https://www.sworldjournal.com/index.php/swj/article/view/swj30-04-041
DOI: 10.30888/2663-5712.2025-30-04-041

УДК 81’33-047.44:81-112+81-115+81-139

CORPORA AND THE STUDY OF LANGUAGE EVOLUTION: A HISTORICAL LINGUISTICS PERSPECTIVE

Tsanko I.I. / Цанько І.І.
Teaching assistant / асистент
ORCID: 0009-0006-8835-9844
State University “Uzhhorod National University”,
Uzhhorod, Universytetska, 14, 88000
ДВНЗ “Ужгородський національний університет”,
Ужгород, Університетська, 14, 88000

Abstract. *This article explores the integration of diachronic corpora into historical linguistic research, focusing on their role in analyzing language change over time. The study outlines key diachronic corpora, such as the Helsinki Corpus of English Texts and the Corpus of Historical American English, and examines their applications in phonological, morphological, syntactic, and semantic analysis. It discusses various corpus-based methodologies, including frequency analyses, collocation analysis, and syntactic change analysis, demonstrating their effectiveness in tracing linguistic evolution. Additionally, the article highlights the advantages of quantitative approaches in historical linguistics and considers future directions for corpus-based research, including the use of big data and computational tools. By incorporating diachronic corpora into linguistic studies, researchers can gain deeper insights into language development, advancing both theoretical and empirical perspectives in the field.*

Key words: *corpus linguistics, historical linguistics, diachronic corpora, language change, comparative method, corpus-based analysis, computational linguistics.*

Introduction.

The study of language change over time has been a central focus of historical linguistics, with traditional methods such as the comparative method (Campbell, 2013) and internal reconstruction (Ringe, 2005) serving as key analytical tools. However, the increasing availability of large-scale linguistic data has led to the emergence of corpus-based approaches, transforming historical linguistic research (McEnery & Hardie, 2011). Corpus linguistics, which enables systematic and data-driven analysis, has proven particularly valuable for studying diachronic language change. The development of historical corpora, such as the Helsinki Corpus (Kytö & Rissanen, 1992) and the Corpus of Historical American English (Alatrash et al., 2020), has provided researchers with extensive datasets to track linguistic evolution in phonology, morphology, syntax, and semantics.

A growing body of research highlights the effectiveness of corpus-based methodologies in historical pragmatics, where scholars investigate the contextual use of language in historical texts (Kytö, 2011). Recent studies have also explored the role of big data analytics and computational tools in refining corpus-based linguistic analysis (Weisser, 2015; Dalieva, 2024). Moreover, corpus-driven studies contribute to language preservation and the reconstruction of lesser-documented languages, as demonstrated in research on endangered languages (Jenifer, 2022). The integration of corpus methodologies with traditional linguistic analysis continues to shape historical linguistics, offering new insights into language variation and change (Breitbarth et al., 2023).

This article examines the role of corpus linguistics in historical linguistic research, discussing its applications in tracking linguistic change, analyzing discourse structures, and identifying pragmatic shifts. By reviewing key diachronic corpora, methodological advancements, and the impact of computational tools, this study highlights the transformative potential of corpus-based approaches in understanding the evolution of language.

Main text.

Historical linguistics is the field of linguistics concerned with "…the study of language change, of how and why languages change" (Campbell and Mixco, 2007, p. 77). It examines the origins and development of individual languages and language groups through methods such as internal reconstruction and, when linguistic data is unavailable, external reconstruction. Furthermore, historical linguistics aims to develop a typology of processes leading to language change, including phonological, morphological, syntactic, and semantic shifts. These changes can be analyzed through various theoretical lenses, including articulatory phonetics, cognitive linguistics, sociolinguistics, and communication theory (Bussmann, Kazzazi and Trauth, 2006, p. 513).

Historically, scholars have employed several methods to analyze language change, reconstruct earlier language stages, and classify languages into families. One of the most central techniques in historical linguistics, as noted by Campbell (2013, p. 107), is the comparative method. This method involves comparing related languages within the same family to reconstruct a proto-language, assuming that languages in a family descend from a single common ancestor, which evolves through dialectal changes and later diverges into distinct languages. The ultimate goal of the comparative method is to reconstruct as much of the proto-language as possible by analyzing changes in phonology, vocabulary, and grammar.

Ringe (2005, p. 244) explains that another important technique is internal reconstruction, which involves using patterns found in the synchronic grammar of a single language or dialect to infer its historical development. Unlike the comparative method, internal reconstruction does not rely on external data from related languages, making it less reliable. Instead, it relies on assumptions about which changes likely occurred over time. Despite its limitations, internal reconstruction is essential when a language is isolated or when comparative reconstruction is difficult or impossible. While comparative analysis helps linguists reconstruct a proto-language and trace its evolution through related languages, internal reconstruction provides insight into the development of a single language over time, often in cases where external data is limited or absent.

These methods, though distinct, work together in the study of language change. While comparative analysis helps linguists reconstruct a proto-language and trace its evolution through related languages, internal reconstruction provides insight into the development of a single language over time, often in cases where external data is limited or absent.

The increasing availability of digitized textual data has introduced new possibilities for historical linguistic research. In contrast to traditional methods, which often rely on manually compiled data from historical texts, corpus linguistics enables the systematic analysis of large-scale linguistic data. According to Weisser (2015, p. 13), a corpus is “any collection of texts that has been systematically assembled in order to investigate one or more linguistic phenomena.” Similarly, McEnery & Hardie (2011, p. 1) define corpus linguistics as a method “… dealing with some set of machine-readable texts which is deemed an appropriate basis on which to study a specific set of research questions”.

The essential characteristics of corpus-based analysis include its reliance on real patterns of usage found in natural texts. This method utilizes a large and systematically curated collection of texts, known as a "corpus," as the foundation for linguistic analysis. Furthermore, corpus-based analysis extensively uses computer technology, incorporating both automated and interactive techniques for data analysis. Both quantitative and qualitative methods are employed to interpret the data, providing a comprehensive view of language patterns.

Linguistic corpora can be categorized based on various criteria, including their temporal scope. One such classification differentiates between synchronic and diachronic corpora. The latter refers to “collections of texts from different periods that allow for longitudinal analysis” (Dalieva, 2024, p. 58). These corpora enable researchers to track linguistic evolution in syntax, morphology, semantics, and phonology across centuries (Dalieva, 2024, p. 58).

One of the well-known examples of a diachronic corpus is the Helsinki Corpus of English Texts. Kytö and Rissanen (1992, p. 7) describe this corpus as a digital resource hosted by the Norwegian Computing Centre for the Humanities and the Oxford Text Archive, aimed at supporting the study of the historical development of English morphology, syntax, and vocabulary. The Helsinki Corpus consists of approximately 1.5 million words and is categorized into three main sections: Old English, Middle English, and Early Modern (Southern) British English. The systematic division into century-long and shorter subperiods enables detailed chronological analysis of linguistic developments.

Another significant diachronic corpus is the Corpus of Historical American English (COHA), created by Brigham Young University. COHA is a well-organized compilation of carefully chosen English texts from a variety of sources, including newspapers, popular magazines, as well as fiction and nonfiction books, spanning the years 1810 to 2009. The corpus contains approximately 406 million words and includes around 107,000 texts, making it a valuable resource for examining linguistic changes in American English over time (Alatrash et al., 2020, p. 6959).

Corpus-based historical linguistics employs various analytical techniques to examine language change, including:

- Word frequency analysis – tracking the rise and decline of words over time.
- Collocation analysis – identifying historical shifts in word associations.
- Semantic shift analysis – examining how meanings of words evolve.
- Syntactic analysis – investigating changes in grammatical structures.
- N-gram (lexical bundle) analysis – studying patterns of multi-word expressions in historical texts (Weisser, 2015, p. 428).

These methods enable researchers to track linguistic evolution in syntax, morphology, semantics, and phonology across different historical periods.

Quantitative methods play a crucial role in diachronic corpus analysis, offering significant advantages in understanding language change. Firstly, language change is inherently complex, involving multiple layers of variation. To analyze this complexity, researchers require analytical tools that provide a nuanced and systematic view of linguistic developments. Understanding how a change occurred is the first step toward explaining why it happened. Secondly, quantitative analytical methods help uncover patterns and linguistic shifts that might otherwise remain unnoticed. By applying computational techniques, researchers can gain an objective perspective on familiar linguistic phenomena that are still not fully understood.

These quantitative techniques complement traditional corpus-based approaches by allowing researchers to systematically track changes in word frequency, collocations, syntax, and semantic shifts over time. For instance, diachronic frequency analysis enables the tracking of word usage trends across different periods, while collocation analysis identifies historical changes in word associations. Similarly, semantic shift analysis examines changes in meaning over time, and syntactic analysis investigates grammatical transformations.

Breitbarth et al. (2023, p. 2) highlight that one of the most promising developments in corpus-based historical linguistics is the application of big data analytics. The increasing availability of large-scale historical corpora, coupled with advanced statistical methods, has transformed the study of language change. Researchers can now analyze extensive datasets spanning centuries, identifying subtle linguistic trends that may have gone unnoticed in smaller, manually curated corpora.

Furthermore, big data techniques enable the visualization of language change through interactive models, such as diachronic heatmaps and evolutionary tree diagrams, offering a more intuitive representation of linguistic evolution.

Corpus-based research is further enhanced by specialized software tools designed for linguistic analysis. Programs such as WordSmith Tools, MonoConc Pro, Corpus Presenter, and Xaira allow researchers to conduct frequency-based and distributional analyses of historical corpora. Some corpora, like the Penn Parsed Corpora of Historical English and the Parsed Corpus of Early English Correspondence, come with built-in search engines tailored for linguistic research. Similarly, COHA provides an interface for investigating diachronic changes in vocabulary, grammatical structures, and collocations (Kytö, 2011, p. 428).

Beyond tracing linguistic evolution, corpus-based historical linguistics has significant implications for language preservation. Many endangered languages lack extensive historical documentation, making it difficult to reconstruct their linguistic histories. However, corpus methodologies, when applied to digitized manuscripts, inscriptions, and recorded speech data, can help linguists recover lost linguistic features and better understand the developmental trajectories of these languages. For instance, in India, the Scheme for Protection and Preservation of Endangered Languages (SPPEL), supervised by the Central Institute of Indian Languages (CIIL), focuses on documenting endangered languages spoken by fewer than 10,000 people. A key aspect of this research is the corpus-building project for the Malayan Tribe’s language in South India, aimed at collecting and documenting linguistic data to support language preservation efforts (Jenifer, 2022, p. 645).

Conclusions.

This study has examined the role of diachronic corpora in historical linguistics, highlighting their significance in tracking linguistic change across phonology, morphology, syntax, and semantics. By integrating corpus-based methodologies, such as frequency analysis, collocation analysis, and syntactic change analysis, researchers can systematically investigate language evolution with greater precision than ever before. The availability of large-scale corpora, such as the Helsinki Corpus of English

Texts and the Corpus of Historical American English, has significantly expanded the analytical possibilities within the field.

The methodological advancements brought by corpus linguistics not only enhance our understanding of past linguistic shifts but also open new avenues for exploring ongoing and future changes. As computational techniques and big data analytics continue to develop, historical linguistics stands to benefit from increasingly sophisticated tools, allowing for more nuanced and comprehensive studies of language variation and change.

References:

1. Alatrash, R., Schlechtweg, D., Kuhn, J. and Im Walde, S. S. (2020) 'CCOHA: Clean corpus of historical American English', *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 6958-6966.

2. Breitbarth, A., Ghyselen, A.-S., van Hout, R. and Wieling, M. (2023) 'Big data: New perspectives for research on language variation and change', *Taal & Tongval*, 75(1), pp. 1-6.

3. Bussmann, H., Kazzazi, K. and Trauth, G. (2006) *Routledge dictionary of language and linguistics.* Routledge.

4. Campbell, L. (2013) *Historical linguistics.* Edinburgh: Edinburgh University Press.

5. Campbell, L. and Mixco, M. J. (2007) *A glossary of historical linguistics.* Edinburgh: Edinburgh University Press.

6. Dalieva, M. (2024) 'Diachronic corpora and language evolution over time', *Web of Teachers: Inderscience Research*, 2(10), pp. 58-60.

7. Jenifer, D. (2022) 'Building corpus for endangered language: A case study of Malayan Tribe', *International Journal of Innovative Research in Technology (IJIRT)*, 9(7), pp. 645-648.

8. Kytö, M. (2011) 'Corpora and historical linguistics', *Revista Brasileira de Linguística Aplicada*, 11, pp. 417-457.

9. Kytö, M. and Rissanen, M. (1992) 'A language in transition: The Helsinki

Corpus of English texts', *Computers in English Linguistics*, 16, pp. 7-27.

10. McEnery, T. and Hardie, A. (2011) *Corpus linguistics: Method, theory and practice.* Cambridge: Cambridge University Press.

11. Ringe, D. (2005) 'Internal reconstruction', in Joseph, B. D. and Janda, R. D. (eds.) *The handbook of historical linguistics.* Oxford: Blackwell Publishing Ltd.

12. Weisser, M. (2015) *Practical corpus linguistics: An introduction to corpus-based language analysis.* Chichester: John Wiley & Sons.

Анотація. *Ця стаття досліджує інтеграцію діахронічних корпусів у історико-лінгвістичні дослідження, зосереджуючись на їхній ролі в аналізі мовних змін через час. В роботі розглядаються основні діахронічні корпуси, такі як Helsinki Corpus of English Texts та Corpus of Historical American English, а також їх застосування у фонологічному, морфологічному, синтаксичному та семантичному аналізах. Обговорюються різні методи, засновані на використанні корпусів, зокрема аналіз частотності, аналіз колокацій та синтаксичних змін, що доводять їхню ефективність у відслідковуванні еволюції мови. В статті також підкреслюються переваги кількісних підходів у історичній лінгвістиці, наголошуючи на важливості цих методів для точнішого вивчення мовних процесів. Крім того, розглядаються перспективи подальших досліджень, зокрема використання великих даних та обчислювальних інструментів, які можуть значно покращити методику аналізу.*

Ключові слова: *корпусна лінгвістика, історична лінгвістика, діахронічні корпуси, мовні зміни, порівняльно-історичний метод, корпусний аналіз, комп'ютерна лінгвістика.*

Article sent: 20.03.2025