INTRODUCTION TO CORPUS LINGUISTICS AND ITS HISTORY

Nurulloyeva Zarina
Bukhara State University, Faculty of Foreign Languages

ABSTRACT

In the field of language studies, a corpus refers to a collection of written texts or transcribed speech that is used for linguistic analysis and description. The development and examination of corpora stored in computerized databases over the past thirty years have given rise to a fresh academic field called corpus linguistics. While corpus linguistics is not a standalone objective, it serves as a valuable resource for enhancing explanations of language structure and usage, as well as for a range of practical applications such as natural language processing by computers and gaining insights into language learning and teaching methodologies.

Key words: Linguistics, corpus, analisys, language, methodology, computer.

Introduction

Corpus linguistics encompasses the compilation and analysis of collections of spoken and written texts as the source of evidence for describing the nature, structure, and use of languages.[1] Corpus linguistics is a branch of linguistics that involves the study of language using large, structured collections of texts known as corpora. These corpora serve as databases that linguists analyze to investigate language patterns, usage, and structure:

1. History:

The history of corpus linguistics demonstrates a continuous drive towards refining methodologies, creating more extensive and diverse corpora, and developing sophisticated analytical tools to extract rich insights from linguistic data. By integrating advances in technology and computational linguistics, corpus linguistics has enhanced our understanding of language at a granular level, offering empirical evidence and data-driven approaches to linguistic analysis. “As corpus linguistics continues to grow, collaborative efforts among researchers, institutions, and technology providers have led to the creation of large-scale corpora like the British National Corpus and the Corpus of Contemporary American English, strengthening the foundation of corpus-based research.”[2] Looking ahead, the future of corpus linguistics holds promising opportunities for further exploration, innovation, and application in a wide range of

[1] International Encyclopedia of the Social & Behavioral Sciences, 2001

[2] Wikipedia.org

disciplines, contributing to a deeper comprehension of language and communication in diverse contexts.

A landmark in modern corpus linguistics was the publication of Computational Analysis of Present-Day American English in 1967. Written by Henry Kučera and W. Nelson Francis, the work was based on an analysis of the Brown Corpus, which was a contemporary compilation of about a million American English words, carefully selected from a wide variety of sources.[3]

- The use of corpora in linguistics gained significant momentum with the development of computer technology, which enabled the storage, search, and analysis of vast amounts of text data.

Corpus in the 1960s, a seminal collection of English texts used for linguistic analysis. In addition to its traditional linguistic applications, corpus linguistics has expanded into various academic and professional domains. For instance, researchers are now exploring the intersection of Law and Corpus Linguistics, a field dedicated to analyzing legal texts utilizing corpus data and analytical tools. Moreover, datasets like the DBLP Discovery Dataset focus on computer science, housing relevant publications enriched with detailed metadata such as author details, citations, and research areas. Another specialized dataset, NLP Scholar, amalgamates papers from the ACL Anthology with metadata from Google Scholar. Corpora also play a role in supporting translation endeavors and aiding foreign language instruction.

2. Methodology:

Corpus linguistics has generated a number of research methods, which attempt to trace a path from data to theory. Wallis and Nelson (2001)[4]first introduced what they called the 3A perspective: Annotation, Abstraction and Analysis.

- Annotation consists of the application of a scheme to texts. Annotations may include structural markup, part-of-speech tagging, parsing, and numerous other representations.
- Abstraction consists of the translation (mapping) of terms in the scheme to terms in a theoretically motivated model or dataset. Abstraction typically includes linguist-directed search but may include e.g., rule-learning for parsers.
- Analysis consists of statistically probing, manipulating and generalising from the dataset. Analysis might include statistical evaluations, optimisation of rule-bases or knowledge discovery methods.

[3] Francis, W. Nelson; Kučera, Henry (1 June 1967). Computational Analysis of Present-Day American English. Providence: Brown University Press.

[4] Wallis, S. and Nelson G. Knowledge discovery in grammatically analysed corpora. Data Mining and Knowledge Discovery, 5: 307–340. 2001.

Most lexical corpora today are part-of-speech-tagged (POS-tagged). However even corpus linguists who work with 'unannotated plain text' inevitably apply some method to isolate salient terms. In such situations annotation and abstraction are combined in a lexical search.

The advantage of publishing an annotated corpus is that other users can then perform experiments on the corpus (through corpus managers). Linguists with other interests and differing perspectives than the originators' can exploit this work. By sharing data, corpus linguists are able to treat the corpus as a locus of linguistic debate and further study.[5]

- Corpus linguistics involves collecting, annotating, and analyzing large corpora of text data to uncover patterns of language use.
- Linguists use software tools and statistical techniques to process and analyze corpus data, identifying word frequencies, collocations, syntactic structures, and other linguistic features.
- Corpus linguistics allows researchers to study language empirically, based on real-world language use, rather than relying solely on intuitions or examples.

3. Applications:

- Corpus linguistics has diverse applications in various fields, including:
  - Language Teaching and Learning: Corpora are used to develop language teaching materials, assess language proficiency, and study language acquisition processes.
  - Lexicography: Corpora aid lexicographers in compiling dictionaries, defining word meanings, and identifying usage patterns.
  - Natural Language Processing (NLP): Corpora play a crucial role in training and improving NLP algorithms for tasks such as machine translation, text classification, and sentiment analysis.
  - Stylistics and Discourse Analysis: Corpus linguistics is used to study stylistic features, discourse structures, and language variation in different contexts.

4. Advantages

- Corpus linguistics offers a data-driven approach to studying language, allowing researchers to make evidence-based claims about linguistic phenomena.
- Corpora provide a representative sample of language use, enabling researchers to explore variability within languages and across different genres and registers.
- By analyzing patterns in large data sets, corpus linguistics can uncover subtle linguistic features, track language change over time, and inform theoretical models of language structure and use. “Corpus inguistics is thus now inextricably linked to the computer, which has introduced incredible speed, total accountability, accurate

[5] Baker, Paul; Egbert, Jesse, eds. (2016). Triangulating Methodological Approaches in Corpus-Linguistic Research. New York: Routledge.

replicability, statistical reliability and the ability to handle huge amounts of data. With modern software, computer-based corpora are easily accessible, greatly reducing the drudgery and sheer bureaucracy of dealing with the increasingly large amounts of data used for compiling dictionaries and other information sources. In addition to greatly increased reliability in such basic tasks as searching, counting and sorting linguistic items, computers can show accurately the probability of occurrence of linguistic items in text."[6]

Conclusion

Corpus linguistics has evolved from its inception in the mid-20th century to become a foundational methodology in modern linguistics research. The development of corpora like the Brown Corpus and subsequent specialized collections has revolutionized the study of language structure, usage, and variation. Over time, corpus linguistics has expanded beyond pure linguistic inquiry into interdisciplinary domains such as law, computer science, and translation studies, showcasing its versatility and applicability across diverse fields.

References

1) International Encyclopedia of the Social & Behavioral Sciences, 2001
2) Francis, W. Nelson; Kučera, Henry (1 June 1967). Computational Analysis of Present-Day American English. Providence: Brown University Press.
3) Wallis, S. and Nelson G. Knowledge discovery in grammatically analysed corpora. Data Mining and Knowledge Discovery, 5: 307–340. 2001.
4) Baker, Paul; Egbert, Jesse, eds. (2016). Triangulating Methodological Approaches in Corpus-Linguistic Research. New York: Routledge.
5) Graeme Kennedy,(2014). An introduction to Corpus linguistics. Routlage. p-5
6) Wikipedia.org

[6] Graeme Kennedy,(2014). An introduction to Corpus linguistics. Routlage. p-5