DOI: https://doi.org/10.18454/RULB.2021.25.1.20

КОРПУСНАЯ ЛИНГВИСТИКА В ЯЗЫКОВОМ ОБРАЗОВАНИИ

Научная статья

Балданова Е.А.[1] , Дондокова Н.Б.[2]

[1, 2] Забайкальский институт железнодорожного транспорта – филиал федерального государственного бюджетного образовательного учреждения высшего образования «Иркутский государственный университет путей сообщения»,
Чита, Россия

* Корреспондирующий автор (jenny_july[at]mail.ru)

Аннотация

Целью работы является возможность применения языковых корпусов в изучении иностранных языков. Объектом исследования выбрана идея создания корпуса английского языка, отражающего региональные и диалектные особенности языка, связанные с глобализацией. Авторы рассматривают этапы создания языковых корпусов, их типы и структуры, а также современные тенденции, влияющие на изучение английского языка. Корпусная лингвистика предлагает изучение коммуникативного использования английского языка на основе больших массивов речи. Корпусные исследования могут послужить основой для составления программ и написания учебников по английскому языку, при чтении лекционных курсов и спецкурсов по межкультурной коммуникации, лингвострановедению, культурологии, когнитивной лингвистике, методике обучения иностранному языку.

Ключевые слова: иностранный язык, лингвистика, корпус, компетенция, речеупотребление, интернет-источник.

CORPUS LINGUISTICS IN THE LANGUAGE EDUCATION

Research article

Baldanova E.A. [1] , Dondokova N.B. [2]

[1, 2] Transabaikal'sky Railway Transport Institute of the Irkutsk Transport University, Chita, Russia

* Corresponding author (jenny_july[at]mail.ru)

Abstract

The aim of the work is the possibility of using linguistic corpora in the study of foreign languages. The object of the research is the idea of creating English language corpus, reflecting its regional and dialectal features associated with globalization. The authors consider the stages of creating language corpora, their types and structures, as well as current trends affecting the study of the English language. Corpus linguistics offers the study of the communicative use of the English language based on large volumes of speech. Corpus research can serve as the basis for compiling programs and writing textbooks in English, when reading lectures and special courses on intercultural communication, linguistic and cultural studies, cognitive linguistics, methods of teaching a foreign language.

Keywords: foreign language, linguistics, corpus, competence, speech usage, internet source.

Introduction

Nowadays one of the leading areas of scientific research in the field of the English language learning is "corpus linguistics". Corpus linguistics is a branch of linguistics dealing with the development, creation and use of text corpora. The term appeared in the 1960s in connection with the development of the practice of creating of corpora, which in its turn, was facilitated by the introduction of computing technology since the 1980s.

The need for such research in modern English studies is caused by reasons of both linguistic and specifically historical nature. The latter, in particular, includes the need to improve the practice of translation, the level of proficiency in foreign languages. The movement of internationalization and integration of society in European countries has created the preconditions for a revival of interest in comparative research in the context of the higher requirements for proficiency in foreign languages imposed by the international community.

Results and Discussion

Definition of Corpus linguistics

In linguistics, corpus is a collection of texts selected and processed according to certain rules, used as a basis for the study of a language. They are used for statistical analysis and testing of statistical hypotheses, confirmation of linguistic rules in a given language. Sometimes a corpus, "corpus of the first order", is simply called any collection of texts united by some common feature, such as language, genre, author, period of creation of the texts, etc.

By definition of V.P. Zakharov and S.Yu. Bogdanova: "A linguistic or language corpus of texts is a large, machine-readable format, unified, structured, labeled, philologically competent array of linguistic data intended for solving specific linguistic problems… The main features of the modern corpus are the machine-readable format, representativeness, and the presence of metalinguistic information" [3].

Representativeness is achieved through a special text selection procedure. The expediency of creating text corpora is explained by:

— presentation of linguistic data in a real context;

— sufficiently large representativeness of data (with a large volume of the corpus);

— the ability to reuse a once created corpus for solving various linguistic problems, such as, for example, the implementation of graphematic and lexical-grammatical analysis of text, etc. [2].

Among the many definitions of a corpus, one can single out its main properties:

— electronic — in the modern sense, the body must be in an electronic form;
— representative — must "represent" well the object that it models;
— marked — the main difference between a corpus and a collection of texts;
— pragmatically oriented — must be created for a specific task.

The role of English

As we know, English is not just the main means of international communication. The scale of distribution of the English language is unique in various parts of the world. The central position in this process is occupied by the English-speaking countries, Great Britain proper, the United States, Australia, etc. Further, the circle expands, including the states where English is the "second" official language in education and administration (India, Kenya, Nigeria, Singapore). And, finally, the greatest opportunities for replenishing the "English-speaking" population are realized in countries that have different national languages, but widely use English as a foreign language [8].

Although there is now no consensus among experts on the question of what is the ratio of native English speakers to foreign speakers, most agree on a 1: 2 ratio in favor of the latter.

This state of affairs could not but affect the general strategy of studying and understanding the English language in all the variety of its functions and options.

The concept of "non-native English" (English as a foreign language) has emerged and has become firmly established, gaining more and more weight, both in linguistic research and in teaching practice. The British themselves often express the idea that soon "English as a native language" (mother-tongue English) will cease to be the dominant form of the English language, being forced out of its usual positions by "English as a foreign language" ("foreigners' English"). Thus, the fate of the preservation of a single linguistic culture largely depends on the quality of the latter.

The idea of creating a corpus of the English language largely stemmed from the need to generalize the diverse manifestations of the English language in its various versions, both dialectal and stylistic (or register). Revealing the peculiarities of dialects, scientists faced the problem of ensuring the reliability of comparative analysis data. It is possible to deduce a tendency in the development of dialects on the basis of individual observations which depend on how the comparison is carried out. After all, in order to search for reliable results, it is necessary that the material should be collected in compliance with uniform criteria and considered in a single way.

The International Corpus of the English Language

With the purpose to ensure the "comparability" of the text samples under consideration, the project for the creation of the International Corpus of the English language was primarily subordinated. Continuing the traditions of British descriptive linguistics, project manager S. Greenbaum has significantly expanded the traditional practice of registering English speech use, including new materials reflecting regional and dialectal characteristics. In conditions where countries that officially recognize English as a second state language are striving for "linguistic independence" and are considering their own version of the language as a criterion for correct use, it is important that the options do not diverge too far. The creation of an international database was intended to help preserve the identity of at least the written form of the English language [6].

The presence of a large number of texts in electronic form greatly facilitated the task of creating large representative corpora of tens and hundreds of millions of words, but did not eliminate the problems: collecting thousands of texts, removing copyright problems, bringing all texts into a single form. Balancing the corpus by themes and genres take a lot of time. Representative corpora exist or are being developed for German, Polish, Czech, Slovenian, Finnish, Modern Greek, Armenian, Chinese, Japanese, Bulgarian and other languages.

The first large computer corpus is considered the Brown Corpus (BC, Brown Corpus), which was created in 1960 at Brown University and contained 500 text fragments of 2 thousand words each, which were published in English in the United States in 1961. As a result, it sets the standard of 1 million tokens for creating representative corpora in other languages.

According to a model close to the BC, in the 1970s, a frequency dictionary of the Russian language by L. N. Zasorina built on the basis of a corpus of texts with a volume of 1 million words was created. It included approximately equal proportions of socio-political texts, fiction, scientific and popular science texts from various fields and drama.

The Russian corpus, created in the 1980s at the Uppsala University, Sweden, was built on a similar model. Its volume numbering of one million words is sufficient for the lexicographic description of only the most frequent words, since words and grammatical constructions of the average frequency occur several times per million words. From a statistical point of view, a language is a large set of rare events.

The British Corpus of the English language

Among the projects that are currently intensively developing, a special place is occupied by the British Corpus of the English language, numbering more than 100 million words. It possesses a carefully balanced material, including more than 4 thousand texts, which represent a variety of genres and varieties of language: from spoken English and newspaper articles to full-text novels. The corpus can be used primarily as a source of examples of live speech in teaching English and for research purposes to identify new trends in language development.

So, each of such everyday words as *polite* or *sunshine* occurs in the bookmaker only 7 times, the expression *polite letter* only once, and such collocations as *polite conversation, smile, request* — never.

For these reasons, as well as in connection with the growth of computer power capable of working with large volumes of texts, in the 1980s several attempts were made around the world to create larger corpora.

However, the most ambitious project, which arouses the greatest interest, is the creation of the Bank of English, carried out by the largest publisher Cobuild Collins. Already, this corpus contains 200 million words, including 15 million oral texts. A huge array of collected material feeds a new generation of scientific and educational publications by Cobuild Collins, such as, for example, the Cobuild English Language Dictionary, published in 1995. A number of research centers have created the development of corpus linguistics in such areas as:

1. Creation of corpuses of translations into English from the languages of the countries of the European Community;

2. Oral communication in English between speakers of different nationalities;

3. Euro-English — the English language of the official publications of the European Commissions.

The creation of international databases covering large corpora of speech and registering the real existence of the language in the contexts of speech use predetermined the priority development in British linguistics of such directions as text organization, pragmatics and discourse. A number of works [10] appeared, indicating the limited consideration of only the internal semantics of a word in creating a full-fledged model of language, that is, the need to refer to the act of communication, to real speech data.

Thus, the focus of attention was on the speaker's speech activity, based not only on his own “linguistic competence”, but also on the pragmatic ability to build speech in a given communication situation. The concept of context, as well as the frequency of the use of a linguistic unit in determining the meaning of a word and its place in the structure of the vocabulary of the language, has acquired special significance.

In this regard, it is interesting to note that the Collins Cobuild English Language Dictionary is one of the first English dictionaries created on the basis of a computerized corpus — it registers the meanings of words in order of frequency of their use in the most typical and established contexts. This is its significant difference from other explanatory dictionaries of the English language, which put the direct nominative meaning of the word in the first place in the dictionary entry. So, for example, the description of the meaning of the noun “*way”* in Cobuild Collins begins with “*a way of doing something or a way to do it”* (*way, method, manner*), that is, with one of the figurative meanings of the word. The direct meaning “*way” (“the way” – “the way to a particular place”*), which in most dictionaries is given first, is at the end by Cobuild Collins.

In this case, it is not just one of the meanings in the internal semantic structure of the word that is brought to the fore, but the meaning realized in typical and repetitive contexts. And this meaning acquires special social significance due to its high frequency in the composition of discourse. It is clear that in the development of the semantics of a word, the nominative meaning turns out to be primary, however, in speech reality, in accordance with the needs of the speakers, the priorities can change.

The reality of speech use requires a broader view of the word, knowledge of the “grammar of the word” and the syntactic features of its functioning. On this basis, a textological (i.e., speech study) study of lexical units arises, which presupposes not the opposition of vocabulary and syntagmatics, but their correlation. So thus, a particular interest is acquired by the word in the text, that is, in a single speech plexus, the main properties of which are the unity of the semantic connection and integrity.

For the development of this direction, corpus linguistics offers the following possibilities. First of all, this is the study of the communicative (already established) use of the language on the basis of large arrays of speech, in contrast to the strictly regulated data about the language used in normative publications to illustrate grammatical rules. Discourse as “speech immersed in life” [1] presupposes the development of a new methodology, according to which the analysis is recognized as correct, considering both the grammatical-syntactic and semantic-pragmatic aspects of the text, and also giving due contextually justified interpretation of each statement.

The figure of the speaking subject, generating speech, is “enlarged in the eyes of linguists according to its role in the speech process” [4]. Such an analysis gives noticeable advantages in identifying the peculiarities of registers and functional styles. Studies of the corpus of English have revealed significant discrepancies between the spoken and written forms of the language. In the description of oral speech, stable grammatical deviations from the standard (usually written) speech use are revealed, which in the future, possibly, will require the creation of a new grammar of the English language, reflecting its oral form.

Teaching English

Corpus linguistics currently has the greatest impact on teaching English as a foreign language. Frontal comparison of authentic English speech and texts reveals the features associated with the influence of the native language on the speech activity of speakers. The international corpus “English as a Foreign Language” is an irreplaceable source of material for conducting a comparative or contrastive study in two directions [7].

Access to reliable information on the comparison allows us to investigate the peculiarities of speech use of various groups of students, depending on their first (i.e., native) language, and thereby identify both invariant (general) errors in English, characteristic of all students, and special phenomena caused by the interference of specific languages. In other words, it becomes possible to compare certain varieties of English as a foreign language, representing the speech activity of German, French, Russian, and other students.

For example, it is known that the passive voice causes certain difficulties in teaching English for Russian-speaking students, since it occurs in English at least twice as often as in Russian. In support of this, computer comparisons of texts record the insufficient reduced use of the passive voice in the Russian part of the corpus. However, in this case, the error should be considered general or invariant, and not due to the influence of the native language, since the same problem is found in other parts of the corpus, for example, in French.

Thus, in the development of corpus studies aimed at improving pedagogical practice, two main directions have emerged:

1) contrastive analysis, which allows students to focus on difficult elements of a foreign language that reveal significant structural differences with their native language, and

2) discourse analysis, which fixes the most frequent, typical, well-established and therefore “socially significant” features of real speech use. The latter seems to be especially important, since, in contrast to systematized rules, the corpus shows what happens in living speech, and what tendencies in the use of linguistic units can manifest themselves in certain contexts.

Internet corpora

Modern technologies make it possible to create “web corpora”, that is, corpora obtained by processing Internet sources. A web corpus is a special type of linguistic corpus that is created by gradually downloading texts from the Internet using automated procedures.

These procedures determine the language and encoding of individual web pages on the fly, remove templates, navigation elements, links and ads (the so-called boilerplate), carry out transformation into text, filtering, normalization and deduplication of the received documents. And then the received data can be processed by traditional tools of corpus linguistics (tokenization, world-syntactic and syntactic annotation) and introduced into the search corpus system. Creation of a web corpus is not only much cheaper, but first of all, its size can even be an order of magnitude larger than traditional corpora [2].

Many texts available on the Internet, that is, billions of word usage for the main world languages, can be used as a corpus. For linguists, the most common way to work with the Internet is to compose queries to a search engine and interpret the results either by the number of pages found or by the first links returned. This methodology is called Googleology [9]. It should be noted that this approach is suitable for solving a limited class of problems, since the text markup tools used on the web do not describe a number of linguistic features of the text, such as specifying stress, grammatical classes, phrase boundaries, etc.

The second method consists of automatically extracting of a large number of pages from the Internet and using them as a regular corpus, which makes it possible to mark it up and use linguistic parameters in queries. This method allows you to create a representative corpus for any language sufficiently represented on the Internet very quickly, but its genre and thematic diversity will reflect the interests of Internet users [5]. The use of Wikipedia as a corpus of texts is becoming increasingly popular in the scientific community.

Computer processing of the material makes it possible to unify the parameters of the analysis and provides a systematic description of aspects of the language as part of the discourse, taking into account the peculiarities of the text organization and data of the pragmatics of speech use.

Corpora can be classified according to various criteria: the purpose of the corpus, the type of linguistic data, “literary”, genre, dynamism, type of markup, volume of texts, and so on. According to the criterion of parallelism, for example, corpora can be divided into monolingual, bilingual and multilingual. Multilingual and bilingual are divided into two types:

1) parallel — many texts and their translations into one or more languages;

2) comparable (pseudo-parallel) — original texts in two or more languages.

Markup consists in attributing special tags to texts and their components: linguistic and external or extralinguistic. The following linguistic types of markup are distinguished: morphological, semantic, syntactic, anaphoric, prosodic, discourse, etc. Further structural levels of analysis are applied to some corpora. In particular, some small corpora may be fully syntactically marked up. Such corpora are usually called deeply annotated or syntactic corpora. Manual marking (annotation) of texts is an expensive and time-consuming task. At the moment, various software tools for marking enclosures are available in the public domain [3]. They can be conditionally divided into stand-alone and web-based.

At the same time, the focus of developers in recent years has shifted towards web applications. These systems have a number of advantages:

— possibility of simultaneous marking of one document by several people;

— do not require installation of additional software, except for the browser;

— flexible differentiation of access rights;

— displaying the current progress of the marking process;

— the possibility of modifying the markup body.

Conclusion

Thus, the linguistic corpus it is a collection of texts collected in accordance with certain principles, marked up according to a certain standard and provided by a specialized search engine. So, the corpus of texts is the subject of research in corpus linguistics. Corpus research can serve as the basis for programming and writing English textbooks. Specialists, first of all, should see those linguistic phenomena that require the most attention. The analysis of the corpus of students' speech activity will reveal not only typical errors, but also determine their place in the structure of the course, which corrects both individual grammatical constructions and syntactic constructions, and more general properties of “lack of idiomacy” manifested in the expanded text.

Corpus linguistics certainly has a great future in improving the teaching of foreign languages. And perhaps in the near future the international database will serve as the basis for a new generation of educational literature, truly student-oriented.

Конфликт интересов

Не указан.

Conflict of Interest

None declared.

Список литературы / References

1. Арутюнова Н. Д. «Дискурс» / Арутюнова Н. Д. // Лингвистический энциклопедический словарь. – М. Советская энциклопедия, 1990.

2. Довнар П.Ю. Лингвистический процессор китайского языка. Особенности разработки / Довнар П.Ю., Воронцов А.В. // Международный конгресс по информатике: информационные системы и технологии: материалы международного научного конгресса 31 окт. - 3 нояб. 2011 г. – Минск: БГУ, 2011.

3. Захаров В. П. Корпусная лингвистика: Учебник для студентов направления «Лингвистика» / Захаров В. П., Богданова С. Ю. - 2-е изд, перераб. и дополн. - СПб: СПбГУ, РИО. Филологический факультет, 2013. - 148 с.

4. Золотова Г. А. Труды В. В. Виноградова и проблемы текста / Золотова Г. А. // Вести Моск. ун-та. Сер. 9. Филология. 1995. N 4.

5. Baroni M. WaCky! Working papers on the Web as Corpus / Baroni M. and Bernardini S. (editors). Gedit, Bologna, 2006.

6. Greenbaum S. CE: the International Corpus of English / Greenbaum S. // English Today. 28. October, 1991.

7. Granger S. Leaner. English Around the World. English World Wwide / Granger S. Oxford University Press, 1995.

8. Kachru B. Suandards. Codification and Sociolinguistic Realism: the English Language in the Outer Circle / Kachru B. // Quirk R., Widdowson H.G. English in the World. Cambriadge, 1985.

9. Kilgarriff A. Googleology is bad science / Kilgarriff A. // Computational Linguistics, 33(1), 2007.

10. Willis D. The Lexical Syllabus. A New Approach to Language Teaching. Collins ELT. London & Glasgow, 1990; Sinclair J. Corpus, Concordance, Collocation. Oxford University Press, 1991; B. Hoye M. Patterns of Lexis in Text. Oxford University Press, 1991.

Список литературы на английском / References in English

1. Arutyunova N.D. Diskurs [Discourse] / N.D. Arutyunova // Lingvistichesky entsiclopedichesky slovar' [Linguistic Encyclopedic Dictionary]. M. Sovetskaya entsiclopediya [Soviet Encyclopedia], 1990. [in Russian]

2. Dovnar P.Yu. Lingvistichesky protsesor kitaiskogo yazika. Osobennosti razrabotki [Chinese linguistic processor. Features of the development] / P.Yu. Dovnar, A.V. Vorontsov // Mezhdunarodny congress po informatike: informatsionnie systemi I tekhnologii: materialy mezdunarodnogo nauchnogo congressa [International Congress on Informatics: Information Systems and Technologies: Proceedings of the International Scientific Congress]. October 31 – November 3, 2011. Minsk: BSU, 2011. [in Russian]

3. Zakharov V.P. Corpusnaya lingvistica: Uchebnik dlya studentov napravleniya "Lingvistika" [Corpus linguistics: A textbook for students of the direction "Linguistics"] / V.P. Zakharov, S. Yu. Bogdanova // 2nd ed., Rev. and add. - SPb: SPbSU, RIO. Faculty of Philology, 2013. 148 p. [in Russian]

4. Zolotova G.A. Vinogradova i problem teksta [Works of V.V. Vinogradov and text problems] / G.A. Zolotova // Vesti Moskovskogo universiteta [Bulletin of Moskow university]. Ser. 9. Philology. 1995. № 4. [in Russian]

5. Baroni M. WaCky! Working papers on the Web as Corpus / Baroni M. and Bernardini S. (editors). Gedit, Bologna, 2006.

6. Greenbaum S. CE: the International Corpus of English / Greenbaum S. // English Today. 28. October, 1991.

7. Granger S. Leaner. English Around the World. English World Wwide / Granger S. Oxford University Press, 1995.

8. Kachru B. Suandards. Codification and Sociolinguistic Realism: the English Language in the Outer Circle / Kachru B. // Quirk R., Widdowson H.G. English in the World. Cambriadge, 1985.

9. Kilgarriff A. Googleology is bad science / Kilgarriff A. // Computational Linguistics, 33(1), 2007.

10. Willis D. The Lexical Syllabus. A New Approach to Language Teaching. Collins ELT. London & Glasgow, 1990; Sinclair J. Corpus, Concordance, Collocation. Oxford University Press, 1991; B. Hoye M. Patterns of Lexis in Text. Oxford University Press, 1991.