www.volsu.ru

DOI: https://doi.org/10.15688/jvolsu2.2025.1.8

UDC 81’33
LBC 81.1

Submitted: 28.04.2024
Accepted: 21.10.2024

THE ROLE OF CORPUS LINGUISTICS IN CONTEMPORARY LINGUISTICS RESEARCH AND TRANSLATION STUDIES [1]

Pei Haitong

Heilongjiang University, Harbin, China

Abstract. The article presents a systematic review of research papers on corpus linguistics as an innovative direction in empirical linguistics. It reveals the theoretical and methodological foundations of the field, defines the specifics of corpus linguistics in comparison with computational linguistics, and highlights modern trends in corpus research as well as the advantages of employing textual corpus data in linguistic studies. It is noted that the corpus method and its resources are actively used in various linguistic researches into many world languages today; that language corpora are differentiated by volume, type, structure, content, purpose, etc. The article provides an overview of such corpora as the British National Corpus (BNC), the American National Corpus (ANC), the Corpus of Contemporary American English (COCA), the Russian National Corpus (RNC), and the Modern Chinese Language Corpus created in the Center for Chinese Linguistics at Beijing University, the Balanced Corpus of the Chinese Language. The specifics of parallel and comparative corpora are noted, including the Child Language Data Exchange System (CHILDES), the International Comparable Corpus (ICC), and the Corpus of English Wikipedia (CEW). The differences between parallel and comparative corpora are also outlined. Prospects for national corpora development lie in new research into almost every area of applied and theoretical linguistics, as well as in scrutiny and further development of translation theory and practice. The characteristics of monolingual, comparative, and parallel corpora are highlighted in the context of their role in linguistic research. It is mentioned that, in addition to parallel corpora, translators’ tools also include monolingual corpora, which provide additional material on the subject of translation and enhance the translator’s background knowledge.

Key words: applied linguistics, corpus linguistics, corpus, national corpus, parallel corpus, comparative corpus, corpus method.

Citation. Pei Haitong. The Role of Corpus Linguistics in Contemporary Linguistics Research and Translation Studies. *Vestnik Volgogradskogo gosudarstvennogo universiteta. Seriya 2. Yazykoznanie* [Science Journal of Volgograd State University. Linguistics], 2025, vol. 24, no. 1, pp. 95-106. DOI: https://doi.org/10.15688/jvolsu2.2025.1.8

УДК 81’33
ББК 81.1

Дата поступления статьи: 28.04.2024
Дата принятия статьи: 21.10.2024

РОЛЬ КОРПУСНОГО ЯЗЫКОЗНАНИЯ В СОВРЕМЕННЫХ ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЯХ И ПЕРЕВОДОВЕДЕНИИ [1]

Пэй Хайтун

Хэйлунцзянский университет, г. Харбин, Китай

Аннотация. Статья представляет собой систематизированный обзор работ по корпусной лингвистике как инновационному направлению эмпирического языкознания. Раскрыты теоретико-методологические основы направления, определена специфика корпусной лингвистики по сравнению с компьютерной, обозначены современные направления корпусных исследований и преимущества привлечения данных текстового корпуса в языковедческих работах. Отмечено, что корпусный метод и его ресурсы сегодня активно используют в различных лингвистических исследованиях многих мировых языков; что языковые корпусы различают по объему, типу, структуре, наполнению, назначению и др. В статье обзорно рассматриваются

такие корпусы, как Британский национальный корпус (The British National Corpus), Американский национальный корпус (The American National Corpus), Корпус современного американского английского языка (COCA), Национальный корпус русского языка, Корпус современного китайского языка (The Modern Chinese Language Corpus), созданный в Центре китайской лингвистики при Пекинском университете, Сбалансированный корпус китайского языка. Выявлена специфика параллельных и сравнительных корпусов, среди которых CHILDES – корпус детской речи, Международный сравнительный корпус (The International Comparable Corpus), Корпус англоязычной Википедии (Corpus of English Wikipedia), приводятся различия между параллельными и сравнительными корпусами. Установлено, что перспективы развития национальных корпусов связаны с новыми исследованиями практически в каждой области прикладной и теоретической лингвистики, с дальнейшей разработкой и углублением теории и практики перевода. Выделены особенности одноязычных, сравнительных и параллельных корпусов в контексте их роли в лингвистических исследованиях. Показано, что в инструментарий переводчика, помимо параллельных корпусов, входят и одноязычные, дающие дополнительный материал о предмете перевода и актуализирующие фоновые знания переводчика.

Ключевые слова: прикладная лингвистика, корпусная лингвистика, корпус, национальный корпус, параллельный корпус, сравнительный корпус, корпусный метод.

Цитирование. Пэй Хайтун. Роль корпусного языкознания в современных лингвистических исследованиях и переводоведении // Вестник Волгоградского государственного университета. Серия 2, Языкознание. – 2025. – Т. 24, № 1. – С. 95–106. – (На англ яз.). – DOI: https://doi.org/10.15688/jvolsu2.2025.1.8

Introduction

The development and use of electronic text corpora at the present stage is one of the most promising areas in modern linguistics. Within the framework of this particular direction, it is possible to predict achievements both in the field of theoretical linguistics (gaining new knowledge about the structure of language) and in the field of applied linguistics (acquiring new generation technologies for automatic text processing and accelerated modernization of linguistic research methods). This is due to the fact that only an electronic corpus makes possible the collection of results in real time that require processing an array quantity of text that an ordinary researcher is not able to cope with. Obtaining the same data manually, e.g. by simply reviewing texts and writing down examples, can take a very long time.

The main goal of corpus linguistics is understanding the rules governing language as a whole, or a specific aspect of language, and allowing for every aspect of language to be studied. Furthermore, electronic corpora provide opportunities for the analysis of linguistic phenomena that are not easily accessible through traditional methods, such as rare or colloquial expressions, dialects, and spoken language.

In addition, an electronic text corpus not only makes it possible to speed up the process of language research and increase its efficiency, probability and verification, but it also helps to solve tasks that linguistics of previous eras did not even raise because of their complexity or impracticability. These include, for example, various types of statistical and other quantitative studies of language, which were partially carried out in the pre-corpus era, but have been actively developing recently.

Corpus linguistics methods are used to track changes in the history of language. As a result, previously considered insoluble problems can now be more easily solved. Each language is in the process of stable but slow changes, the results of which for the most part become noticeable on a scale of several centuries. Understanding the mechanisms of such changes can provide new knowledge about the essence of natural language as a whole. However, research in this area will be most effective when using data from historical, or diachronic, corpora containing texts created over a long period of time, usually no less than five to seven centuries. The result of this diachronic expansion is the creation of general language corpora of various languages. Thus, national corpora of English (British and American variants), many Slavic languages (in particular, Russian, Polish, Czech, Croatian) and the like have already been created. Such corpora usually contain hundreds of millions of word usages or other linguistic units. For example, the Corpus of Historical American English (COHA) is a diachronic corpus that is often used for diachronic research on the English language. Corpus linguistics also provides tools for documenting endangered languages and creating resources that

can be successfully used in lexicography, and other areas of linguistics, such as translation, language teaching, etc.

Materials and methods

The research methods are description and comparative analysis with the aim of characterizing a number of national and parallel corpora and classification when identifying types of corpora. The descriptive method has been applied to identify the characteristics of national corpora, while the comparative method has been used to identify similarities and differences in approaches to compiling a particular national or parallel corpus. The source base of the article consists of a number of national corpora (English, American, Russian, Chinese), as well as parallel corpora (Russian-Chinese Corpus of NRU, Corpus of Russian-Chinese Translations).

Discussion

From the point of view of modern theoretical and descriptive linguistics, a corpus is not only a special tool for studying a language, but also a necessary component of its integral description. To the classic "dictionary-grammar" pair, conceptualised by F. de Saussure, modern science has added a third element – the corpus. In this case, a complete description of the language is understood as including the corpus and the dictionary and grammar built on its basis. Based on this, it is possible to talk about "corpus dictionaries" and "corpus grammars" of a new generation, created and verified in relation to a specific fixed corpus. The corpus-based nature of dictionaries and grammars increases their reliability and verifiability and avoids the subjectivity and incompleteness that can characterize traditional descriptions.

It is obvious that the use of electronic text corpora data is gradually becoming an inseparable component of linguistic theory and practice. This explains the relevance of questions about the specifics and prospects for using text corpora to solve a wide range of problems in various fields of linguistics, as well as in related sciences.

An analysis of recent studies and publications indicates a constant research interest in the problem of corpus linguistics and its technologies. Scientists believe that the genesis of corpus linguistics is the awareness of an objective linguistic concept – "the dichotomy language – speech". The concept according to which language and speech are considered different real objects that interact in the process of speech activity has won and continues to win more and more supporters. Based on this approach, language has only speech potential as well as various possibilities for discourse formation. At the same time, linguists consider the possibilities of discourse as a person's adaptation to the surrounding linguistic situation. Some difficulties at the current stage of development of corpus linguistics are caused by the lack of a unified approach to the terms used in it.

Today we can already talk about a rather diverse scientific heritage of modern linguists on the problems of corpus linguistics. Critical understanding, systematization and popularization of these developments will contribute to the productive mastery of the capabilities of text corpora by the linguistic community.

At first glance, theoretical studies of corpus linguistics have less productive results compared to the successful analytical products obtained in applied research. But theoretical problems developed on the basis of text corpora make it possible to present new knowledge and facts about the structure of language in their results. Thus, scientists have repeatedly noted that many grammatical structures and phenomena are discovered only when working with text corpora, and the study of typological phenomena in linguistics, grammatical (syntactic) analysis along with lexical analysis, is the most common type of research for which corpora are used.

Upon the first inspection, the differences in the theoretical and applied directions of corpus linguistics should be significant and even opposite to each other. However, in reality this is not the case, since corpus linguistics is based on a certain understanding that language is a completely social phenomenon, it can be described by data based on experience, that is, on speech acts. Everyone who speaks and writes a language necessarily adapts to social circumstances. Therefore, the difference between these directions (theoretical, descriptive and applied) does not mean their isolation. Here we can talk about the triune

orientation of all corpus linguistics. Moreover, applied research itself always contains theoretical elements.

Consequently, corpus linguistics is at the intersection of theoretical and applied problems. We can say that theoretical research prepares solutions for practical problems and forms their basis.

In corpus linguistics the key concept is the corpus, and it is still defined differently by researchers. E. Finegan calls a corpus a representative collection of texts presented in a machine-readable format, including information about the situation in which the text was created (in particular, information about the author, addressee, audience) [Finegan, 2015]. T.E. McEnery, A. Hardie believe that a corpus is a collection of language fragments selected in accordance with clear language criteria for use as a language model [McEnery, Hardie, 2011]. In Russian corpus linguistics, one of the first to define a corpus was V.P. Zakharov, characterizing it as a large, electronically presented, structured and labeled, philologically representative array of linguistic data designed to solve certain linguistic problems [Zakharov, 2011].

The presence of quite diverse definitions of corpus concepts, however, shows that they do not contradict each other, but simply provide a more detailed, complete description of it. Thus, a linguistic corpus of texts can be understood as a large-volume, unified, structured, marked-up and philologically competent array of texts in natural language, presented in digital format, supplemented by a management system – universal software tools for searching and processing various linguistic information.

Technologies for building the corpus are generally uniform. The corpus is compiled on the basis of a large representative group of texts, processed in such a way that the language material is arranged according to the principle of concordance. When compiling a corpus, each text is labeled and divided into fragments, illustrated units and grouped into a selection of contexts. Unlike ordinary texts, characterized by a linear, horizontal organization, genre and style specificity, authorial individuality, structural and semantic integrity, and the presence of a specific communicative goal, the corpus does not have a general communicative goal (the communicative goal can be traced only at the level of one sentence). The contexts in the corpus, presented in the concordance, certify the functioning of a speech unit in various styles and areas of use, giving a broad picture of real language practice. Thus, the use of a corpus approach provides the opportunity to study any linguistic units in various speech genres, and also allows us to determine the specifics of the use of these units in various types of discourse.

In essence, the corpus approach is a technique or set of techniques, rather than a theory for describing language. Most often, linguistic corpora are used in morphological, lexicological and syntactic studies. Corpus databases allow searching and statistically counting the use of given roots, affixes and inflections, thereby studying ways to create a linguistic unit. Linguistic corpora make it possible to obtain data on specific word forms and entire grammatical categories. For example, linguists are conducting a systematic-structural study of verb morphology based on the corpus (N.V. Buntman, A.S. Borisova, Yu.A. Darovskikh, Yu.A. Sukhorukova, and others).

Linguistic corpora provide material for studying the frequency of language units in general, as well as in their immediate functioning in texts of various styles. For instance, on the base of Longman Spoken and Written English Corpus (LSWE), American linguist Douglas Biber identified the 12 most common verbs in the English language: *say*, *get*, *go*, *know*, *think*, *see*, *make*, *come*, *take*, *want*, *give*, and *mean* [Biber, 2001, p. 104].

The corpus-based approach is also actively applied to research the specifics of the structure and usage of various types of grammatical constructions. Research on the verbal category of aspectuality is well-known. For example, D. Biber notes that the simple form of the verb in English is used 20 times more often than the progressive or continuous form; however, there are several verbs that primarily occur in the continuous form, namely: *bleeding*, *chasing*, *shopping*, *starving*, *joking*, *kidding*, and *moaning* [Biber, 2001, p. 106].

The methodology of corpus studies is also applied to discourse analysis. The foundation for this approach was laid by J. Sinclair and theoretically justified by P. Baker [Baker, 2006]. The practical application of corpus linguistics

methods for analyzing discursive phenomena can be noted, for example, in the work of E. Semino and M. Short, which is dedicated to ways of representing speech and thought in English texts [Semino, Short, 2004].

The main condition for a corpus to become a useful tool is its relevance to the research problem. Moreover, it should have a maximum deep markup and extensive annotation, to provide the researcher or translator with all the necessary data to achieve the set goals. Depending on the targets set by the research, corpora can be built based on the following indicators and parameters, including, but not limited to: corpora of complete texts or fragments; research corpora: illustrative or interpretive; written, spoken, or mixed corpora; monolingual or multilingual corpora; synchronous or diachronic corpora; static or dynamic corpora; balanced or monitoring corpora; parallel or comparative corpora; small, medium, or large corpora, etc.

Descriptions of some versions of monolingual and multilingual corpora, parallel and comparative corpora are presented in the following parts of the article.

Typically, the corpus analysis procedure involves three main steps: identification of language data using categorical analysis, correlation of language data using statistical methods, and intelligent interpretation of the results. Since corpus research is based primarily on an empirical approach to the analysis of speech material, this makes it possible to achieve maximum objectivity in language learning, making the subjective views of the researcher impossible. They provide a unique language learning tool, thanks to which you can search in large bodies of text, obtain data on linguistic units and phenomena at various language levels (phonetic, morphological, lexical-semantic and syntactic), examine the frequency of word forms, lexemes, grammatical categories, syntactic constructions, identify atypical grammatical phenomena and constructions, establish the immediate lexical and grammatical environment of the word, with the help of which you can analyze the use of the word in all its characteristic collocations, colligations and syntaxes.

The principles of annotation are uniform for different corpora and represent a specific algorithm, which is described below within the review of corpora and their characteristics.

Primary automatic markup of text for further processing within the corpus involves the following basic operations: text segmentation (definition of paragraphs and sentences), tokenization, morphological analysis: lemmatization and determination of grammatical categories. Tokenization refers to the process of segmenting text, which is a sequence of characters such as letters, spaces, punctuation marks and numbers, into words and phrases [Teodorescu, 2017, p. 5].

The task of tokenization is to separate words from syntactic signs, numbers, complexes of letters and numbers, internet addresses, nicknames, signs, etc. At the same time, the creation of a perfect algorithm for machine identification and delimitation of tokens is an unresolved issue due to the presence in the texts of a large number of units and combinations, the unambiguous automatic classification of which is impossible at this stage, for example: multi-word tokens; names containing signs; numbers containing spaces; the presence in the text of a period that is not a sign of the end of a sentence (direct speech, abbreviations), punctuation marks containing more than one character (for example, three dots), etc. At the present stage, these and many other cases cannot be classified automatically, such procedures are performed automatically. Solving these issues will make it possible to achieve an appropriate level of primary automatic processing of text material and subsequent high-quality recognition of the meaning of individual units and the entire text within a certain corpus.

The use of a text corpus provides the opportunity to simultaneously process a large amount of textual information selected according to certain criteria. Thus, the user can analyze language material according to the following criteria: chronological parameters (time of creation of the text, time of translation), frequency of use of each translation option in texts of different genres, etc. It is clear that the effectiveness of using the corpus when translating any language units directly depends on the representativeness of the corpus itself.

Research in recent years has led to the development of numerous algorithms and approaches for creating computer dictionaries (V.M. Andryushchenko, L.N. Belyaeva,

A.S. Herd, I.I. Ubin, and others). However, some features and characteristics of language units still cannot be formalized due to lexical and grammatical polysemy, extralinguistic specifics of individual units and entire texts, as well as the ongoing development and variability of language. Because of this issue, the development of universal algorithms and standards, as well as the creation of perfect text corpora, remains relevant and requires further study and exploration of solutions.

Results

From general statements, we will turn to several characteristics of monolingual corpora of three languages: English, Russian, and Chinese, that are viewed as a tool for representing linguistic material with the aim of highlighting the specific features of their structure, then we are planning to state functional values of comparative and parallel corpora as significant tools for translators.

Monolingual corpora as a tool for processing speech material

The long history of the development of the corpus approach to the systematization of linguistic units in Europe and the United States of America has ensured the emergence and rapid development of English language corpora. Today, the most representative and authoritative corpora of the English language are the British National Corpus (BNC) and the American National Corpus (ANC).

One of the best-known and most representative corpora of modern English (its British version) is the British National Corpus. Its main characteristics are easy and understandable access, the presence of a large collection of texts of different genres and a well-developed annotation, which together make it a convenient research tool and a source of reliable linguistic information.

The corpus usage is convenient. When users enter the site, they get access to a window with four main functions: SEARCH/SEARCH, FREQUENCY/FREQUENCY, CONTEXT/CONTEXT (left and right combination of a lexical item with different levels of depth) and OVERVIEW/GENERAL INFORMATION. The search enables five options: List/list, Chart/diagram, Collocates/collocations, Compare/compare, KWIC (Keyword in Context)/keyword in context. These parameters are associated with a specific display of data on the screen, depending on the criteria that interest the user. The List function allows you to search for contexts with individual words.

Using this function, you can get a list of text fragments with the word you are searching for. You can also obtain data on the frequency of use of a unit, and the Chart function provides information about the overall frequency for a specific genre of the corpus. The results can be presented in tables with horizontal or vertical orientation.

Collocates function allows you to search for a language unit with a given combination to the left and/or right of the unit being analyzed and provides results sorted by the amount of use of each unit.

When using the Compare function, you can obtain information about the functioning of the two units being compared, indicating the frequency of use for each unit compared.

KWIC function (Keyword in Context) allows contexts of varying depth to be tabulated by specifying the number of units to the left and/and to the right of the unit of interest.

In addition to these features, the British National Corpus provides the ability to create custom, personalized virtual corpora.

The Corpus of Contemporary American English (COCA) is the largest existing corpus of American English today. It currently contains one billion words and is freely available for use. Every year (since 1990) it is expanded by 20 million words, and the number of texts already exceeds 160 thousand.

This corpus provides three main ways to search for language units:

1) perform a search on a single word, obtaining results by collocations, topics, clusters, websites, concordances and related words for each individual unit;

2) performing a search by phrases and strings;

3) view the frequency list among the most frequent 60,000 words.

This search allows you to drill down by word form, part of speech, range of 60,000 words, and

even pronunciation. According to the compilers, such tools are especially useful for those who study and teach languages.

The corpus also allows for comparisons between genres and years of creation of the corpora. Thanks to the powerful speed of operations and the large volume of corpus data, search results are characterized by a high degree of objectivity and reliability.

Another representative corpus of modern English in its American version is the American National Corpus (ANC). This corpus is characterized by the presence of modern (since 1990) texts of all genres and transcripts of oral speech. It is also completely open for use. Today, the second version of this corpus is available for use, which contains 22 million words of written and spoken American English, which are annotated by lemmas, parts of language, noun inflections and verb conjugations. Moreover, 500,000 words of this corpus are manually annotated.

The National Corpus of the Russian Language is a large, balanced electronic corpus, the core of which consists of Russian-language texts, and includes a parallel corpus made up of a multilingual component. Unlike English-language corpora, there is only one national corpus for the Russian language, but its structure integrates several different subdivisions.

The text content of the NCRL includes main corpus, syntactic corpus, newspaper, parallel (official business, legal, legal blocks), dialect, poetic, oral, accentological, multimedia, educational, historical corpus, as well as Russian classics, panchronic, social networks and from 2 to 15.

A direct search in the NCRL allows for precise sampling. A more complex and specialized lexical and grammatical search in the corpus is carried out at grammatical, semantic and additional (in particular, punctuation marks) levels. You can search for several words with the ability to set the distance between them. Creating your own subcorpus involves narrowing the metatext features (author and title of the text, time of creation of the text, genre characteristics, etc.).

Word-formation markup in NCRL is considered in two versions, the first of which is implementation as part of semantic markup; determination of the parameters of word-formation markup in this case is carried out by selecting the "semantic features" window in the "lexico-grammatical search" form and then by selecting the parameters of the "word formation" group available in this window. In this type of markup, the set of word-formation parameters corresponds to the following types of characteristics: morphological-semantic word-formation features; word-creating digit; lexico-semantic (taxonomic) type that creates words; usual morphological type of word formation [Tagabileva, 2010]. This version of word-formation markup is available only in semantically marked NCRL corpora: main, newspaper, parallel, poetic, oral, accentological, multimedia.

Options providing parallel multilingual NCRL corpora are: WebCorp, Word Filter, IntelliText. WebCorp works on the selected information retrieval system, processing a list of URLs, selecting concordance strings from the pages found for the query. Using the operator, you can perform a simultaneous search for several words. Square brackets are used to group query elements.

The Word Filter option allows you to attach additional words that should or should not appear in the concordance lines stored for a search query. In the "Site" field, you can define the search area through a set of domain zones or URL fragments. You can also specify domains that should not be included in the search results by writing them with a minus sign.

WebCorp has results processing functions. When the search is completed, the results page provides the ability to analyze the collocations of the search term, that is, the words that appear most often in its surroundings. It is also possible to group collocations alphabetically and according to time characteristics. There are two options for sorting by time: you can select a time period from the drop-down menu. The IntelliText function has a special function called Affixes that allows you to search for prefixes or suffixes. If it is necessary to find a prefixoid, then a prefix search is used.

Characterizing the NCRL, A. Mustajoki states that the National Corpus is characterized by a balanced composition of texts [Mustajoki, 2007, p. 158]. This means that the corpus contains, as far as possible, all types of written and oral texts represented in a given language (works of fiction of different genres, newspaper and magazine articles of various subjects, advertising, special texts, diaries, correspondence), and that all these texts are included in the corpus if possible,

in proportion to their share in the language of the corresponding period.

The compilers of the NCRL differentiate the texts of the corpus as follows: modern literary prose of various genres and directions, modern drama, memoirs and biographical literature, magazine journalism and literary criticism, newspaper journalism and news, scientific, popular science and educational texts, religious and religious-philosophical texts , production and technical texts, official business and legal texts, everyday texts (including texts not intended for publication: personal correspondence, diaries, etc.). At the same time, the NCRL texts are presented in a certain proportion, reflecting their share in the total body of modern texts. In general, the corpus data are representative of written texts, including transcripts of oral speech, related only to institutional communication, to public genres of oral official communication. Oral communication is included in the NCRL as an independent subcorpus.

In NCRL, measures of stability of collocations, the absolute frequency of occurrences, and the number of documents in which the unit occurred are calculated. The typology of the proposed markup includes lemmatization, parts of speech, grammatical markup, markup of additional parameters (presence of punctuation, capitalization). The user receives a preliminary analysis of the output of the corpus (clustering of contexts), an assessment of the stability of collocations, an assessment of the probability of the appearance of language units (lemmas, parts of speech, forms of a certain case) in the nearest context. The functionality includes sorting by statistical measures, uploading data online and going to the NCRL (providing examples that meet the selected criteria). The resource ensures the development of quantitative corpus research and becomes the basis for fundamental research in the field of Russian grammar. Modern research in the field of Russian linguistics assumes a wide application of corpus-based databases. For example, contemporary linguists use them to study various lexical groups [Chesnokova, Manshin, 2018], with the aim of conducting linguistic expertise [Kotov, Mineeva, 2013], in linguistic didactics [Kornienko, 2023] and others.

Among the Chinese corpora, we note the Modern Chinese Language Corpus of the Center for Chinese Linguistics at Peking University, which became the first corpus in the country. The search in it is based on the actual distance between characters/syllables, which allows you to build a concordance. As noted by E.N. Kolpachkova, this corpus is essentially more of a “text archive” than a language corpus itself, since it has only meta-markup, and lacks such mandatory features of a corpus as morphological and syntactic markup [Kolpachkova, 2019].

Compared to other national Chinese language corpora, the Corpus of Contemporary Chinese at Peking University has several advantages. These advantages include its larger scale, longer period of corpus collection, and more detailed classification and search functions. At the same time, the corpus has a high degree of reliability and can provide more accurate and comprehensive linguistic data for language research, dictionary compilation, language teaching, and other purposes. However, the Corpus of Contemporary Chinese at Peking University also has some drawbacks. For example, the speed of updating its corpus may not be fast enough to reflect recent changes in the language. Additionally, using the corpus requires certain technical and professional knowledge, which may not be convenient for all users [Yu Shiwen et al., 2002].

The General Balanced Corpus of Contemporary Chinese Language by the State Language Commission of China contains two sub-corpora, about 13 million words of modern Chinese and about 100 million characters of the corpus of ancient texts. The corpus has part-sentence markings, including an indication of the part of speech and its grammatical categories. Only the subcorpus of modern Chinese is marked up; the subcorpus of ancient texts does not contain markup.

The timespan of the covered by corpus ranges from 1919 to 2002. It contains approximately 40 subcategories across three main categories: humanities and social sciences, natural sciences, and general education. The annotated corpus is a subset of the overall Chinese Language Corpus of the National Language Committee of the People’s Republic of China [2].

Compared to other corpora, the advantages of the General Balanced Corpus of Contemporary Chinese Language by the State Language Commission of the PRC are:

1. Large scale: The corpus has a total size of 100 million words, making it one of the largest modern Chinese corpora in China.

2. Long-term coverage: The corpus covers the period from 1919 to 2002, which characterizes the development process of modern Chinese language.

3. Optimal balance: The annotated corpus is a balanced selection based on previously developed principles for material selection, segmented by words and part-of-speech tagging to ensure balance and representativeness of the corpus.

4. Wide application: This corpus is widely used in areas, such as processing modern Chinese language and written information; formulating language and writing norms and standards; academic research on language and writing; teaching the Chinese language; social application of written and spoken Chinese.

The drawbacks of the General Balanced Corpus of Contemporary Chinese Language by the State Language Commission of China at this stage include:

1. Potential issues with data quality: some parameters may be inaccurate or contain noise.

2. Limited by subjective factors in construction, some specific linguistic phenomena may be incomplete or lack objectivity.

3. Some data may become outdated or irrelevant over time [Yang Erhong].

These are not the only variants of the Chinese language corpus, but these are the largest of them.

The national corpora mentioned above differ in volume, text composition, historical periods covered, and the presence of predominantly modern texts (British corpus) or both modern texts and those from earlier periods (Russian, Chinese). What unites the national corpora of the languages presented is their accessibility (they are freely available) and the widespread use of corpus databases.

Comparative and parallel corpora in comparative linguistic studies

It is important to note that both comparative and parallel corpora are composed to enhance translation activities, though they differ from one another in their functionality.

A comparative corpus of texts contains units selected with a specific requirement: identical samples of texts from the same genres should be associated within the same communicative spheres of the studied languages and over the same time period. What unites them is some common content that is presented in different languages which is useful for conducting contrastive and translation studies.

The texts included in the comparative corpora are not translations of the same text (unlike parallel corpora), though they belong to the same field and share the same metadata (such as year of writing, publication, author's name, publisher, etc.). It should be noted that, to date, the criteria for determining the similarity of texts are not clearly defined, but the purpose of this type of corpus is to compare languages or linguistic units functioning in similar authentic contexts and textual works, without the changes and transformations that appear in translated texts contained in parallel corpora. Such corpora are collections or catalogs of links to monolingual corpora that are not physically connected to each other.

An example of a comparative corpus is the corpus of children's speech (CHILDES). It contains transcripts of children's speech productions, most of which are spontaneous communicative events. The texts have been produced by mono- and bilingual children, brothers and sisters, children with various speech defects etc. To date, this corpus includes texts in 24 languages. This corpus can be used as a method of collecting material for studying the features of child language, which can be applied later in Linguistic didactics.

In 2014, the work of the Corpus of English-language Wikipedia (CEW) was initiated; it is based on collection of texts from the open online encyclopedia Wikipedia. The corpus contains 1.9 billion words in over 4.4 million articles that have undergone lemmatization procedures for morphological analysis. This corpus offers several options:

– word sketch – to determine the grammatical connections of English colocations;

– thesaurus – for selecting synonyms for each word;

– word lists – for compiling lists of the frequency of English nouns, verbs, adjectives, etc.;

– n-grams – for compiling frequency lists of multi-word complexes;

– concordance – for selecting examples in contexts.

Thus, comparative corpora are valuable sources for searching cross-linguistic information and translation theory and practice. The texts that are included in such corpora make it possible to check the frequency and compatibility, to state shades of meaning and distributional environment of linguistic units in authentic contexts, taking into account their situational characteristics.

Different functional options are connected with the parallel corpus. The structure of the corpus of parallel texts differs significantly from monolingual corpora. It contains the original text and its translation(s) into other language(s). In this case, the texts are segmented and aligned by paragraphs and/or sentences, that is, each paragraph and/or sentence of the original corresponds to a sentence and/or paragraph of the translation.

The base of the corpus of parallel texts must be compiled and organized in such a way as to provide direct user access to two (several) sub-corpora simultaneously, that is, the system must simultaneously display ordered units of the source text and their translation equivalents, aligned by paragraphs and sentences and formatted in accordance with linguistic and extraverbal information used by the compilers, such as part-speech affiliation, type of grammatical form, syntactic function, type of connection in a phrase, and the like. Creating parallel corpora requires special software that allows you to align paragraphs/sentences of original texts with their translated versions into other languages. According to A.N. Baranov, the most effective technologies here should be machine translation systems with their universal language, which would be an intermediary language [Baranov, 2001]. However, such a universal language has not yet been created.

Data from parallel corpora can be used in translation lexicography, comparative lexical and grammatical works, while studying the theory and practice of translation, language teaching, as well as for the development of machine translation systems.

From the perspective of theoretical translation studies, researchers are interested in the use of corpora to study translation processes and understand the cognitive essence of the translator's activity; that is, their attention is focused on analytical work with existing corpora. From a practical standpoint, there are many more questions, primarily concerning the effective use of existing corpora to create new high-quality texts and corpora. Among the main possible directions for using a parallel text corpus, translators identify the following: checking whether a certain form is used in the target language; ensuring that such a form is used in specific contexts, whether linguistic or social; verifying in which phrases and contexts such a form is used [Piotrowski, 2008, p. 118]. That is, a parallel text corpus provides the translator with a tool for quick analysis of a linguistic unit and its translation options by utilizing a large amount of textual and extralinguistic information while simultaneously immersing in the context.

Thus, using a parallel text corpus allows the translator to understand how a word or construction should be correctly translated in a specific context, which often differs from the content and options provided by traditional dictionaries. It is clear that the translator's task is not merely to find an analogue or a word equivalent in the dictionary, or to work with each individual word, but to analyze the entire text as a whole, checking the viability of each specific translation option in a given context. As McEnery states, the parallel corpus must show "how an idea formulated in one language is conveyed in another" [McEnery, 2012, p. 22]. Words devoid of contextual encirclement hold no value for the translator. A parallel text corpus helps establish such equivalents that will function organically in the new translation text alongside the authentic text, taking into account existing contexts and options, and provides the translator with the opportunity to track and consider the subtle nuances of meaning and usage of each text unit. A good example may be the Russian-Chinese corpus of the National Corpus of Russian Language (NCRL), created in 2016. At the moment it contains more than 3.5 million words of more than 1000 texts, the bulk of which are works of art by Russian and Chinese authors of the 19th – 21st centuries, news and official business texts. In addition to Russian and Chinese, the corps website is available in English. It has grammatical markings and a convenient search system. In turn, a Corpus of Russian-Chinese translations is being developed in China.

Conclusion

The high potential of parallel text corpora for translation activities does not diminish the role of monolingual corpora as a tool for translators. It is monolingual corpora that provide the translator with additional information about the subject of translation, aid in the improvement of their understanding, and update the translator's background knowledge. Such corpora are also important and serve as a source for clarifying the meanings of non-standard expressions, technical terms, occasionalisms, newly borrowed words, and for grasping the subtle stylistic and emotionally expressive nuances, and so on. Opposing it stays the corpus of parallel texts that enables search for identity patterns in translation of a certain language unit by means of another language, with the subsequent application of the following patterns to new translation units, clarification and enrichment of the register of translation equivalents in bilingual dictionaries. Quite other functionality characterizes a comparative corpus of texts which contains word structures being identical samples of texts associated with the identical communicative spheres and genres; the textual content is characterized by as information presented in different languages. Any of the three types of corpuses are useful tools for translation theory and practice.

NOTES

[1] The article is written as part of a grant in the field of philosophy and social sciences from Heilongjiang Province for the year 2024, project number 24YYC006, titled "Comparative Study of the Worldviews of the Chinese and Russian Peoples within the Framework of Cognitive Linguistics."

[2] "National Languages Committee" likely refers to a government agency responsible for language affairs in China, and "State Language Commission" also refers to a government agency responsible for language affairs in China.

REFERENCES

Baker P., 2006. *Using Corpora in Discourse Analysis*. London, Continuum, 2006. 208 p.

Baranov A.N., 2001. *Vvedenie v prikladnuyu lingvistiku* [Introduction to Applied Linguistics]. Moscow, Editorial URSS Publ., 2001. 360 p.

Biber D., 2001. Using Corpus-Based Methods to Investigate Grammar and Use: Some Case Studies on the Use of Verbs in English. *Corpus Linguistics in North America*, 2001, pp. 101-115.

Chesnokova I.D., Manshin M.E., 2018. Natsionalnyy korpus russkogo yazyka kak osnovnoy instrument poiska pri lingvisticheskikh issledovaniyakh (na primere poiska antonimov v publicisticheskikh tekstakh) [National Corpus of the Russian Language as the Main Search Tool in Linguistic Research (Based on the Search for Antonyms in Journalistic Texts)]. *Izvestiya VGPU* [Ivzestia of the Volgograd State Pedagogical University], no. 5 (128). URL: https://cyberleninka.ru/article/n/natsionalnyy-korpus-russkogo-yazyka-kak-osnovnoy-instrument-poiska-pri-lingvisticheskih-issledovaniyah-na-primere-poiska-antonimov-v

Finegan E., 2015. *Language: Its Structure and Use*. Stamford, CT, Cengage Learning. 575 p.

Kolpachkova E.N., 2019. *Korpusy kitayskogo yazyka: sovremennoe sostoyanie i osnovnye problemy* [Chinese Language Corpora: An Overview and Major Problems]. URL: https://orient.spbu.ru/images/document/2019/Kolpachkova_Chinese_Corpus_overview.pdf

Kornienko A.V., 2023. Natsionalnyy korpus russkogo yazyka kak istochnikovaya baza sotsio-gumanitarnykh issledovaniy [National Corpus of the Russian Language as a Source Base for Social and Humanitarian Research]. *Peterburgskaya sotsiologiya segodnya* [Petersburg Sociology Today], no. 21, pp. 58-71. DOI: 10.25990/socinstras.pss-21.8sqp-bb68

Kotov A.A., Mineeva Z.I., 2013. Ispolzovanie natsionalnogo korpusa russkogo yazyka pri provedenii lingvisticheskoy ekspertizy [Use of Russian National Corpus in Linguistic Examination]. *Filologicheskie nauki. Voprosy teorii i praktiki* [Philology. Theory & Practice], no. 8 (26): in 2 parts. Pt. 2, pp. 99-103.

McEnery T., Hardie A., 2011. *Corpus Linguistics: Method, Theory and Practice*. Cambridge, Cambridge University Press. 312 p.

McEnery T., Hardie A., 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge, Cambridge University Press. 294 p.

Mustajoki A., 2007. Rol korpusov v lingvisticheskikh issledovaniyakh yazykov [Role of Corpora in Linguistic Research and Language Teaching]. *Natsionalnyy korpus russkogo yazyka i problemy gumanitarnogo obrazovaniya: materialy Mezhdunar. nauch. konf.* [National Corpus of the Russian Language and the Problems of Humanitarian Education. Proceedings of the International Scientific

Conference]. Moscow, Nats. issled. un-t «Vyssh. shk. ekonomiki», pp. 152-166.
Piotrowski T., 2008. The Translator and Polish-English Corpora. *Incorporating Corpora. The Linguist and the Translator*. Clevedon, Multilingual Matters Ltd., pp. 117-132.
Semino E., Short M., 2004. *Corpus Stylistics: Speech, Writing and Thought Presentation in a Corpus of English Writing*. London, New York, Routledge. 272 p.
Tagabileva M.G., 2010. Word Formation Markup of the National Corpus of the Russian Language: Tasks and Methods. *Conference Dialogue 2010"*. URL: http://www.dialog-21.ru/digests/dialog2010/materials/pdf/73.pdf
Teodorescu M.H., 2017. *Machine Learning Methods for Strategy Research*. HBS Working Paper, no. 18-011. August 2017. (Revised October 2017). 61 p.
Yang Erhong National Language Commission Modern Chinese Language Corpus. *Encyclopedia of China (Third Edition)*. Beijing. URL: https://www.zgbk.com/ecph/words?ID=393640&SiteID=1&Type=bkzyb&webview_progress_bar=1&show_loading=
Yu Shiwen et al., 2002. Basic Processing Norms of the Beijing University Modern Chinese Language Corpus. *Journal of Chinese Information Science*, vol. 16, iss. 5, pp. 51-56.
Zakharov V.P., 2011. *Korpusnaya lingvistika: ucheb.-metod. posobie* [Corpus Linguistics. Tutorial]. Irkutstk. 161 p.

SOURCES

CHILDES. URL: https://www.sketchengine.eu/childes-corpora/
Corpus of English Wikipedia. URL: https://www.sketchengine.eu/english-wikipedia-corpus/
National Corpus of the Russian Language. URL: https://ruscorpora.ru/?ysclid=lxhsi6qgi7535286247
The American National Corpus. URL: https://anc.org/
The British National Corpus. URL: https://www.english-corpora.org/bnc/
The International Comparable Corpus. URL: https://www.researchgate.net/project/International-Comparable-Corpus-ICC

Information About the Author

Pei Haitong, Candidate of Sciences (Philology), Head of the Department of Senior Courses, Heilongjiang University, Harbin, China, peihaitong@mail.ru, https://orcid.org/0000-0002-5031-6586

Информация об авторе

Пэй Хайтун, кандидат филологических наук, заведующий кафедрой старших курсов, Хэйлунцзянский университет, г. Харбин, Китай, peihaitong@mail.ru, https://orcid.org/0000-0002-5031-6586