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

https://doi.org/10.1017/9781108565080.001 Published online by Cambridge University Press

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.

https://doi.org/10.1017/9781108565080.001 Published online by Cambridge University Press

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.

https://doi.org/10.1017/9781108565080.001 Published online by Cambridge University Press