

USING CORPORA TO STUDY NEOLOGISMS AND LEXICAL INNOVATION

Tojiboyeva Farangiz

Sherqulova Nodira

O'razaliyeva Madina

a 4th-year student of the Faculty of Philology, Jizzakh State Pedagogical University

Senior teacher of JSPU: Hakima Abdullajonova

a teacher of the Faculty of Philology, Jizzakh State Pedagogical University

Abstract: *The study of neologisms—newly coined words and expressions—has long fascinated linguists for its reflection of cultural, technological, and social change. In recent decades, the emergence of corpus linguistics has revolutionized the way neologisms and lexical innovation are identified, analyzed, and interpreted. This paper explores how corpus-based methodologies facilitate the systematic study of new words, tracing their frequency, distribution, and semantic evolution across authentic language data. By employing corpora such as the Corpus of Contemporary American English (COCA), the Oxford English Corpus (OEC), and specialized social media datasets, researchers can uncover linguistic creativity in real time. This paper reviews key theoretical perspectives on lexical innovation, outlines corpus-based tools and methods used for neologism detection, and discusses major findings from applied studies. Results indicate that corpora provide a robust empirical framework for studying neologisms, offering insights into linguistic productivity, sociocultural trends, and the dynamic relationship between technology and language. Ultimately, corpus analysis transforms the study of neology from anecdotal observation into a systematic, data-driven scientific discipline.*

Keywords: *Neologism, Lexical Innovation, Corpus Linguistics, Language Change, Computational Linguistics, Lexicography*

Language change is both inevitable and essential to the vitality of linguistic systems. New words emerge to name new objects, express evolving ideas, and reflect cultural shifts. The study of neologisms—recently coined lexical items or newly adopted senses of existing words—offers insight into the mechanisms of linguistic creativity and cultural adaptation. As language evolves under technological and social pressure, so too do the methods used to study it. Traditional lexicographic approaches relied on intuition, observation, and limited textual evidence, which often lagged behind actual usage. The

development of corpus linguistics in the latter half of the twentieth century transformed this process, enabling linguists to analyze vast collections of authentic text data.

The primary aim of this paper is to examine how corpora are used to study neologisms and lexical innovation. It investigates methodological approaches, practical applications, and theoretical implications of corpus-based neology research. The paper also considers how computational tools, frequency analysis, and contextual interpretation contribute to identifying, tracking, and understanding new words in various linguistic environments.

Neologisms are not merely linguistic curiosities—they mirror social change, technological advancement, and ideological transformation. Words such as selfie, cryptocurrency, ghosting, and metaverse exemplify how linguistic innovation accompanies cultural evolution. Corpus-based research provides the empirical foundation necessary to document these phenomena with precision, transparency, and reproducibility.

Neologisms are broadly defined as words, phrases, or expressions that are newly created or recently introduced into a language (Newman, 2019). Bauer (2001) distinguishes between lexical innovation—the creation of a new form or meaning—and lexical diffusion, the process by which innovations spread across a speech community. While some neologisms quickly enter mainstream usage and dictionaries, others remain ephemeral, limited to subcultures or digital communities.

Linguists classify neologisms into several categories, including word-formation neologisms (e.g., crowdfunding), semantic neologisms (new meanings for existing words, like cloud in computing), and loanwords borrowed from other languages (e.g., emoji from Japanese). The rate of neologism formation often correlates with social or technological shifts, illustrating language's responsiveness to human innovation. Before the rise of corpus linguistics, lexicographers and linguists identified neologisms manually, using literary texts, newspapers, or spoken records. Dictionaries such as the Oxford English Dictionary relied on editorial expertise and citation slips to track new words. This process, while meticulous, was time-consuming and subjective. Many neologisms were recorded only after gaining widespread acceptance, leaving early and transient forms undocumented (Metcalf, 2002).

The introduction of electronic corpora in the 1980s transformed linguistic research by enabling large-scale, automated text analysis. The Brown Corpus (1967), the British National Corpus (BNC), and later the Corpus of Contemporary American English (COCA) provided researchers with representative samples of real-world language, allowing for empirical investigations of emerging vocabulary.

Corpus linguistics offers a systematic approach to studying linguistic change over time. As McEnery and Hardie (2012) explain, corpora allow researchers to analyze linguistic phenomena based on frequency, collocation, and contextual distribution. In the context of neologisms, corpora provide temporal and genre-based data, enabling the detection of new lexical items and the measurement of their diffusion. Scholars such as Renouf and Banerjee (2007) pioneered corpus-based methods for tracking lexical innovation. Their “Neology Detector” algorithm compared corpora from different time periods to identify previously unseen word forms. Such approaches have been extended through web-based corpora and computational linguistics, allowing researchers to analyze linguistic change on an unprecedented scale.

The internet has accelerated lexical innovation, generating a flood of new words through social media, online gaming, and digital communication. Studies by Ziegler (2020) and Fuchs (2022) show that Twitter, Reddit, and TikTok serve as linguistic laboratories where new expressions emerge and spread. Words such as stan, doomscrolling, and finsta illustrate how online discourse fosters creativity and rapid lexical diffusion. Corpus-based analysis of these platforms reveals patterns of morphological innovation, semantic extension, and pragmatic function in digital language use.

This study references multiple well-known corpora for the analysis of neologisms, including the Corpus of Contemporary American English (COCA), the Oxford English Corpus (OEC), and web-derived corpora such as EnTenTen. These corpora were chosen due to their scale, recency, and diversity of text types. For real-time analysis, specialized social media corpora (e.g., GloWbE, Twitter datasets) provide invaluable data on emerging lexical forms.

Corpus-based neologism studies typically begin by comparing different temporal snapshots of a corpus. For example, by contrasting COCA data from 2000–2010 with data from 2011–2024, researchers can identify words that appear in the later dataset but not in earlier ones. Automated frequency lists and keyword extraction techniques help isolate candidate neologisms. Statistical filters, such as hapax legomena (words occurring only once), are used to refine results and exclude typographical errors.

After identifying potential neologisms, researchers examine their collocational behavior, semantic fields, and contextual usage through concordance lines. This allows for qualitative interpretation—determining whether a new form represents genuine innovation or a variation of an existing expression. Frequency data help distinguish between transient slang and stable lexical additions.

Tools like AntConc, Sketch Engine, and WordSmith Tools are commonly employed for corpus analysis. They provide functions such as n-gram extraction, collocation

analysis, and trend visualization. For diachronic analysis, dispersion plots illustrate how a word's frequency changes over time, revealing patterns of adoption or decline.

Corpus comparison reveals that neologisms often emerge in specific domains—technology (cryptocurrency), popular culture (stan), and politics (Brexit). These words initially appear with low frequency, often restricted to niche communities, before diffusing across media and genres. For example, selfies appeared sporadically in online forums in the early 2000s, then rose sharply in frequency between 2012 and 2014, as recorded in COCA and OEC. Corpus evidence allows researchers to quantify this diffusion and trace its semantic stabilization.

Corpus studies highlight several mechanisms of neologism formation:

- Blending: combining parts of words (brunch, manspreading, hangry).
- Compounding: joining full words (climate-smart, data-driven).
- Affixation: adding prefixes or suffixes (unfriend, microtargeting).
- Conversion: shifting word class (to Google, adulting).

These processes are evident in corpus data through morphological and syntactic patterns. For instance, the verb to friend emerged as a productive verb during the social media era, demonstrating a functional shift through widespread use.

Corpus analysis offers objectivity, scalability, and replicability. However, it faces challenges related to representativeness and data bias. Internet-based corpora may overrepresent certain demographics or linguistic styles, while automated algorithms can misclassify creative spellings or hashtags as distinct words. Moreover, identifying the boundary between innovation and error remains a persistent challenge, requiring human interpretation alongside computational methods. Corpora have transformed the study of neologisms from an anecdotal pursuit into an empirical science. By providing extensive, searchable linguistic data, they allow researchers to observe lexical innovation as it unfolds, quantify its spread, and analyze its semantic and social significance. The integration of corpus methods with computational linguistics enables large-scale, real-time monitoring of language evolution across digital media and print genres.

Neologisms reflect the adaptability of language and its capacity to capture cultural and technological progress. Through corpus-based methodologies, linguists can not only identify new words but also understand the mechanisms, motivations, and implications behind their emergence. Future research will benefit from even larger and more dynamic corpora, integrating multimodal and multilingual data to capture the complexity of lexical innovation in a globalized, digital world.

Conclusion

This study demonstrates that corpus-based methodologies deliver a scalable, data-driven framework for analyzing neologisms and broader patterns of lexical innovation. By leveraging large, diverse corpora and computational tools, researchers can operationalize the identification, tracking, and interpretation of emerging vocabulary with a level of rigor unattainable through traditional observational methods. The empirical insights generated through frequency analysis, collocational profiling, and diachronic comparison illuminate not only the structural mechanisms underlying new word formation but also the sociocultural forces propelling linguistic change.

As digital communication continues to accelerate the production and dissemination of new lexical items, corpus linguistics provides a strategic pathway for monitoring real-time language evolution across domains, platforms, and communities. Although challenges related to representativeness, data noise, and algorithmic misclassification persist, ongoing advancements in computational linguistics and corpus design are mitigating these constraints. Future research will benefit from integrating multimodal, multilingual, and continuously updated corpora, enabling a more holistic understanding of how innovation propagates through a rapidly shifting communicative landscape.



REFERENCES

1. Bauer, L. (2001). Morphological productivity. Cambridge University Press.
2. Bybee, J. (2010). Language, usage, and cognition. Cambridge University Press.
3. Fuchs, R. (2022). Tracking lexical innovation in social media corpora. *Journal of Linguistic Data Science*, 4(2), 113–135.
4. McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
5. Metcalf, A. (2002). Predicting new words: The secrets of neologism success. Houghton Mifflin.
6. Newman, P. (2019). *Neologisms and linguistic creativity*. Oxford University Press.
7. Renouf, A., & Banerjee, J. (2007). A corpus-based study of lexical innovation. *International Journal of Corpus Linguistics*, 12(3), 301–332.
8. Ziegler, J. (2020). Digital discourse and the rise of internet neologisms. *Language and Communication*, 75, 45–58.