УДК 81-114.4

Ибадова Нигяр Салех кызы
педагог
Бакинский Славянский Университет
г. Баку, Азербайджан

Ibadova Nigar Saleh
Lecturer
Baku Slavic University
Baku, Azerbaijan
ibadowa.nigar@gmail.com

КОРПУСНАЯ ЛИНГВИСТИКА: СОВРЕМЕННЫЙ ПОДХОД К ИЗУЧЕНИЮ ЯЗЫКА

В статье рассматриваются теоретические и практические аспекты корпусной лингвистики – современного направления в языкознании, основанного на систематическом сборе и анализе больших массивов текстов. Особое внимание уделяется характеристикам корпусов, функциям метаданных и их значению для научных исследований. Рассматривается опыт Азербайджана в развитии корпусной лингвистики в контексте глобализации и цифровизации языка.

Ключевые слова: корпусная лингвистика; лингвистический корпус; аннотация; метаданные; частотный анализ; азербайджанский язык.

CORPUS LINGUISTICS: A MODERN APPROACH TO LANGUAGE STUDY

This article examines the theoretical and practical aspects of corpus linguistics – a contemporary field of linguistic research based on the systematic collection and analysis of large text corpora. Special attention is paid to the characteristics of corpora, the functions of metadata, and their significance for scientific research. The article also explores Azerbaijan's experience in developing corpus linguistics in the context of globalization and digitalization of language.

Key words: corpus linguistics; linguistic corpus; annotation; metadata; frequency analysis; Azerbaijani language.

Corpus linguistics is a field of linguistics based on the systematic gathering, storage, and analysis of extensive text corpora using computer technologies. It provides researchers with a unique opportunity to study language based on real data rather than only intuitive judgments or isolated examples. Today, corpus linguistics is one of the fastest-growing fields in linguistic science.

What is a Corpus? A corpus is an electronic database of texts collected for scientific analysis. The texts may include fiction, scientific articles, newspaper publications, spoken speech, subtitles, and even social media content. Corpora are usually annotated – grammatical, semantic, and other tags are added to the words, facilitating analysis. According to Lemnitzer and Zinsmeister, a corpus is a collection of written or spoken utterances, typically stored in a machine-readable digital format.

Each text in a corpus includes not only the linguistic data itself but may also contain metadata (information about origin, genre, author, etc.) and linguistic annotation that describes grammatical, semantic, and other features of the text.

Why is Corpus Linguistics Important? Traditionally, linguistics has relied on descriptive theories and the intuition of the researcher. Corpus linguistics allows us to:

- Obtain objective data on the frequency and distribution of words and constructions;
- Study actual linguistic trends and changes;
- Analyze word and phrase usage in context;
- Create dictionaries and reference works based on real language use;
- Support language teaching and machine translation.

Lemnitzer describes corpus linguistics as a scientific discipline that deals with the description of natural language utterances and their elements and structures. These descriptions are based on the analysis of authentic (real) texts collected in corpora, and linguistic theories are derived from them. Corpus linguistics, as a scientific field, must adhere to scientific principles and standards.

Language descriptions based on corpora are used in various fields, such as language teaching, documenting endangered languages, dictionary development, and natural language processing.

Corpus Collection and Construction A corpus is a carefully assembled collection of texts or speech data. Researchers determine the corpus's goal and scope: for example, it may consist solely of scientific articles, interviews, or social media texts.

The preparation steps include:

1. **Data cleaning** – removing errors, duplicates, or non-informative elements (ads, code);
2. **Standardization** – converting texts into a unified format (encoding, fonts);
3. **Annotation** – tagging words with parts of speech (POS tags), lemmas, grammatical features.

According to Lemnitzer, linguistic corpora usually have the following characteristics:

- Representativeness;
- Metadata support;
- Linguistic annotation.

The first feature – representativeness – distinguishes corpora from other linguistic data collections. While large corpora typically have all the listed features, smaller research corpora may not include all of them.

It is particularly important to prioritize oral texts, as spoken language is the primary medium of human communication.

These criteria cover both extralinguistic and pragmatic-semantic aspects. For system-oriented linguists, it is essential that structural variation is represented in

the corpus. In well-documented languages, completeness can be ensured through targeted selection. In under-documented languages, this is impossible because the desired structures are not yet known. However, it is assumed that structural variation will naturally emerge if functional variation is adequately covered – following the principle "form follows function".

Modern corpora usually consist of full texts or dialogues – for example, newspaper articles or chat transcripts. Text length varies widely: from short messages like SMS to long texts such as novels. For many studies, wide textual contexts are required to analyze phenomena like pronoun reference or discourse structure.

The Role of Metadata Metadata serve several key purposes:

- Record contextual aspects of text creation (e.g., time, place, author);
- Allow filtering and categorization for research.

Lemnitzer provides the example: "If the films in an archive are annotated with appropriate metadata, you can find all movies where Woody Allen was involved only as a director."

He also notes two levels of metadata:

1. For the digital file itself (e.g., its name, format);
2. For the original source material (e.g., a newspaper article, diary page).

Metadata enable targeted classification, both manual and automated search, and analysis in corpora.

Quality in Corpus Compilation

Text quality involves phonetic, grammatical, and stylistic accuracy as well as meaningful content. Documentation is created for posterity, so substandard texts are neither useful nor resource-worthy. Quality also includes the compiler's work: recording quality, transcription accuracy, annotation, and metadata reliability. Quality must take precedence over quantity.

Some corpora are based on transcriptions of medieval manuscripts in different dialects and scripts. Editions may include:

- Adapted transcription close to modern norms;
- Diplomatic transcription preserving original form with minor clarifications;
- Facsimile reproductions close to the original layout.

Corpus Linguistics Methods Documenting language varieties and genres is not new. However, documenting an entire language arose in the 1980s with the realization that many languages were endangered. Full-scale documentation became a priority to preserve linguistic heritage.

Specialized tools – corpus managers and search systems – allow researchers to:

- Search word usage cases;
- Analyze collocations;
- Study grammatical structures.

Analytical methods include:

- **Frequency analysis** – how often a word or structure appears;
- **Concordance** – viewing all contexts of a word;
- **Statistical analysis** – identifying patterns and correlations.

Corpus searches often utilize morphological descriptors, which consider grammatical categories and find all relevant forms. Searches can span the entire corpus or specific context-defined subsets. This is useful in empirical material collection, dictionary development, and grammatical analysis.

Examples of Corpora:

- British National Corpus (BNC);
- Russian National Corpus (RNC);
- Corpus of Contemporary American English (COCA);
- Specialized spoken or genre-based corpora.

Data reflect phenomena from scientific domains and are accepted as evidence based on methodological rigor. In scientific reasoning, data substitute the actual phenomenon, often inaccessible directly. In inductive approaches, data serve as evidence; in deductive methods, they test hypotheses.

For centuries, traditional grammar preserved syntactic and morphological rules. Corpus-based research often yields surprising results about actual language use. Corpus linguistics is just one of many methods – it complements, but does not replace, elicitation techniques like minimal pairs. As corpora grow, their empirical value increases.

If a form does not appear in a speaker's lifetime exposure, it is inferred by analogy – the same applies to linguists using finite corpora. Larger corpora increase the reliability of such inferences.

Corpus Linguistics in Azerbaijan

In Azerbaijan, corpus linguistics was introduced through a state program to promote the national language in the context of globalization. The program included the creation of various digital dictionaries – orthographic, explanatory, terminological, and frequency-based.

Building a national corpus is a strategic state priority. It helps preserve the language, monitor lexical changes over time, and facilitates research without manual data collection. Researchers can extract and analyze linguistic units in natural contexts.

Compiling frequency dictionaries and concordances for classic Azerbaijani authors – Nizami Ganjavi, Mirza Alakbar Sabir, and Muhammad Fuzuli – is a key step toward creating individual corpora for these poets. These projects serve as a foundation for the Azerbaijani National Corpus.

A comprehensive national corpus must represent all functional styles and genres, including classical literature, folklore, and oral traditions. Frequency dictionaries are also critical for machine translation, summarization, authorship attribution, publishing, and forensic linguistics.

Conclusion

Corpus linguistics is a powerful tool for modern language research, enabling deep, objective insights into language structure and use. As technology and data availability grow, corpus methods gain increasing relevance in research, education, translation, and linguistic technology development.

REFERENCES

1. Lemnitzer, L., & Zinsmeister, H. (2010). *Korpuslinguistik: Eine Einführung*. Tübingen: Narr Francke Attempto Verlag.
2. Rayskina, V. A., & Dubnyakova, O. A. (n.d.). *Современные методы корпусной лингвистики при анализе текста (на примере корпуса BFM)*. Retrieved from https://cyberleninka.ru/article/n/sovremennye-metody-korpusnoy-lingvistiki-pri-analize-teksta-na-primere-korpusa-bfm/viewer
3. Lehmann, C. (n.d.). *Daten–Korpora–Dokumentation*. Retrieved from https://www.christianlehmann.eu/publ/daten.pdf
4. Khalilova, G. A. (n.d.). *Развитие азербайджанской корпусной лингвистики. Создание корпуса текстов азербайджанских авторов (поэтов и писателей) и национального корпуса азербайджанского языка*. Retrieved from file:///C:/Users/Администратор/Downloads/razvitie-azerbaydzhanskoy-korpusnoy-lingvistiki-sozdanie-korpusa-tekstov-azerbaydzhanskih-avtorov-poetov-i-pisateley-i-natsionalnogo-korpusa-azerbaydzhanskogo-yazyka.pdf
5. Azerbaijani National Corpus. (n.d.). Retrieved from https://korpus.azerbaycandili.a