Natural Language Processing (2025), 1–4
doi:10.1017/nlp.2024.64

CAMBRIDGE UNIVERSITY PRESS

Book Review

Data Analytics for Discourse Analysis with Python: The Case of Therapy Talk, by Dennis Tay. New York: Routledge, 2024. ISBN: 9781032419015(HB: USD $135.00), ISBN 9781003360292 (eBook: USD $41.24), xiii+182 pages.

In the past decade, the emergence of "big data" and the proliferation of new media have widely propelled the dissemination and utilization of data analytics in the field of discourse studies (Nartey and Mwinlaaru, 2019). Specifically, the large-scale and data-driven exploration can mitigate the "cherry-picking" effect commonly observed in discourse studies, thereby enhancing the persuasiveness and reliability of textual analysis through an objective and close textual examination in a quantitative way (Baker and Levon, 2015). In response to the rise of large language models and machine learning technology, Dennis Tay's latest book, *Data Analytics for Discourse Analysis with Python: The Case of Therapy Talk*, emerges as a cutting-edge guidance and introductory-level demonstration for data analytics by unitizing machine learning and Python in discourse analysis, particularly through the illustrative case study of psychotherapy talk. It is more exploratory than theory-oriented in that it follows a data-driven method and mines the text as it is without pre-specified predictors.

This book comprises six chapters. Chapter 1 functions as an introduction section to prime the readers with a foundational knowledge of data analytics in the context of discourse studies. Chapters 2–5 demonstrate the utilization of four data analytic techniques in discourse studies, namely, Monte Carlo simulations (MCS), cluster analysis, classification, and time series analysis, with each presented in a separate chapter to ensure clarity and facilitate a focused examination. As the concluding part, Chapter 6 offers supplementary constructive guidance for the future learning and comprehensive application of data analytics in scholarly research. It is worth noting that this book is exceptionally accessible, even to those with limited experience in data analytics within the domain of natural language processing, as it adopts a methodical and step-by-step approach to demonstrate the data processing techniques. Most importantly, it elucidates the fundamental logic and concepts associated with each technique while illustrating the data analytic process.

The initial introductory chapter focuses on the crucial role of data analytics and expounds on some quantification methods of textual examination in discourse analysis. Subsequent to addressing the four types of data analytics, namely, descriptive, diagnostic, predictive, and prescriptive analytics, the author underscores the significance of applying data analytics to psychotherapy talk research by emphasizing the significant potential for enhancing discourse analysis, both academically and in terms of its social practical implications for therapists. Furthermore, following a concise introduction to the key topics covered in this book, the procedure of doing the analytics and some quantification methods of language are comprehensively elaborated through illustrative examples, encompassing word embedding, Linguistic Inquiry and Word Count (LIWC) scoring, and alternative ways of computing tf-idf scores. Finally, detailed instructions for the installation and utilization of Python are provided, which would be very helpful and instructive for the novice Python users or researchers alike.

Chapter 2 expounds on application of MCS to address the problem of missing data in discourse studies in a stepwise fashion. Initially, the author briefly introduces the origin and the nature of MCS, emphasizing its practical potential to tackle the problems of missing or incomplete data with several real-life scenario analysis. By conducting a scenario analysis of the "birthday problem," the author clearly illustrates the accuracy of probability prediction and the law of large numbers by comparing the difference in probability prediction between an analytic approach and MCS. By taking the real-life scenario "spinning the casino roulette" as an example, the author explains two features of central limit theorem (i.e., the law of large numbers) and puts forward the application potential of MCS in simulating the missing data and accuracy testing in therapy talk transcripts. Last, also the most importantly, forty psychoanalysis session transcripts of therapy talk are examined as a case study to demonstrate the three steps carrying out MCS, including LIWC scoring, simulation runs with a train-test approach, and validation of the aggregated outcomes of MCS, followed by a discussion of the findings and insights into how to take the findings as the entry point for the theoretical hypothesis formulation.

Chapter 3 demonstrates the utilization of cluster analysis in discourse analysis to examine linguistic (a)synchrony in three types of therapist–client interaction. As a real-life case, COVID-19 pandemic datasets of sixty countries/regions with four potential properties serve as a pertinent example to illustrate and compare agglomerative hierarchical clustering (AHC, a major subtype of hierarchical algorithm) and k-means clustering (a specific type of nonhierarchical algorithm) by demonstrating Python operation for each one. AHC is used to determine the clustering division in a bottom-up way and produce clusters ordered in a hierarchy by Euclidean distance and Ward's linkage, while k-means clustering involves the optimization process of identifying cluster centroids and determining the clustering groups by Euclidean space as well as intuitive optimal valuation of the number and positions of clusters. As in Chapter 2, psychotherapy talk is presented as a case study again to demonstrate the systematic approach for quantifying linguistic synchrony between therapists and clients in a step-by-step way. This comprehensive process involves computing LIWC scores, k-means clustering with model validation for measuring distribution patterns of synchrony on a sessional and dyadic basis, validating the clustering solutions, and qualitative analysis of context-based synchrony construction.

Chapter 4 looks primarily at classification, a technique of supervised machine learning for modeling the mapping relationship between properties of objects (e.g., language or discourse) and existing groups or category labels (nonlinguistic groups). Following a conceptual introduction to classification, k-nearest neighbors (k-NN) is introduced as a classification algorithm to predict the group labels of targeted objects based on their Euclidean distances to the existing nearest neighbors that are already assigned a group label. In a similar vein, this chapter demonstrates the application of k-NN in examining psychotherapy talk as a case study, to predict therapy types from therapist–client language in a stepwise way, including LIWC scoring, deciding the k number of nearby clusters and different accuracy measures of the k-NN model.

Chapter 5 targets the fourth and final technique: time series analysis. The author initially introduces and highlights its potential usefulness in discourse studies for discovering patterns underlying the fluctuation, thereby facilitating modeling and enabling longitudinal or cross-sectional prediction. Meanwhile, the statistic logic is explicitly explained by an overview of structure and components of time series data as well as the elaboration of key related conceptual ideas, including structural signature, autoregressive and moving average models, and ARIMA models (autoregressive integrated moving average models). Subsequently, the author proceeds to demonstrate the "juice-extraction-from-sugarcane" process to model and forecast psychotherapy language patterns in a stepwise fashion, again in the case study of cross-sessional psychotherapy talk analysis. The significance of this chapter cannot be overstated, as it offers profound insights into enhancing discourse analysis beyond mere descriptive analysis towards a more predictive and prescriptive approach.

Chapter 6, the concluding part for this book, also carries a profound and illuminating importance. Followed by a brief summary of the crux of each chapter, the application potential of data

analytics in discourse analysis is elaborated in a metaphorical way. On the one hand, it acts as "a rifle" to enhance the efficacy and reliability of targeting at and analyzing the research questions; on the other hand, it also functions as "a spade" to cultivate new and unexpected findings, formulate new hypotheses, or even unveil the new direction for further researches. Most importantly, some learning resources and learning directions are put forward as a guidance for readers to enhance their data analytics mastery by combining technique learning and practical implementation. The author finally points out the potentials of applying data analytics to addressing specific questions in other discourse contexts, especially those with distinct dialogic and temporal nature, such as the widely researched contexts of social media, politics, and education.

This book's salient features, including real-life scenario exemplification, machine learning-based discourse study, and case demonstration of psychotherapy talk study, render it an essential and useful resource for scholars and students in discourse studies, natural language processing, psychology, social media, and computer science. Its significance can be summarized as follows: First, it excels in skillfully and metaphorically illustrating the intricate and profound logic of data analytics by blending analogy, metaphor, and simple real-life examples, thus enhancing the accessibility of this book to a larger readership across diverse research backgrounds. Second, it maintains a practical and instructive stance by relating the theoretical discussion to practice. Each chapter follows a consistent and streamlined structure comprising a key technique, a case study of therapy talk, and annotated Python code, which fosters a nuanced comprehension of the essence of data analytics and its application in discourse studies among readers. Furthermore, it adopts a reader-oriented perspective and reinforces readers' nuanced understanding of Python operating process, with detailed Python code and explanations in each step of demonstrating target technique application, along with a list of Python code offered at the end of each chapter for the convenience of reference.

Admittedly, there are also certain limitations to be considered. In contrast to the elaborate illustration of quantitative methods, the qualitative analysis of therapy talk seems to be less adequately demonstrated and lacks an in-depth exploration. Additionally, it is intended to be a brief primer on data analytics and only covers some basic data analytic methods. Therefore, readers with detailed knowledge of data analytics may find it overly introductory and insufficiently informative for tackling intricate text mining tasks.

Overall, Dennis Tay's new book, *Analytics for Discourse Analysis with Python: The Case of Therapy Talk*, is undoubtedly a pertinent and valuable contribution in the backdrop of increasing urgency of integrating new techniques into traditional discourse studies in the new era. It could serve not merely as the guidebook for the novice researchers navigating the nuances of data analytics but also as a cross-fertilization between disciplines, thus being insightful and practically significant for the scholars, students, and professionals alike in diverse fields related to linguistics, discourse studies, and computer programming languages.

Acknowledgments. This work was funded by the Second Round Chongqing Municipality First-Class Discipline Research Grant for Foreign Language and Literature (SISUWYJY202303) and the Chongqing Graduate Student Research Innovation Project (CYB240271).

Fengmei Cai
School of English Studies,
Sichuan International Studies University,
Chongqing, China

Xingbing Liu (Corresponding author)
School of Business English,
Sichuan International Studies University,
Chongqing, China
E-mail: liuxingbing@sisu.edu.cn

References

Nartey M. and **Mwinlaaru I. N.** (2019). Towards a decade of synergising corpus linguistics and critical discourse analysis: a meta-analysis. *Corpora* **14**(2), 203–235.

Baker P. and **Levon E.** (2015). Picking the right cherries? A comparison of corpus-based and qualitative analyses of news articles about masculinity. *Discourse & Communication* **9**(2), 221–236.