Natural Language Engineering (2023), **29**, pp. 842–845
doi:10.1017/S1351324922000201

CAMBRIDGE UNIVERSITY PRESS

Book Review

Natural Language Processing for Corpus Linguistics by Jonathan Dunn. Cambridge: Cambridge University Press, 2022. ISBN 9781009070447 (PB), ISBN 9781009070447 (OC), vi+88 pages.

Corpus linguistics is essentially the computer-based empirical analysis that examines naturally occurring language and its use with a representative collection of machine-readable texts (Sinclair, 1991; Biber, Conrad and Reppen, 1998; McEnery and Hardie, 2012). The techniques of corpus linguistics enable the analyzing of large amounts of corpus data from both qualitative (e.g., concordances) and quantitative (e.g., word frequencies) perspectives, which in turn may yield evidence for or against the proposed linguistic statements or assumptions (Reppen, 2010). Despite its success in a wide range of fields (Römer, 2022), traditional corpus linguistics has become seemingly disconnected from recent technological advances in artificial intelligence as the computing power and corpus data available for linguistic analysis continue to grow in the past decades. In this connection, more sophisticated methods are needed to update and expand the arsenal for corpus linguistics research.

As its name suggests, this monograph focuses exclusively on utilizing NLP techniques to uncover different aspects of language use through the lens of corpus linguistics. It consists of four main chapters plus a brief conclusion. Each of the four main chapters highlights a different aspect of computational methodologies for corpus linguistic research, followed by a discussion on the potential ethical issues that are pertinent to the application. Five corpus-based case studies are presented to demonstrate how and why a particular computational method is used for linguistic analysis. Given the methodological orientation of the book, it is not surprising that there are substantial technical details concerning the implementation of these methods, which is usually a daunting task for those readers without any background knowledge in computer programming. Fortunately, the author has made all the Python scripts and corpus data used in the case studies publicly available online at https://doi.org/10.24433/CO.3402613.v1. These online supporting materials are an invaluable complement to the book because they not only ease readers from coding but also make every result and graph in the book readily reproducible. To provide better hands-on experience for readers, a quick walkthrough on the accessing of online materials is presented prior to the beginning of the main chapters. With just a few clicks, readers will be able to run the code and replicate the case studies with interactive code notebooks. Of course, readers who are familiar with Python programming are encouraged to further explore the corpus data and expand the scripts to serve their own research purposes.

Chapter 1 provides a general overview of the computational analysis in corpus linguistics research and outlines the key issues to be addressed. It first defines the major problems (namely, categorization and comparison) in corpus analysis that NLP models can solve, and explains why computational linguistic analysis is needed for corpus linguistic research (namely., reproducibility and scalability). The author then introduces all five case studies to be presented in the forthcoming chapters. These studies, ranging from usage-based grammar to corpus-based sociolinguistics, demonstrate how NLP methods can be applied to investigate real-world linguistic phenomena. As for the key issues, the categorization problems and comparison problems are discussed in two

separate sections. On the one hand, a categorization problem can be considered as a text classification task in which a classifier is trained to predict the category that a text belongs to. The training of a text classifier requires a human annotator to define the categories beforehand. On the other hand, a comparison problem is concerned with measuring the distance (or similarity) between words or documents by using a text similarity model, and then clustering the words or documents into groups based on their similarity. Unlike a text classifier, a text similarity model finds categories on its own and does not need human annotation. But what makes language data understandable to computers? The answer is vector space, the numeric representations for language "in which the relationship between vectors mirrors the linguistic relationships that we are interested in" (p. 9). At the end of this chapter, the author raises the ethical issue of data rights. Readers are reminded to be cautious on the representativeness and ownership of the corpus data used, because the linguistic knowledge learnt by a model depends on the data provided.

Chapter 2 concentrates on how different types of linguistic signals are used as features for text classification. The author starts with the issue of classifier evaluation, using an example where a classifier is trained (using construction frequencies as features) to distinguish different dialects of English used online. Here the classifier is evaluated by three commonly used metrics of prediction quality: precision, recall, and f-score. The author demonstrates how these measures can be implemented and how the results of evaluation can be interpreted accordingly. In each of the following four sections, he focuses on a different vector space represented by a specific aspect of language such as content/topic, grammatical structure, syntactic context, and pragmatic sentiment. In the "representing content" section, the objective of classification is to predict the geolocation that a tweet was sent from using the content (i.e., lexical words, multi-word expressions) of that tweet. The author illustrates how a document is transformed into a vector space that is composed of its own semantic content. In the "representing structure" section, function word n-grams are used as stylistic features to determine the authorship of over 1,100 books published in the nineteenth century. In the "presenting context" section, a positional vector represented by one-hot encoding is created as the syntactic context for a given word to predict its part of speech. In the "presenting sentiment" section, a sentiment lexicon is employed to convert hotel review texts into vectors, based on which a sentiment classifier is trained to differentiate between good and bad hotel reviews. Regardless of the type of linguistic features chosen for each case study, it is shown that classification models trained on these features can yield quite good prediction results in general.

The author elaborates further with two separate sections on the mechanisms underlying two types of classification models, that is, logistic regression, and feed-forward networks. Although both models can be used for text classification, they are largely different. Logistic regression is "shallow" and straightforward and allows users to check which feature counts, hence the result is more interpretable. On the contrary, feed-forward network is "deep" and complicated partly because it embeds multiple hidden layers. These hidden layers are not inspectable, which has made its result less interpretable. Moreover, a logistic regression usually works better on small datasets, while the training of a feed-forward network may require a large amount of data. Finally, the problem of implicit bias (i.e., when a model learns unrelated cues from a dataset and makes incorrect generalizations) is discussed in the ethic section.

Chapter 3 aims to address comparison problems using text similarity models. It starts with a short analysis of the differences between categorization and comparison problems. A categorization problem deals with predefined discrete categories, while a comparison problem works on the scalar relationship between samples. The following four sections investigate measuring pairwise similarities at corpus, document, and word levels. Corpus similarity informs us to what extent two corpora or datasets resemble each other. A case study on register across 12 languages is presented to demonstrate how corpus similarity measures (e.g., Spearman correlation) are utilized to compare corpora and explore register variation. It is shown that context-specific features such as word frequencies and character n-gram frequencies are robust indicators of register variation. Document similarity identifies the most similar texts based on linguistic signals that represent

https://doi.org/10.1017/S1351324922000201 Published online by Cambridge University Press

content, author (or style), or sentiment of the target texts. The similarity of two documents is measured as the Euclidean distance between vectors of the corresponding texts. Word similarity tells us how close two words are in terms of meaning. The measurement of word similarity is discussed in two consecutive sections, with the first focusing on association (e.g., ΔP association, or word association matrix) and the second focusing on distributional semantics (e.g., cosine distance in word embeddings). The next section goes beyond comparing pairwise similarities, and discusses the clustering of related words using k-means algorithm. In the remainder of this chapter, the ethical issue of model discrimination (i.e., the negative stereotype) is discussed.

Chapter 4 focuses on the validation and visualization of the results of computational linguistic analysis. The evaluation of computational models can sometimes be tricky, since one cannot know whether a classifier performs well according to the f-score alone without considering its context. To illustrate this point, the author offers an example evaluating the classifiers trained for political speech prediction, demonstrating how different classifiers can be compared and how the results validated against the baseline. Another issue in model validation is overfitting, which usually occurs when a model is trained on smaller datasets. In such cases, techniques such as cross-validation (for "shallow classifiers" such as logistic regression) and validation sets (for "deep classifiers" such as feed-forward networks) can be employed. The author then discusses how different types of plots and graphs can be used to visualize the results of such analyses, thereby facilitating their understanding. For example, line plots are used to present the unmasking performance for authorship attribution; scatterplots and heat maps are created to compare word embeddings; while choropleth maps are used to show linguistic diversities across the globe. The author ends this chapter with a discussion on the ethical issue of imbalanced data availability (i.e., the lack of equal access of language data).

Chapter 5 draws a brief conclusion on the main points covered in the book. The author summarizes the issues that have been addressed in categorization and comparison problems. However, one question remains: is there always a clear distinction between these two families of methods? He then discusses possible applications in which interactions of these two may take place. It is noteworthy that most of the analyses presented in this book are based on the *text_analytics* python package that is dedicated to corpus-based text analysis using NLP techniques. Readers are encouraged to explore further this and other tools to answer questions of their corpus linguistic studies.

To sum up, this book would be insightful to students and researchers who are interested in text-based linguistic analyses, by bridging the gap between corpus linguistics and natural language processing. It provides a useful reference for corpus linguists who are eager to know how recent NLP techniques can be employed to address linguistic questions, and for computer scientists who are curious about the linguistic assumptions and limitations behind the techniques in NLP for corpus linguistics. Therefore, it is a valuable addition to the essential reading list of corpus linguistics research and natural language engineering.

Research funding. The research work on this review is supported by Sichuan Foreign Languages & Literature Research Center under Grant No. SCWYH18-06, and the Social Science Fund of Sichuan Province under Grant No. SC21WY002.

Ju Wen[1]* and Lan Yi[2]
[1]School of Liberal Education, Chengdu Jincheng College,
Sichuan 611731,
P. R. China and
[2]School of Foreign Languages, Chengdu Jincheng College,
Sichuan 611731,
P. R. China
*Corresponding author. E-mail: jupiter@cqu.edu.cn

https://doi.org/10.1017/S1351324922000201 Published online by Cambridge University Press

References

Biber D., **Conrad S.** and **Reppen R.** (1998). *Corpus Linguistics: Investigating Language Structure and Use*. New York: Cambridge University Press.

McEnery T. and **Hardie A.** (2012). *Corpus Linguistics: Method, Theory and Practice. Cambridge Textbooks in Linguistics*. Cambridge: Cambridge University Press.

Reppen R. (2010). Building a corpus: What are the key considerations? In *The Routledge Handbook of Corpus Linguistics*. London: Routledge. https://doi.org/10.4324/9780203856949.ch3

Römer U. (2022). Applied corpus linguistics for language acquisition, pedagogy, and beyond. *Language Teaching* **55**(2), 233–244. https://doi.org/10.1017/S0261444821000392

Sinclair J. (1991). *Corpus, Concordance, Collocation: Describing English Language*. Oxford: Oxford University Press.

https://doi.org/10.1017/S1351324922000201 Published online by Cambridge University Press