is discussed in detail in Chapter 4, with several examples from languages such as Danish, Spanish, and Portuguese. Closing the first part of the book, Chapter 5 presents a concise yet comprehensive overview of dialectology for computational linguists.

Part II covers methods and resources for data collection, preprocessing, and annotation for similar languages, varieties, and dialects. Chapter 6 deals with data collection and representation, covering social media and speech transcripts. Chapter 7 discusses preprocessing and adaptation of taggers used to annotate similar languages. Finally, Chapter 8 deals with methods to learn dependency parse trees from one language and to project them to a related language.

The last part of the book, Part III, covers applications and language-specific issues when processing similar languages, varieties, and dialects. Chapter 9 presents an overview of computational methods for similar languages and dialect identification. Chapter 10 presents an account of computational methods applied to diatopic language variation in social media. Chapter 11 deals with machine translation between similar languages, varieties, and dialects. Chapter 12 discusses speech processing applications. The last two chapters deal with language-specific issues when processing dialects and varieties of two major languages: Arabic (Chapter 13) and Chinese (Chapter 14).

We would like to take this opportunity to thank all chapter authors for their valuable contribution and the colleagues who kindly helped us by giving the authors feedback and suggestions. We are grateful to Željko Agić, Patricia Cukor-Avila, Charlotte Gooskens, Wilbert Heeringa, Vincent J. van Heuven, Chu-Ren Huang, Menghan Jiang, Steffen Klaere, Jingxia Lin, Nikola Ljubešić, Miriam Meyerhoff, John Nerbonne, Dong Nguyen, Jelena Prokić, Tanja Samardžic, Dingxu Shi, Rachael Tatman, Jörg Tiedemann, Pedro Torres-Carrasquillo, James A. Walker, Martijn Wieling, and Hongzhi Xu.

Finally, we would like to thank the series editor, Chu-Ren Huang for the interest in our volume and for the continuous support, and Kaitlin Leach and Amy He at Cambridge University Press for their support throughout the editorial process.

References

Chambers, J. K. and Trudgill, P. (1998). *Dialectology.* Cambridge: Cambridge University Press.

Chambers, J. K., Trudgill, and P. Schilling-Estes, N. eds. (2002). *The Handbook of Language Variation and Change.* Oxford: Blackwell.

Meyerhoff, M. (2015). *Introducing Sociolinguistics.* New York: Routledge.