TITLE: “CORPUS LINGUISTICS AND ITS HISTORY”

Baxshullayeva Munisa Ixtiyor qizi
A student of the
Bukhara State University
munisabaxshullayeva0602@gmail.com
Ergashova Oqila Oybek qizi
A student of the
Bukhara State University
ergashova oqila0102@gmail.com
Haydarova Muharram Dilshod qizi
A student of the
Bukhara State University
muharramhaydarova2 @gmail.com

ABSTRACT: This article delves into the fascinating history of corpus linguistics, tracing its evolution from its humble origins to its current status as a transformative force in the study of language. We explore the key milestones in its development, highlighting the pivotal contributions of pioneering researchers who recognized the power of large-scale language data. From the early days of statistical analysis to the advent of digital corpora and sophisticated computational tools, the article charts the trajectory of this field and its impact on various linguistic disciplines. We examine the theoretical underpinnings of corpus linguistics, its diverse applications in areas like language acquisition, lexicography, and language variation, and the ethical considerations surrounding the use of large language datasets. By shedding light on the past, present, and future of this vibrant field, this article underscores the crucial role that corpus linguistics plays in deepening our understanding of the complexities of human language.

Key words: corpus linguistics , origins , statistical analysis , digital corpora , computational tool , lexicography .

INTRODUCTION

For centuries, linguists have sought to understand the intricate workings of human language. They have meticulously analyzed texts, dissected grammatical structures, and traced the evolution of words and sounds. Yet, despite these efforts, a true grasp of the complexities of language remained elusive. The advent of corpus linguistics, however, marked a paradigm shift in our approach to language study. No longer limited to anecdotal examples and small, carefully curated datasets, researchers gained access to vast troves of digitized language, opening up new avenues for investigation. This article delves into the fascinating history of corpus linguistics, exploring its origins, its evolution, and its profound impact on our understanding of how language functions and evolves. We will journey from the early days of statistical analysis to the rise of digital corpora and the development of powerful computational tools, uncovering the key figures and groundbreaking discoveries that have shaped this field. Join us as we unravel the story of how data became the key to unlocking the secrets of language.

METHODS

Historical Texts: Examining seminal works by key figures in the field, such as J.R. Firth, Brown, and Kucera, as well as publications from prominent journals like International Journal of Corpus Linguistics* and *Corpus Linguistics and Linguistic Theory. Historical Events: Identifying key moments in the development of corpus linguistics, such as the emergence of the Brown Corpus, the development of the internet, and the rise of Natural Language Processing (NLP) technologies.

Theoretical Frameworks: Analyzing the underlying theoretical principles that underpin corpus linguistics, including the principles of frequency, distributional analysis, and statistical significance.

Case Studies: Examining specific applications of corpus linguistics in different areas, such as lexicography, language acquisition, and sociolinguistics.

Through a combination of these sources, this article aims to provide a comprehensive overview of the evolution of corpus linguistics, tracing its trajectory from its origins in the mid-20th century to its current status as a dominant force in linguistic research. The article will also address the ethical considerations associated with the use of large-scale language datasets and explore the potential future directions of this dynamic field.

In fact, Kennedy (1998: 13) believes Alexander Cruden to have been the author of the most well-known edition of a biblical concordance, first published in 1737. However, Leech (1992) argues that corpus linguistics is a recent methodology and that it introduces a new approach: “I wish to argue that computer corpus linguistics defines not just a newly emerging methodology for studying language, but a new research enterprise, and in fact a new philosophical approach to the subject” (LEECH 1992: 106–107). In the same vein, Stubbs (1997) points out that corpus linguistics is not merely a tool, but an important concept in linguistic theory: “First, corpus linguistics is a view about data: many different methods can be used to analyse corpus data. Second, a corpus is not just a tool, but a major concept in linguistic theory” (STUBBS 1997: 300).

In the 1940s and 1950s, American structuralists greatly contributed to the flourishing of corpus analysis. However, between 1950 and 1980, these analyses lost their significance due to Chomsky’s criticism, but this was regained with the emergence of computers in 1980–1990s. Chomsky’s criticism was so influential that

McEnery and Wilson (2001), in an attempt to exemplify research prior to Chomsky and whose methodological approach is based upon corpus linguistics, use the term ‘early corpus linguistics’: Early corpus linguistics is a term we will use here to describe linguistics before the advent of Chomsky. In this examination of early corpus linguistics, we will imply that linguistics before Chomsky was entirely corpus-like. This is both true and untrue. The dominant methodological approach to linguistics immediately prior to Chomsky was based upon observed language use. The debate that Chomsky reopened in linguistics was, however, a very old one, as will be seen. Part of the value of looking in some detail at Chomsky`s criticisms of corpus data.

If we extract the most frequent phraseology around frequent words, we can investigate a general hypothesis (proposed by Sinclair et al 1970, Sinclair 1999, Summers 1996): that frequent words are frequent because they occur in frequent phrasal constructions which express conventional pragmatic functions in text. For example, the word-form way is amongst the top ten nouns: somewhere around rank 100 in frequency lists from large corpora. Its general phraseology has been discussed within both pattern grammar (Sinclair 1999, Hunston & Francis 2000: 101-02, 110) and construction grammar (Goldberg 1996).

RESULT

The results of analyzing the portrayal of statistical Analysis: Pioneers like J.R. Firth and Zellig Harris recognized the value of large-scale language data for statistical analysis. Firth's work emphasized the importance of language in context, laying the groundwork for corpus-based methods.

* Early Corpora: The creation of the "Brown Corpus" (1961) marked a turning point, providing researchers with a substantial, standardized collection of English texts for quantitative analysis.

* Rise of Computational Tools: The development of computers and software programs enabled researchers to process and analyze vast amounts of data efficiently, paving the way for more sophisticated corpus-based studies.

DISCUSSION

Across the whole corpus the collocates make explicit the pragmatic force of the construction. To the left are words from the semantic fields of "achievement" and "progress". To the right are words from the fields of "plans" and "struggle". In individual texts the specific collocates contribute to textual cohesion. Once we have this candidate for a phrasal construction, we can use the p-frame software plus concordance lines to identify related recurrent strings. This depends on subjective decisions: but these decisions are based on replicable quantitative data. A wider search produces other, also highly conventionalized, ways of expressing this complex speech act.

• (see / know) which way the wind is blowing; has become a way of life; if that's the way you want it; laughing all the way to the bank; let me put it this way; only one way to find out; that's one way of putting it; that's the way I look at it; there is no way of knowing / telling; well on the way to recovery

Although native speakers cannot generate comprehensive lists of such phrases from introspection, they recognize them as idiomatic and conventional ways of expressing culturally important meanings. Several have have strong pragmatic meanings, including speech acts (e.g. threat) and discourse markers (which open or close discourse sequences).

CONCLUSION

The findings of this research article demonstrate the techniques I have discussed are very good at revealing communicative acts of specific kinds. If recurrent word-strings are both frequent and fairly evenly distributed across a corpus, then it follows that they have little to do with the content of individual texts, but rather with general communicative functions which speakers frequently express, independently of what they are talking about. The techniques can discover conventional ways of managing information and of expressing 'the typical meanings that human communication encodes' (Francis 1993: 155). There are no purely automatic inductive discovery methods for identifying phrasal constructions. However, automatic methods can find recurrent strings with limited formal variation and provide empirical quantitative data on phraseology. All methods have their limitations, but the question is not: Do they tell us everything we want to know about phraseology? (Clearly no.) But are they better than trying to discover patterns by introspection? (Clearly yes.

The history of corpus linguistics reveals a journey from humble beginnings to a dominant force in linguistic research. It is a testament to the power of data and computational tools to unlock the hidden secrets of language. By embracing vast amounts of digitized text, researchers have shifted from subjective interpretations to objective, data-driven analysis, revealing the intricate patterns and variations within human communication.

While corpus linguistics has revolutionized language study, its impact extends far beyond academia. Its applications in lexicography, language teaching, natural language processing, and even social sciences highlight its versatility. However, the field faces ongoing challenges, particularly in navigating the ethical implications of utilizing large datasets. As we move into a future defined by even greater data availability, researchers must prioritize responsible data collection, ensuring fairness, privacy, and transparency in their work.

REFERENCES

1. Allén, S. et al 1975. Nusvensk frekvensordbok. Stockholm: Almqvist & Wiksell.

2. Ayscough, S 1790. An Index to the Remarkable Passages and Words Made Use of by Shakespeare. London: Stockdale.
3. Bally, C. 1909. Traité de stylistique française. Geneva: Librairie Georg & Cie.
4. Cruden, A. 1737. A Complete Concordance to the Holy Scriptures. London: Tegg.
5. Firth, J. R. 1957. A synopsis of linguistic theory 1930-1955. Transactions of the
6. Philological Society. Special Volume. Studies in Linguistic Analysis. Oxford:
7. Blackwell. 1-32.
8. Fletcher, W. 2003-05. PIE Phrases in English.. http://pie.usna.edu.
9. Francis, G. 1993. A corpus-driven approach to grammar. In M. Baker et al (eds.) Text and Technology. Amsterdam: Benjamins. 137-56.
10. Francis, G., Hunston, S. & Manning, E. 1996. Grammar Patterns. 1: Verbs. London: HarperCollins.
11. Goldberg, A. E. 1996. Making one's way through the data. In Shibatani, M. & Thompson, S. A. (eds) Grammatical Constructions. Oxford: OUP. 29- 53.
12. Hunston, S. (2002) Corpora in Applied Linguistics. Oxford: OUP.
13. Hunston, S. & Francis, G. 2000. Pattern Grammar. Amsterdam: Benjamins.
14. Luhn, H. P. 1960. Keyword-in-context index for technical literature. American
15. Documentation, xi, 4: 288-95.
16. Morris, C. W. (1938) Foundations of the theory of signs. In O. Neurath, R. Carnap & C. W. Morris eds International Encyclopedia of Unified Science. Chicago: Chicago University Press. 77-138.
17. Palmer, H. E. 1933. Second Interim Report on Collocations. Tokyo: Kaitakusha.
18. Reed, A. 1986. DOC: CLOC V00309. Available at http://www.decus.org/libcatalog/document_html/v00309_1.html. Accessed Nov 2005.
19. Sinclair, J. 1998. The lexical item. In E. Weigand (ed) Contrastive Lexical Semantics. Amsterdam: Benjamins. 1-24.
20. Sinclair, J. 1999. A way with common words. In H. Hasselgård & S. Oksefjell (eds) Out of Corpora. Amsterdam: Rodopi. 157-79.
21. Sinclair, J. 2005. The phrase, the whole phrase and nothing but the phrase. Plenary lecture to Phraseology 2005, Louvain-la-Neuve, October 2005.
22. Sinclair, J. M., Jones, S. & Daley, R. 1970/2004. English Collocation Studies: The OSTI Report. (Ed.) R Krishnamurthy. London: Continuum. [Originally circulated as a mimeoed report in 1970.]