Contributions to the Computational Processing of Diachronic Linguistic Corpora

Computer-assisted corpus linguistics is one of the main points of convergence between linguistic and computational methods. In particular, the use of diachronic linguistic corpora provides opportunities for the quantitative analysis of phenomena concerning language change through time. This dissertation offers contributions to three of the stages of the research involving diachronic corpora: (a) corpus building and compilation; (b) designing of tools and algorithms for data exploration; and (c) data analysis for linguistic, cultural and historical research. Two resources are first presented: a Web scraper of comments from news portals; and a diachronic corpus composed of comments published in a major Brazilian news website. These resources are relevant not only for linguists, but also for professionals concerned with the public perception of news and the relationship between media and society. Then, we propose a generalizable method to assist the identification of periods of establishment and obsolescence of linguistic items in a diachronic corpus based on the frequency of these items in the corpus. This method may be employed for the analysis of any collection of linguistic items, regardless of language or historical period. Finally, we describe how diachronic corpora might be used for quantitative linguistic investigation by proposing a framework centered on the investigation of vocabulary through a diachronic approach. The applicability of this framework is demonstrated through the case analysis of the use of the term *fake news* in the media. With these contributions, we expect to advance research on diachronic corpus linguistics and on computational methods for linguistic analysis.

ISBN 978-94-6093-343-1