СОВЕРШЕНСТВОВАНИЕ МЕТОДИКИ ОБУЧЕНИЯ ЯЗЫКАМ: ПЛОЩАДКА ОБМЕНА ПРОГРЕССИВНОЙ ПРАКТИКОЙ

КАЗАНСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ
ИНСТИТУТ ФИЛОЛОГИИ И МЕЖКУЛЬТУРНОЙ КОММУНИКАЦИИ
НОЦ «Институт Каюма Насыри»

СОВЕРШЕНСТВОВАНИЕ МЕТОДИКИ ОБУЧЕНИЯ ЯЗЫКАМ
ПЛОЩАДКА ОБМЕНА ПРОГРЕССИВНОЙ ПРАКТИКОЙ

ТЕЛЛӘРГӘ ӨЙРӘТҮ МЕТОДИКАСЫН КАМИЛЛӘШТЕРҮ
АЛДЫНГЫ ТӘҖРИБӘ БЕЛӘН УРТАКЛАШУ МӘЙДАНЧЫГЫ

МАТЕРИАЛЫ VII МЕЖДУНАРОДНОГО НАУЧНО-МЕТОДИЧЕСКОГО ОНЛАЙН-СЕМИНАРА

КАЗАНЬ
2023

ON THE USE OF CORPORALS IN LINGUISTIC RESEARCH

Avezov S.S.
Bukhara state university
1990senigama@gmail.com

Аннотация. В этой статье рассматривается вопрос использования текстовых коллекций в лингвистических исследованиях. В ней объясняются основные принципы корпусной лингвистики, различные типы текстовых коллекций и методы их создания. Наконец, исследуется потенциал корпусной лингвистики в сравнительных исследованиях.

Ключевые слова: методика лингвистического исследования, корпусная лингвистика, параллельный корпус текстов, репрезентативность.

The way in which linguistic data is collected and analyzed has long been a central concern in the field of linguistics. We believe that currently, there is a tension between older, more traditional methods and newer, emerging approaches to collecting linguistic data.

To illustrate the characteristics of traditional methods, the text is quoting E.V. Paducheva, who states that linguistic data is often collected from dictionaries and other linguistic literature, and examples are sometimes taken from a card file of the Dictionary of the Russian Language at the Leningrad Branch of the Institute of Linguistics of the USSR Academy of Sciences. The source of the example is only cited if it is of particular interest or if the correctness of the sentence is in question. Additionally, artificial examples are also frequently used[Paducheva: 228]. The traditional approach relies on the intuition of native speakers rather than working with actual texts. The primary role of a linguist is to explain the rules that native speakers intuitively understand, and to translate these nebulous concepts into more formal and logical forms. This is achieved by methodically observing and reflecting on language use. This approach is evident in many recent linguistic works, where empirical data is only used to verify hypotheses, and examples used are often random and sporadic. It appears that there is not a deliberate disregard for empirical data, but rather a lack of emphasis on it, which can lead to a breakdown in the logical process of data collection, hypothesis formation, verification, and theory development. It's also noted that the actual theory of knowledge is about revealing and describing natural formations that exist independently and the observation of them allows us to formulate laws as necessary relations arising from the nature of things.

Corpus linguistics, which has seen significant growth in recent decades, allows for a more comprehensive and objective understanding of language phenomena by using large, electronically stored, structured, and annotated collections of linguistic data. These corpora can be used to test linguistic hypotheses and theories and can also be used as a source of examples for difficult language phenomena. The use of a large, representative corpus ensures that the data is typical and comprehensive. The corpus allows for the study of data in its natural context, which is not possible with traditional methods such as introspection, questionnaires, or interviews with informants. The corpus enables the collection of data that is not available with

traditional methods and the generalizations drawn have the status of an empirically observed fact, rather than an introspective guess.

Corpus Linguistics, being a relatively new field, is characterized by some terminological confusion. The use of terms such as «language corpus» and «linguistic corpus» is also debated. A.A. Polikarpov argues against the use of the latter term, as in his opinion, a linguistic corpus is a type of corpus related to the study of language, rather than language itself. He argues that a qualified linguist would not make this mistake in using the term «linguistic». It's also crucial to make a distinction between corpus linguistics as a theory and as a method. Computational linguistics as a theory is a branch of linguistics that develops general principles for creating and utilizing linguistic corpora using computer technology. When using corpora as a reliable source of data on the phonetic, morphological, syntactic, and semantic structure of a language, we are referring more to the corpus approach as a method of linguistic research. In this case, representativeness is a particularly important characteristic of the corpus[Polikarpov: 114].

A corpus is a simplified version of a language or sublanguage, and its representativeness affects the accuracy of the data obtained from it. Therefore, the question of corpus representativeness can also be seen as the problem of properly selecting, adapting, and integrating large amounts of text into a smaller corpus. Representativeness is not just about the quantity of material, but also the proportionality of the representation of the language or sublanguage being studied. Increasing the size of the corpus does not necessarily improve its reliability; instead, careful selection of texts during the planning and use of the corpus is more important. This leads to the problem of classifying different types of corpora.

There are two main types of corpora that are distinguished based on the criteria of representativeness and selection of texts:

1. Corpora that pertain to the entire language
2. Deliberately chosen corpora, related to a specific sublanguage such as genre, style, or language used by a particular social group, as described by W.E. Francis[Francis: 334-352].

The first type of corpora are built using the principle of deduction, which refers to moving from a general corpus of texts to a specific corpus that reflects this general corpus. These corpora are universal in nature and aim to reflect the entire range of speech activity, regardless of the researcher. Such corpora are available to the public, either in whole or in part, through the internet. Some well-known examples of these traditional corpora include the British National Corpus, which contains around 100 million word usages, and the Mannheim Corpus of the German language, which contains around 1 billion word usages. Recently, the corpus of modern German created by the University of Leipzig has also gained popularity among German language scholars. In Russia, the creation of corpora has been identified as an important task in computational linguistics. The National Corpus of the Russian Language, which can be found at http://ruscorpora.ru, is currently in operation. Additionally, work is underway to create a representative national corpus of the

Russian language, called the Large Corpus of the Russian Language (BoKR), which will have a volume of at least 100 million word forms.

Second-class corpora are created with the purpose of representing a specific linguistic or cultural phenomenon. The selection of texts for the corpus is determined by the corpus creator and is based on their goals for practical or scientific research. The methodologies used for constructing these corpora are inductive, focusing on the accuracy of the representation of the chosen phenomenon within the corpus. Examples of Russian-language corpora of this type include the Computer Corpus of Russian Newspaper Texts of the Late 20th Century and the Corpus of Political Metaphors.

There are various types of hulls that can be classified based on the material they are made of, their structure, and their intended use. For instance, data type can be used as a classification attribute to differentiate between written, speech, and mixed corpora. Additionally, parallelism can be used to distinguish between monolingual, bilingual, and multilingual corpora. From the perspective of linguists, the most important classification criteria include research, illustrative, static, dynamic types of corpora, as well as parallel text corpora.

To conclude, corpus linguistics has a great potential in contrastive studies. One of the most promising areas is the development of parallel corpora of texts in different languages, which consist of original texts and their translations. These parallel corpora not only have the advantages of monolingual corpora in studying a single language, but also provide optimal conditions for researching the problems of conveying different linguistic meanings across languages, and finding equivalents in translation practice. The need for this type of research has been highlighted by V. G. Gak, who stated that «comparing translations with the original, we often find such lexical substitutions that are not covered by dictionaries and cannot be explained with their help,» and that «speech parallels can only be identified through linguistic experiments or by comparing translations»[Gak: 264].

Working with electronic corpora can offer new opportunities and improve the level of objectivity in linguistic research. However, it's important to remember that when using a corpus for lexical analysis, it is not possible to completely capture the entire vocabulary of a language. The lexicon of a language is vast and almost infinite, with many possible combinations that cannot be fully represented in a corpus. Additionally, the lexicon is an open system, meaning that no matter how much a corpus is expanded, there will always be words that are not yet represented in it.

References

1. Gak VG Comparative lexicology // International relations. – Moscow, - 1977. – P. 264.
2. Francis W. E. Problems of formation and machine representation of a large body of texts // New in foreign linguistics. Issue XIV // Problems and methods of lexicography. Progress, – Moscow, - 1983. – P. 334-352.
3. Paducheva E. V. Statement and its correlation with reality // Referential aspects of the semantics of pronouns // Editorial URSS. – Moscow, - 2004. – P. 228.
4. Polikarpov A.A. About one review. Access mode: //http://www.linguide.com.ua/ content. (accessed 17.01.2023). – P. 114.