Innovative Technologica: Methodical Research Journal Vol: 3, No 3, 2024, Page: 1-6

Methods, Challenges, and Ethical Considerations in Data Collection of Corpus Compilation

Madina Dalieva

PhD, Associate Professor Uzbekistan State World Languages University Uzbekistan, Tashkent

DOI: https://doi.org/10.47134/innovative.v3i3.122
*Correspondence: Madina Dalieva
Email: m_dalieva@gmail.com

Received: 07-07-2024
Accepted: 18-08-2024
Published: 29-09-2024

Copyright: © 2024 by the authors. Submitted for open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

Abstract: Corpus compilation is a critical process in linguistics that involves gathering and organizing large datasets for language analysis and model training. This article examines key aspects of corpus compilation, with a particular focus on data collection. It explores the sources of data, strategies for ensuring representativeness, and challenges such as copyright constraints and data quality issues. Ethical considerations, such as anonymization and consent, are also discussed. By understanding these factors, researchers can build effective and ethically sound corpora for linguistic research and computational applications

Keywords: Corpus Compilation, Data Collection, Linguistics, Natural Language Processing, Representativeness, Copyright, Ethics In Research, Data Preprocessing

Introduction

A corpus, which is a structured collection of written or spoken texts, serves as the foundation for empirical studies on language use, and is indispensable for both qualitative and quantitative analyses. Linguists, language educators, and computational researchers use corpora to explore everything from syntax and semantics to sociolinguistic variation and language change over time(Jablotschkin, 2024; Masua, 2024; Paquot, 2024).

The process of corpus compilation involves carefully selecting, collecting, and organizing text data to ensure that the resulting corpus is representative, balanced, and useful for the research questions or computational models at hand. Given the wide range of applications for corpora, the compilation process must account for several factors: the type of data (e.g., written versus spoken), the diversity of sources, ethical considerations, and legal constraints such as copyright(Alfuraih, 2024; Chen, 2023; Goncharova, 2024; Lanza, 2023; Rackevičienė, 2023).

Methodology

This article will explore the critical aspects of corpus compilation, focusing on data collection as the cornerstone of the process. We will examine the types of data sources

available, the methods for ensuring representativeness, and the challenges that arise when constructing a high-quality corpus. Special attention will be given to ethical concerns and the practical difficulties associated with gathering diverse linguistic data. By understanding these complexities, researchers can build corpora that are both linguistically robust and legally and ethically sound.

Result and Discussion

In the Results section, summarize the collected data and the analysis performed on those data relevant to the issue that is to follow. The Results should be clear and concise. It should be written objectively and factually, and without expressing personal opinion. It includes numbers, tables, and figures (e.g., charts and graphs). Number tables and figures consecutively in accordance with their appearance in the text(Alfraidi, 2022; Maffei, 2023; Oushiro, 2023).

Data collection is the backbone of corpus compilation, as the quality, representativeness, and scope of the collected texts directly influence the utility of the resulting corpus. The aim is to gather a diverse and balanced set of texts that accurately reflect the linguistic phenomena under study. This step is especially critical because the choice of texts will ultimately shape the insights drawn from linguistic or computational analyses.

Sources of Data

The types of sources used in corpus compilation can vary depending on the intended purpose of the corpus. Common sources include both written and spoken texts. Written texts are often drawn from newspapers, books, academic papers, blogs, and websites, each providing different registers and styles of language use (Biber, Conrad, & Reppen, 1998). For example, the British National Corpus (BNC) was designed with a wide variety of written sources to capture both formal and informal registers of British English (Leech, 1992). Such diversity is essential to ensure the corpus is comprehensive enough to support various types of linguistic inquiries.

In contrast, spoken corpora require recordings of real-world conversations, which can include interviews, debates, and spontaneous dialogue. Capturing natural spoken data is more challenging than compiling written texts, as it involves transcription and often ethical considerations related to privacy. Nonetheless, spoken language corpora are invaluable for studying the dynamic, interactive nature of language (McEnery & Hardie, 2012). An example of this is the Spoken BNC, which focuses exclusively on everyday spoken English (Aston & Burnard, 1998).

In some cases, corpus compilers focus on domain-specific language. For instance, specialized corpora are constructed to investigate the vocabulary, grammar, and usage in fields like medicine, law, or engineering. Such corpora are often based on professional or academic publications within a particular discipline, which makes them invaluable for understanding specialized jargon and terminologies (Bowker & Pearson, 2002).

Criteria for Selection

When collecting data, ensuring that the corpus is representative of the target language or dialect is a primary concern. Achieving representativeness involves careful consideration of the texts' diversity, genre balance, and temporal scope.

One of the key strategies for ensuring linguistic variety in a corpus is to include texts from multiple genres. Different genres exhibit different linguistic features, and capturing this variation is crucial for a comprehensive analysis of language. Biber (1993) emphasized the importance of including diverse text types, such as news articles, fiction, essays, and spoken dialogues, to reflect genre-based language differences. This approach is exemplified by the Longman Spoken and Written English Corpus, which was designed to capture variations in English across a broad range of genres (Biber et al., 1999).

In addition to genre diversity, sampling strategies play a significant role in corpus compilation. Stratified sampling is often used to ensure balanced representation of different text categories, such as region, demographics, or language register (McEnery & Hardie, 2012). This approach allows for the construction of a corpus that can be used to explore sociolinguistic variables and linguistic variation. For instance, some corpora aim to represent formal academic registers as well as informal, conversational language, allowing for comparative linguistic analysis (Gries, 2009).

Temporal variation is another important consideration, especially when compiling a diachronic corpus. Including texts from different time periods allows researchers to study how language changes over time. Historical corpora, such as the Helsinki Corpus, which spans from Old English to Modern English, are invaluable resources for diachronic studies of language evolution (Rissanen, Kytö, & Heikkonen, 1996). By capturing language from various time periods, researchers can observe linguistic trends and shifts in usage.

Challenges in Data Collection

Despite the clear benefits of a well-compiled corpus, there are numerous challenges associated with data collection. One significant hurdle is dealing with **copyright and legal considerations**. Many texts, particularly those from books or major news outlets, are protected by copyright, meaning that corpus compilers must obtain permission to use the texts. In some cases, the texts may not be available for public use at all, which limits the scope of the corpus (Kennedy, 1998). This is particularly relevant when working with contemporary written texts or user-generated content, such as social media posts, where permissions and ethical considerations come into play.

Another challenge is ensuring **representativeness and avoiding bias**. Language use varies across social, regional, and demographic groups, and corpora must account for this variation. A common issue is that certain forms of language, such as informal online discourse, may overrepresent specific demographic groups, such as younger, digitally literate individuals, while underrepresenting older or less digitally connected populations (Meyer, 2002). To avoid these biases, corpus compilers must carefully select their data sources and ensure that the corpus reflects a broad range of language users.

Finally, **data quality and noise** can pose significant obstacles, especially when working with data scraped from the web or user-generated content. Such texts often contain irrelevant information, such as HTML tags or advertising content, which must be removed

during preprocessing. Web-scraped data is particularly prone to containing duplicated or poorly formatted text, which can compromise the quality of the corpus (Baisa & Suchomel, 2014). Careful cleaning and preprocessing are necessary to ensure that the data is suitable for linguistic analysis.

Conclusion

In conclusion, corpus compilation is a meticulous yet essential task that underpins both linguistic research applications. Data collection forms the cornerstone of this process, requiring careful consideration of data sources, sampling strategies, and ethical guidelines. Overcoming challenges related to legal constraints, biases, and data quality ensures the corpus's representativeness and reliability. Additionally, adherence to ethical practices—such as anonymization and proper consent—ensures the responsible use of data. By following best practices in data collection and compilation, researchers can create high-quality corpora that yield meaningful insights and support advanced computational models, ultimately driving progress in both language studies and technology development.

References

Aston, G., & Burnard, L. (1998). *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh University Press.

Baisa, V., & Suchomel, V. (2014). Sketch Engine for Noisy Data: Evaluating Word Sketches. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 27-31 May 2014, Reykjavik, Iceland.

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing, 8*(4), 243-257.

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

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Pearson Education Limited.

Bowker, L., & Pearson, J. (2002). *Working with Specialized Language: A Practical Guide to Using Corpora*. Routledge.

Crowdy, S. (1993). Spoken Corpus Design. *Literary and Linguistic Computing, 8*(4), 259-265.

Gries, S. T. (2009). *Statistics for Linguistics with R: A Practical Introduction*. Mouton de Gruyter.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. Longman.

Leech, G. (1992). 100 Million Words of English: The British National Corpus (BNC). *Language Research, 28*(1), 1-13.

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

Meyer, C. F. (2002). *English Corpus Linguistics: An Introduction*. Cambridge University Press.

Rissanen, M., Kytö, M., & Heikkonen, K. (1996). *The Helsinki Corpus of English Texts: Diachronic and Dialectal*. Department of English, University of Helsinki.

Alfraidi, T. (2022). The Saudi Novel Corpus: Design and Compilation. *Applied Sciences (Switzerland)*, *12*(13). https://doi.org/10.3390/app12136648

Alfuraih, R. F. (2024). Exploitation and Evaluation of an Arabic-English Composite Learner Translator Corpus. *International Journal of Arabic-English Studies*, *24*(1), 155–172. https://doi.org/10.33806/ijaes.v24i1.552

Chen, L. (2023). The Knowledge Tools of Ancient China and the Construction of Classical Knowledge Repositories. *Journal of Library Science in China*, *49*(3), 19–40. https://doi.org/10.13530/j.cnki.jlis.2023019

Goncharova, O. V. (2024). Data Mining Efficiency in the ESG Indexes Verbalization Analysis (on the Example of the MSCI Site). *Advances in Science, Technology and Innovation*, 13–16. https://doi.org/10.1007/978-3-031-49711-7_3

Jablotschkin, S. (2024). DE-Lite – a New Corpus of Easy German: Compilation, Exploration, Analysis. *LT-EDI 2024 - 4th Workshop on Language Technology for Equality, Diversity, Inclusion, Proceedings of the Workshop*, 106–117.

Lanza, D. F. (2023). THE SPOKEN CORPORA OF CENTRAL AMERICAN SPANISH: COMPILATION AND EVALUATIVE OVERVIEW. *Normas*, *13*(1), 83–111. https://doi.org/10.7203/Normas.v13i1.27658

Maffei, D. P. (2023). The Database of Hellenistic Inscribed Epigrams from Doric-speaking Areas. *Journal of Open Humanities Data*, *9*. https://doi.org/10.5334/johd.134

Masua, B. (2024). In the heart of Swahili: An exploration of data collection methods and corpus curation for natural language processing. *Data in Brief*, *55*. https://doi.org/10.1016/j.dib.2024.110751

Oushiro, L. (2023). Computational resources for handling sociolinguistic corpora. *The Handbook of Usage-Based Linguistics*, 417–434. https://doi.org/10.1002/9781119839859.ch23

Paquot, M. (2024). The Core Metadata Schema for Learner Corpora (LC-meta). *International Journal of Learner Corpus Research*. https://doi.org/10.1075/ijlcr.24010.paq

Rackevičienė, S. (2023). LITHUANIAN-ENGLISH CYBERSECURITYTERMBASE:

PRINCIPLES OF DATA COLLECTION AND STRUCTURING. *Rasprave Instituta Za Hrvatski Jezik i Jezikoslovlje*, *49*(2), 439–461. https://doi.org/10.31724/rihjj.49.2.12