Corpus Studies in 2024: Emerging Trends and Applications

Aleksandra Tomaszewska, Institute of Computer Science, Polish Academy of Sciences

Maciej Ogrodniczuk, Institute of Computer Science, Polish Academy of Sciences

Abstract

The paper provides an overview of corpus studies as of early 2024, highlighting the trend towards combining data-driven and traditional methodologies. Facilitated by technological advancements, this integration marks a convergence of corpus and computational linguistics. We outline the characteristics of contemporary corpora, such as their modality, metadata, size, specialization, and present their examples. We also trace the historical development of corpus methodologies, identifying emerging trends in corpus design and construction and their increasingly cross-disciplinary academic and non-academic uses. We further explore concordancers, presenting available tools, their evolution, ease of use, fast data processing, multilingual capabilities, and sharing features. The impact of large language models (LLMs) on corpus studies is also examined, acknowledging their advantages while also discussing the challenges they introduce, underscoring the continuous need for creating and annotating corpora. Lastly, we address current challenges in corpus studies, focusing on reducing content bias, adhering to ethical standards, and complying with FAIR data principles such as the importance of replicability, transparency, and the accessibility of corpus data.

Keywords: corpus studies, corpora, corpus tools, computational linguistics, NLP, LLM