BOOK NOTICE

BOOK NOTICE of PAWŁOWSKI, A., MAČUTEK, J., EMBLETON, S., & MIKROS, G. (EDS.), 2021. *LANGUAGE AND TEXT: DATA, MODELS, INFORMATION AND APPLICATIONS*. JOHN BENJAMINS PUBLISHING COMPANY. ISBN 978 90 272 1010 4 (Hb)

by SIDNEY WONG, UNIVERSITY OF CANTERBURY, NEW ZEALAND
SIDNEY.WONG@PG.CANTERBURY.AC.NZ

Te Reo – Journal of the Linguistic Society of New Zealand
Volume 66, Issue 1
Pages 15–16

Language and Text is a product of the 10th QUALICO conference on quantitative linguistics hosted in Wrocław, Poland in 2018. The conference explored the impact on text analysis of the major increases in data repositories, the rise of social media language data and the decline of research on historical and fictional texts. The volume represents the evolution of thinking in text analysis from analytical methods suitable for small text samples to statistical and machine learning methods suitable for very large corpora of written language.Chapters are grouped into two parts.

Part I explores theory and models, and tests established corpus linguistic methods (collocations, co-occurrence, frequency distributions, n-grams etc.) for their "goodness of fit" for the analysis of sample empirical data, including the distribution of noun phrases in written Japanese, syllable properties in Croatian, Serbian, Russian, and Ukrainian, and the position of enclitics in Old Czech.

Part II is introduced by a chapter on the perils of using 'big data' and, through presentation of a cautionary tale, makes three key recommendations. Firstly, it is suggested that researchers should filter large text data with care to ensure that they do not obscure patterns. Secondly, researchers need to be clear with their assumptions and the limitations of text data; and, thirdly, they need to have an in-depth understanding of the subject matter of their data source. The volume then provides a series of empirical studies which explore relationships between source texts and linguistic behaviour and extends more established methods to include those involving machine learning. The various contributions also extend the application of text analysis beyond discrete linguistic structures to include components from adjacent fields such as critical discourse analysis and digital humanities. For example, using a supervised and unsupervised n-gram algorithm, one study provides evidence that it is possible to identify the genre of a book from its text; and the author's gender from the titles in their bibliography. Other studies highlight practical applications of text analysis, including in the political sphere and in a language learning context.

Language and Text provides a valuable summary of text analytic methods applied to concrete examples. However, prior knowledge of linguistic theory and statistical methods

will aid the reader in navigating the wide range of studies in this rapidly changing field. As mentioned by the editors in their introduction, “the methods of quantitative text analysis and artificial intelligence develop so dynamically precisely because they allow language users to better organize mass information processes and restore a sense of order at the cognitive level.” (p. 2). Linguists’ use of computational methods can bypass the limitations of methods in linguistic research which rely on limited language samples and small multivariate datasets. Care needs to be taken, however, to ensure that ethical questions around data quality and bias are addressed; for example binary notions of gender, or limited geo-political sampling.

The studies reported in *Language and Text* reflect a diverse range of languages, language varieties and registers from across the world, in a range of different writing scripts (Latin, Cyrillic, Arabic etc.), and include morphological, phrase-level and an indirect phonological analysis. However, there seems to be an underlying assumption in some of the chapters that the behaviour observed in written language is easily transferable to other modes of language.