Dunn, J. (2022). *Natural Language Processing for Corpus Linguistics*. Cambridge University Press. 84 pp.

Reviewed by Hanna Schmück (Lancaster University)

Natural Language Processing for Corpus Linguistics (Dunn, 2022) is part of the Cambridge *Elements in Corpus Linguistics* series and thus aimed at both experts looking to expand their toolkit and students with an interest in computational linguistics. The book exhibits a strong practical focus which is reflected in the 20 interactive labs containing *Python* code specifically written and commented with replicability, adaptability, and pedagogic use in mind. It can be seen as a handbook for first advances into more computationally complex Natural Language Processing (NLP) assuming a corpus linguistic background as well as a resource for computationally savvy researchers looking to utilise part of the code for their own research projects. Perhaps due to the foundation of the interactive labs stemming from various previous research projects, the book heavily relies on and references Dunn's own previous work. The book is divided up into five major chapters: an introductory part discussing Computational Linguistic Analysis more generally, followed by more technical chapters on text classification, text similarity, and validation and visualisation. Dunn concludes very briefly in a last chapter and provides pointers to further resources. As part of each chapter the author puts particular emphasis on the necessity of invoking ethical considerations in each use case. This review briefly summarises the content covered in each chapter before providing a critique of the discussed content.

The first chapter titled "Computational Linguistic Analysis" explores how the incorporation of NLP approaches can aid both reproducibility and scalability of linguistic studies. A differentiation is made between two major types of research areas:

i. 'Categorization Problems' which are attempted to be solved via text classifiers. Here, the approach entails assigning labels from a predefined set to language elements. Examples include part of speech tagging, document-level topic/genre identification, or dialect classification.
ii. 'Comparison Problems' which are attempted to be solved via similarity models. Examples for comparison problems are determining whether two words share the same semantic category, sentiment comparisons between documents, or authorship comparisons. Text similarity models differ from text

https://doi.org/10.1075/ijcl.00057.sch | Published online: 22 December 2023
International Journal of Corpus Linguistics 29:1 (2024), pp. 123–129. ISSN 1384-6655 | E-ISSN 1569-9811

> classifiers in that they do not rely on predefined labels but require more dynamic approaches to measuring similarities between linguistic elements.

Dunn continues to characterise the case studies explored in greater detail in the later chapters briefly before providing the reader with a fundamental understanding of how vector space representations are used in NLP for language analysis. This section describes how words and sentences are represented numerically via frequency vectors; these are based on the frequency distribution of individual words in sentences or documents. The novelty this provides to the traditional corpus methods is that the vectors enable the use of a variety of mathematical operations on the words which now exist in vector-shape. The section also mentions challenges in handling rare word types and small vocabulary sizes. Finally, the chapter is concluded with a discussion of data rights, particularly highlighting that the increase in model size and computational power results in an increased need for thorough ethical reviews.

Moving closer to the heart of the book, the focus of Chapter two lies on text classification and model evaluation via precision, recall, and F-scores. The use cases for linguistic classifiers presented in this chapter are content/topic classification, classification of syntactic structures both on a document-wide level for authorship attribution and on a word level for Part-of-Speech classification, and sentiment classification. All case studies are accompanied by extensive code labs which enable the reader to put the theory into practice, change parameters, and examine alternative datasets. A core contribution of this chapter is an approachable dissemination of logistic regression and feed forward networks. This section provides great value to the NLP novice in that both concepts constitute crucial pillars of a large range of NLP models both discussed within this book and stretching beyond its scope. Chapter two is then concluded with a discussion of implicit bias, emphasising that models might be non-generalisable in unpredictable ways. This is exacerbated by the fact that they cannot be manually evaluated fully due to their immense size. Addressing implicit bias in text classification thus involves careful dataset curation, ongoing error analysis, and validation against real-world scenarios to ensure that the models perform well across different contexts and for all categories of interest and to generate awareness for weaknesses the model may have.

Similarity models lie at the heart of Chapter three. This chapter thus extends the scope to capture clusters or networks of linguistic items, focusing on how strongly they are connected to one another via pairwise comparisons. Dunn explores this using three layers of analysis. Firstly, the author explores similarity measures when comparing whole corpora composed of different genres. Secondly, document similarity is explored. This, in turn, involves three sub-levels: content

similarity, authorship similarity, and sentiment similarity. Lastly, Dunn introduces word similarity and thus provides a beginner's guide to vector semantics. The word2vec algorithm is introduced and its underlying mechanic of learning word embeddings by predicting word co-occurrence patterns is yet again explored in a highly accessible manner. Use cases are provided, e.g. in the form of a comparison of word embeddings from *The New York Times* and congressional speeches, revealing differences in word associations. The entirety of these analyses are available as accessibly commented code labs and allow the reader to validate results as well as modify the code to be applied to their own research questions. Finally, the chapter introduces clustering methods, particularly the k-means clustering algorithm for generating semantic domains on the basis of the obtained similarity values. A particularly valuable contribution of this book is the critical discussion of standard practices in NLP such as k-means clustering since it requires specifying the number of clusters (k) in advance; the number of desired clusters is, however, largely arbitrary, and often not grounded in linguistic theory. Chapter three also illustrates how certain traditional corpus linguistic properties such as Part-of-Speech can aid semantic clustering by limiting ambiguity. Dunn provides several examples of semantic domains created using similarity models, again expanded on in the accompanying code labs. Lastly, Chapter three is concluded with a discussion of model discrimination. Dunn highlights the risks of models learning undesirable (i.e. racist or sexist) cues since vector semantics relies on the distribution of words within the training data. Should the training data contain negative stereotypes, as is often the case particularly when working with online data, the model is likely to perpetuate these which tends to conflict directly with ethical research guidelines. The need for critical assessment when employing such models and the importance of transparency in linguistic analysis is thus highlighted.

Chapter four discusses validation and visualisation techniques in computational linguistic analysis. The chapter begins by emphasizing the importance of reporting results with baselines to provide context and ensuring the robustness of results. The first subchapter provides an example of classifying congressional speeches by political party and demonstrates how statistical tests can be employed to determine the significance of differences in model performance. The latter sections of the chapter explore visualisation methods in detail, focusing on relational plots, box plots, heat maps, and choropleth maps. Box plots are employed to discuss possible overfitting issues in four classification models that have been generated as part of the preceding chapters. Particular attention is given to unmasking algorithms to partially circumvent the issue of a lack in interpretability of the model. Unmasking involves systematically removing the most predictive features from the model. Doing this sequentially then illustrates when performance drops occur in the model and thus allows partial insights into its inner workings allow-

ing the researcher to assess its reliability. The following subchapter explores the use of principal components analysis (PCA), an increasingly popular method in Corpus Linguistics (Mastropierro, 2017; Wilson Black et al., 2023), to reduce high-dimensional word embeddings to easily visualisable two dimensions. Dunn also discusses the use of Jaccard similarity to measure the overlap of word embeddings from different corpora. This approach allows for the generation of heatmaps highlighting variations in word meaning representations across different datasets. Lastly, Choropleth Maps are introduced as a visualisation option particularly relevant to researchers with an interest in linguistic diversity. Dunn uses data from tweets and web pages, respectively, to quantify linguistic diversity within countries via the Herfindahl–Hirschman Index (HHI) and visualises the output as a colour-coded world map. The final ethical considerations chapter addresses concerns related to data bias and unequal access to computational linguistic analysis in different parts of the world. It emphasises the need to improve representation for low resource languages and non-inner-circle varieties such as Nigerian English. Dunn also highlights projects like *GeoWAC* (Dunn & Adams, 2020) that aim to mitigate these biases. A very brief final chapter summarises the presented methods and points to the code labs as an opportunity for further engagement with the material.

Although the book has made numerous significant contributions towards making NLP methods accessible to a linguistic audience there are still certain constraints that require comment. Considering that the target audience is corpus linguists it is at times surprising to see oversimplifications such as "text classifiers have been shown to make very good predictions about the part of speech of individual words when trained on small amounts of annotated data" (p. 2) or "function words will not be helpful for making predictions about the topic of a document" (p. 6) without further discussion or references pointing towards research exploring POS-tagger accuracy or topic identification in a more nuanced way. This is particularly striking since function words have been found to be relevant predictors not only in authorship attribution contexts – "Function words can have a major effect on separating one manuscript witness from all of the others" (Honkapohja & Suomela, 2022: 776) – but also for register identification (Biber, 2012: 14), and even topic extraction itself (Nakamata, 2019: 230). The same line of argument applies to Part-of-Speech tagger accuracy. There are still major challenges to be overcome such as inconsistencies in accuracy depending on the language – Vries et al. (2022: 7679) report less than 80% accuracy using a state-of-the art POS tagger on Arabic language data – and gender-bias (Garimella et al., 2019). A brief mention of the conceptual placement of function words on the lexicogrammatical continuum (Langacker, 2008) would further have been commendable, especially since Dunn chooses to class *going* and *seem* as function words

(p. 22) which may surprise a traditional corpus linguist. Overall, it may have been desirable to focus on including and unifying corpus linguistic concepts with NLP methods to aid comprehension and emphasise important parallels. Chapter 3.4 in particular would have presented a prime opportunity for this via linking traditional collocation analysis, a concept intimately familiar to the corpus linguist, to word similarity measures.

Dunn's book rightly argues that a stronger focus on ethical questions in corpus/computational linguistics is necessary, particularly with regards to issues of privacy, ownership, and the perpetuation of biases. It is naturally beyond the scope of any single piece of academic writing to exhaustively cover ethical considerations that may influence NLP projects, both due to the great and ever-changing variety of available methods and due to the strong dependence of ethical issues on the dataset at hand. Nevertheless, a more thorough mention of three areas of research ethics would have been of immediate use to a corpus linguist looking to expand their NLP knowledge, i.e. the intended reader:

i. underlying assumptions
ii. interpretability
iii. methodological rigour.

This is particularly striking since a focus on methodological rigour and acknowledgement of complexity is the foundation of using NLP methods in an ethically sound manner.

Points that would have benefitted from greater focus on underlying assumptions in this book are explorations of what has been chosen as the basic unit of analysis (words/lemmas/character n-grams) in each use case and whether there is a theoretical justification for this. Discussions of this nature are often lacking in NLP literature and a linguistic audience is particularly well-equipped to contribute to methodological triangulation in this respect. A similar argument can be made in terms of assumptions made in Chapter two where Dunn equates dialect with country of origin of online data (p. 14). Given the nature of the book as a guide to NLP research a discussion surrounding the limitations, e.g. the possibility for tweets to originate from a certain country but being written in a completely different language/dialect, would have been essential.

While there is a brief mention of interpretability (p. 54) and robustness across different languages (p. 42), this discussion generally falls short considering its integral relevance to the corpus linguist. Dunn, for instance, provides examples from authorship attribution, a common task in applied fields like forensic linguistics which are likely to influence high-stake real-world decisions. The reader would therefore have benefitted from an emphasis on the limitations of black-box

models and the ethical question of who carries the responsibility for decisions made on the basis of ultimately opaque models.

A point of note in terms of methodological rigour is that the evaluation metric plays a decisive part in the interpretation of results from NLP methods, and Dunn exclusively relies on F-scores for this purpose. Even though the introduction of further evaluation metrics inevitably adds complexity to an introduction to NLP, it is essential to leave readers in a position where they are able to critically review what they should be measuring their results with. A brief mention of Kolmogorov–Smirnov tests (Djuric & Miguez, 2009), Cohen's Kappa scores (Powers, 2011), and Area under the ROC Curve (Powers, 2011) as well established, industry-standard alternatives would therefore have been desirable.

Whilst it must be acknowledged that it might have been helpful to add more nuanced discussions of the abovementioned points, this book nevertheless provides a very accessible introduction to NLP methods that are of immediate relevance to the corpus linguist looking to expand their toolkit. Especially the chapters detailing which calculations are carried out under the hood of large language models fill an egregious gap in existing literature and help demystify these approaches. The case studies and code labs in particular can be seen as a resource in their own right since they provide useful practical starting points for researchers from backgrounds as diverse as Corpus-Based Sociolinguistics, Corpus Stylistics, Multilingualism, and Discourse Analysis. This book is ideal for corpus linguists with no pre-existing knowledge of NLP methods but a basic understanding of programming, and whilst at times painting an oversimplified picture of the problems at hand, it provides a good overview of powerful state-of-the-art text classification and comparison techniques.

Funding

Open Access publication of this article was funded through a Transformative Agreement with Lancaster University.

References

Biber, D. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, *8*(1), 9–37.

Djuric, P.M., & Miguez, J. (2009). Model assessment with Kolmogorov-Smirnov statistics. In *IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 2973–2976). IEEE.

Dunn, J., & Adams, B. (2020). Geographically-balanced gigaword corpora for 50 language varieties. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 2528–2536). European Language Resources Association. https://aclanthology.org/2020.lrec-1.308/

Garimella, A., Banea, C., Hovy, D., & Mihalcea, R. (2019). Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. https://aclanthology.org/P19-1339/.

Honkapohja, A., & Suomela, J. (2022). Lexical and function words or language and text type? Abbreviation consistency in an aligned corpus of Latin and Middle English plague tracts. *Digital Scholarship in the Humanities*, *37*(3), 765–787.

Langacker, R. W. (2008). *Cognitive Grammar: A Basic Introduction*. Oxford University Press.

Mastropierro, L. (2017). *Corpus Stylistics in Heart of Darkness and its Italian Translations*. Bloomsbury Academic.

Nakamata, N. (2019). Vocabulary depends on topic, and so does grammar. *Journal of Japanese Linguistics*, *35*(2), 213–234.

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63. https://arxiv.org/pdf/2010.16061

Vries, W. de, Wieling, M., & Nissim, M. (2022). Make the best of cross-lingual transfer: Evidence from POS Tagging with over 100 Languages. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Volume 1: Long Papers*. Association for Computational Linguistics. https://aclanthology.org/2022.acl-long.529/.

Wilson Black, J., Brand, J., Hay, J., & Clark, L. (2023). Using principal component analysis to explore co-variation of vowels. *Language and Linguistics Compass*, *17*(1), Article e12479.

Address for correspondence

Hanna Schmück
Lancaster University
h.schmueck@lancaster.ac.uk

Publication history

Date received: 15 September 2023
Date accepted: 29 September 2023
Published online: 22 December 2023