

TYPES OF CORPORA: GENERAL, SPECIALIZED, AND PARALLEL

Abdullajonova Hakima

Senior teacher of JSPU

Abduraxmonova Iroda

Jizzakh State Pedagogical University

Foreign Languages

Annotation. *This article explores the main types of corpora used in linguistic research: general, specialized, and parallel corpora. It explains their structural features, typical applications, and importance in contemporary language studies. General corpora provide broad, representative samples of language use, while specialized corpora focus on specific domains, genres, or disciplines. Parallel corpora, on the other hand, allow cross-linguistic comparison by aligning texts across languages. The article highlights how these corpora contribute to language teaching, translation studies, lexicography, natural language processing, and linguistic analysis.*

Keywords. *corpus, general corpus, specialized corpus, parallel corpus, linguistics, lexicography, translation studies, NLP, language teaching, language analysis*

Annotatsiya. *Ushbu maqolada tilshunoslikda qo'llaniladigan asosiy korpus turlari: umumiy, maxsus va parallel korpuslar yoritiladi. Unda ularning tuzilishi, qo'llanish sohalari hamda zamonaviy til tadqiqotlaridagi ahamiyati tushuntiriladi. Umumiy korpuslar tilning keng va vakillik namunalarini taqdim etsa, maxsus korpuslar ma'lum soha, janr yoki fanlarga yo'naltiriladi. Parallel korpuslar esa matnlarni turli tillar bo'yicha moslashtirish orqali qiyosiy tadqiqotlar olib borish imkonini beradi. Maqola korpuslarning til o'qitish, tarjima, leksikografiya va NLPdagi o'rnini ko'rsatadi.*

Kalit so'zlar. *korpus, umumiy korpus, maxsus korpus, parallel korpus, tilshunoslik, leksikografiya, tarjimashunoslik, NLP, til o'qitish, til tahlili*

Аннотация. *В статье рассматриваются основные типы корпусов, используемых в лингвистических исследованиях: общий, специализированный и параллельный корпуса. Объясняются их структура, сферы применения и значение в современной лингвистике. Общие корпуса предоставляют широкие, репрезентативные образцы языка, тогда как специализированные корпуса ориентированы на конкретные области, жанры или дисциплины. Параллельные корпуса позволяют проводить межъязыковое сопоставление благодаря выравниванию текстов на разных языках. В статье подчеркивается вклад*

корпусов в преподавание языка, переводоведение, лексикографию и обработку естественного языка.

Ключевые слова. *корпус, общий корпус, специализированный корпус, параллельный корпус, лингвистика, лексикография, переводоведение, NLP, обучение языку, языковой анализ*

A linguistic corpus is a structured and systematic collection of written or spoken texts that represents actual language use. Corpora have become essential tools in modern linguistics, enabling researchers to examine patterns, frequency, context, and variation in natural language. Among the most commonly used types of corpora are general corpora, specialized corpora, and parallel corpora. Each type serves unique academic, pedagogical, and technological purposes, contributing to the advancement of linguistics and language-related fields.

General corpora are broad collections of texts designed to represent a language as it is used across different genres, contexts, and registers. Their goal is to provide a balanced and representative sample of the entire language. Well-known examples include the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA). These corpora contain millions of words from newspapers, fiction, academic writing, spoken language, and online sources. Because of their diversity, general corpora are ideal for identifying trends in grammar, vocabulary, discourse, and stylistic features. They are widely used in lexicography, where dictionary makers rely on corpus evidence to determine definitions, usage examples, collocations, and typical contexts. Moreover, language teachers and researchers use general corpora to understand authentic language usage and develop evidence-based teaching materials.

Specialized corpora, unlike general corpora, focus on a particular field, domain, or genre. These corpora may consist of legal documents, medical research papers, business communication, learner essays, or even a single author's literary works. The purpose of specialized corpora is to analyze language within a specific context and to identify features that may not appear prominently in general corpora. For example, a medical corpus would contain terminology, phraseology, and rhetorical structures unique to medical discourse. Specialized corpora are also common in English for Specific Purposes (ESP), where teachers need data-driven insights about how language operates in technical contexts. Learner corpora, a subset of specialized corpora, allow researchers to track common language errors, developmental stages, and influences of learners' first languages. These insights support the creation of targeted teaching strategies and improved learning materials.

Parallel corpora represent another crucial category: they contain texts in two or more languages that are aligned sentence by sentence or paragraph by paragraph. These corpora enable researchers to study translation shifts, equivalence, cultural adaptations, and cross-linguistic structures. Parallel corpora support translation studies by offering concrete examples of how certain phrases or grammar structures are translated in different contexts. They also play a major role in the development of machine translation systems and multilingual Natural Language Processing (NLP). For instance, modern neural translation models require large-scale parallel corpora to learn patterns of correspondence between languages. Government institutions, the European Union, and international organizations frequently publish multilingual documents, which serve as valuable sources for constructing parallel corpora.

The importance of corpora extends beyond theoretical linguistics. In language teaching, corpora provide empirical data for designing syllabi, exercises, and authentic language input. Teachers can use concordance tools to demonstrate real-life usage of vocabulary or grammar. Students benefit by exploring patterns independently, fostering data-driven learning skills. In translation and interpreting, corpora help professionals maintain consistency, select appropriate terminology, and understand stylistic tendencies in various genres. In technology, corpora fuel NLP tasks such as speech recognition, text classification, sentiment analysis, and information retrieval. Without corpora, developing accurate AI language models would be nearly impossible.

Each type of corpus contributes differently: general corpora provide breadth, specialized corpora provide depth, and parallel corpora provide cross-linguistic insight. Together, they form the backbone of contemporary language analysis and technological innovation. Their continued development ensures the growth of linguistic research, improved language education, and more sophisticated language technologies.

Conclusion.

General, specialized, and parallel corpora each play vital roles in modern linguistics. General corpora capture broad language patterns, specialized corpora reveal domain-specific features, and parallel corpora enable multilingual comparison and translation analysis. Their combined value supports research, teaching, translation, lexicography, and NLP. As the digital world expands, corpora will continue to grow in size and diversity, strengthening empirical approaches to language and enhancing both academic and technological advancements.

REFERENCES

1. McEnery, T. & Hardie, A. *Corpus Linguistics: Method, Theory and Practice*. 2012.

2. Sinclair, J. Corpus, Concordance, Collocation. 1991.
3. Biber, D. Variation Across Speech and Writing. 1995.
4. Baker, M. Corpora in Translation Studies. 1993.
5. Kennedy, G. An Introduction to Corpus Linguistics. 1998.
6. Granger, S. Learner English on Computer. 1998.
7. O'Keeffe, A., McCarthy, M. The Routledge Handbook of Corpus Linguistics. 2010.
8. <https://www.sketchengine.eu>
9. <https://corpus.byu.edu>
10. <https://www.linguisticsociety.org/resource/corpora-linguistics>

