UDC 81’33
SRSTI 16.31.02
https://doi.org/10.48371/PHILS.2023.69.2.011

THEORY AND METHODOLOGY OF THE WORLD'S NATIONAL LINGUISTIC CORPORA

*Pirmanova K.K.[1], Turekhanova A.[2], Ashimova N.[3]
[*1]PhD postdoctoral student at Al-Farabi Kazakh National University, Almaty, Kazakhstan, e-mail: kunsulu.pirmanova@mail.ru,
[2]Master-teacher of Abay Kazakh National Pedagogical University, Almaty, Kazakhstan, e-mail: akmaral.turekhanova@mail.ru,
[3]A senior-teacher of Abay Kazakh National Pedagogical University, Almaty, Kazakhstan, e-mail: 05nurg@gmail.com
[1]https://orcid.org/0000-0003-3783-3199
[2]https://orcid.org/0000-0002-9925-227X
[3]https://orcid.org/0000-0002-1524-9429

Abstract. A comprehensive study of the stages of formation and development of corpus linguistics has been carried out. The aim of the article is to analyze the scientific approaches to the issue of scientific significance of the considered linguistic discipline, as well as to identify a set of concepts and criteria that constitute the foundation of this direction. The relevance of the article is determined by the fact that linguistic corpora have great potential, which has not yet been fully comprehended by the scientific community, if only because the text the main object of corpus linguistics in its various forms of implementation is one of the main components of the language system and the speech activity of a modern speaker. The principal novelty of the results of this study allows us to speak of the legitimacy of creating corpus dictionaries and corpus grammars of a new generation, designed and verified in relation to a particular fixed corpus. The novelty of the analysis lies in the fact that the appropriateness of corpus research as an essential requirement of the time, associated with the new quality of linguistic reality and meeting the needs of modern society is confirmed. The article examines the main stages in the formation of corpus linguistics as a scientific field, describes scientific ideas and approaches, peculiar to each of these stages, and provides an overview of the main principles of corpus linguistics in domestic and foreign linguistics. We analyze in detail the debate between representatives of different scientific schools and identify the advantages of one or another approach, tracing the similarities and differences between the approaches to the study of corpora at different historical stages of the formation of the studied scientific field.

Keywords: corpus linguistics, national corpus, methodology, representativeness, linguistic analysis, meta-information, classification, criteria.

Basic provisions

Corpus linguistics still only provides a very rough guide to a theory that can link individual texts to a corpus of texts, which can use what is common to the corpus to identify typical features of the language, while allowing the linguist to use inferences about frequent patterns to build a theory that combines the routine (everyday) use of language and the creative approach of linguistic analysis.

Introduction

The history of the development of any national language is accompanied by a change in ideas about its nature, essence, structure, social significance, and functioning on a global scale. This situation is relevant and understandable, as the evolution of the scientific linguistic paradigm is in principle characterized by a change of points of view, approaches, and aspects of research. Whereas for 19th-century linguistics language was interesting in itself, for 20th and 21st-century linguistics it is not so much the theoretical knowledge but rather its applied aspect that becomes relevant. In modern linguistics, text and then discourse as an object of study have defined the problem of the labor-intensiveness of research material, which required optimizing the handling of linguistic material. As a consequence, corpus linguistics emerged at the interface between linguistics proper and programming, which is becoming increasingly popular among contemporary linguists. Due to the rapid development of information technology, the collection and analysis of practical material and the whole variety of texts in different languages acquire a new meaning and require carefully developed principles and mechanisms, taking into account the main provisions of corpus linguistics. Professor G.P. Melnikov considers the collection of practical material in different languages to be new. G.P. Melnikov believes that any research carried out by a linguist should be guided by at least the following steps of activity:

1) selection of principles and bases ("etalons") of classification of studied objects;

2) the process of assigning objects to classes according to these bases ("etalons");

3) comprehension, interpretation, interpretation of results of objects assignment to classes, and explanation of reasons for the such assignment [1, p. 29].

The problem of creating a linguistic corpus is one of the topical problems of Kazakh language education, which has not yet been fully solved. Using the achievements of computer (applied) linguistics in the field of world languages for the needs of our national language is a responsible task facing specialists in this field. The research and creation of linguistic corpora are very important not only for specialists but also as a social problem. There is information in the works of foreign and Russian scholars about the reconstruction and rewriting of academic dictionaries and grammar after corpus creation. The research potential of the field of corpus linguistics is enormous. Therefore it is necessary to consider the possibilities of using corpus linguistics for research works on the texts of the Kazakh literary language and to create the scientific-theoretical rationale for the implementation of these possibilities. After all, the implementation of the results obtained in compiling linguistic corpus, in particular, their use in compiling various dictionaries, rewriting scientific grammars, or defining certain linguistic phenomena, is relevant not only for the further development of corpus linguistics but also for the formation of new technology of scientific research.

The users of the corpus, especially journalists, are usually more interested in their meta-textual information and examples of certain linguistic elements and their structural uses than in the context of real texts. The first linguistic studies carried out using corpora were limited to determining the frequency of use of various linguistic

elements in a text. Statistical methods include machine translation, speech recognition, synthesis, spelling and grammar checking, etc. They are used in solving complex linguistic problems. For example, using statistical methods on corpus material, it is possible to find out which words are always used together, indicating that they can be classified as regular phrases. From a semantic point of view, regular phrases constitute an integral (unbreakable) semantic unit, and this provision is very important to consider in the field of lexicography and automatic text processing systems.

In addition, the corpus is a rich source in the study of lexicography and grammar. Lexicographic research and research in the field of semantics are closely related. Based on observing the context of any linguistic unit in the corpus, it is possible to determine the semantic features that characterize such a linguistic unit.

Corresponding theorists use corpus as a tool to test their assumptions and prove their theories. Applied linguists (teachers, translators, etc.) use the corpus to teach languages and solve their professional tasks. Computational linguists are a special group of users of corpora: their goal is to identify and use statistical and linguistic patterns found in texts to build computer language models. Other language professionals (writers, editors) can get satisfactory answers to their questions by using corpora. Social scientists (historians, sociologists) can study their objects of study through language, i.e. through the parameters of a text called period, author, or genre. Literary scholars use the corpus for scholarly research that studies features of styles. Corpus is used to study various automated systems (machine translation, speech recognition, and information retrieval).

The founder of the applied direction in Kazakh linguistics, Professor A.K. Zhubanov, characterizes corpus linguistics as [2] a "computer linguistic base". A priori corpus linguistics has great research potential, however, as the works of Kazakh and foreign authors show, differences in approaches to the creation and use of corpus in Kazakhstan and abroad are evident.

Today, corpus linguistics can be called one of the main resources for the study of language and language descriptors. If we talk about computer linguistics, corpus formation should be considered the foundation of this linguistic discipline, allowing the creation of automated applications for processing texts and other linguistic and speech manifestations [3, p. 147-227].

Material and research methods

Methods for introducing grammatical labels into the text, linguistic-statistical, analysis and summarization methods, logical-semantic, distributive, algorithm theory, methods for creating computer databases, etc.

Regarding the question of how corpus data can contribute to theoretical linguistics, of course, they cannot replace speakers' judgments of vocabulary and grammar, but they provide specialists with a wealth of representative empirical material. Ultimately, corpora can provide three types of data that can be used in language research: *empirical support, frequency information, and extra-linguistic information (meta-information).* Let us consider these mentioned types of data in as much detail as possible.

1. Empirical support. Many reporters use the corpus as an "example bank", that is, they seek empirical support for the assumptions, principles, and rules of their research in the corpus. Of course, the examples found may be far-fetched or found by chance, but the corpus linguistics approach provides a search tool that allows for representative and balanced linguistic material as well as user choice of any corpus.

It also allows for material that proves certain scientific hypotheses correct, i.e. that it demonstrates the veracity of scientific theories. Even in the works of authors who make extensive use of corpora, conclusions from linguistic data contradict corpus data.

At every linguistic level, from the sounds of words to entire conversations and texts, the evidence can be found in corpora [4]. Situations not possible with self-observation can be re-analyzed from the corpus structure and the results can be replicated.

2. Information on frequency.

The qualitative method of using corpus is reflected in empirical support, and at the same time corpora can provide information on the frequency of use of words, phrases, and expressions for quantitative research. Quantitative studies (of course, often based on qualitative analysis) are used in many areas of theoretical and computational linguistics. They show similarities and differences between different groups of speakers or different types of texts and allow the frequency of data to be determined for psycholinguistics and other studies.

3. Metadata. In addition to the linguistic context, a corpus of texts includes the age or gender of the speaker or writer, the genre of the text, the temporal or spatial context in which the text appears, etc. provides extralinguistic information or meta-information about Such metadata allows comparison between different types of texts and different groups of speakers.

According to many scholars, corpus linguistics is not a separate paradigm of linguistics, but rather its methodology (methodology). For example, many well-known English corpora have been compiled and used for ad hoc research by representatives of different areas of linguistic science. For example, the *CHILDES* corpus has been widely used in psycholinguistic research by scholars interested in children's language acquisition through transcripts of children's spoken language in various communicative situations. The *Helsinki corpus* of various types of written texts from the earliest stages of English language development is used to study the process of language history development. The (*Bergen Corpus of London Teenage Language)* COLT collects oral speech texts of 13-17-year-olds and is used to study the languages of youth groups known in the field of sociolinguistics [5]. The rationality of linguistic analysis conducted on "real" linguistic material, allowing for qualitative results of their research, increases reporters' interest in the use of corpus [6].

Literary review

The research work takes as its methodological basis several directions of studying the world language with the help of computer technologies. In particular, 1) theoretical direction of foreign and Russian scholars in the field of computer and corpus linguistics (works of A.K. Zhubanov, G.P. Melnikov, N. Chomsky, D.N.

Ushakov, etc., as well as the first Brownian Corpus created in America in 1960, the Corpus of texts created at Uppsala University in Sweden in 1980, the Czech National Corpus created at Charles University in Prague, the Spanish National Corpus, the "National Corpus of the Russian Literary Language" created in 2001 and placed on the Internet in 2004, etc.); 2) K.B. Bektaev, A. Akhabaev, A.K. Zhubanov, S. Myrzabekov, D. Baitanaev, K. Moldabekov, A. Belbotaev, etc. who conducted lingua statistical research related to the field of morphology in Kazakh linguistics. The lingua statistical direction and practice of creating frequency dictionaries founded by scientists; 3) as I. Khamdamova, A. Akabirov, V. Mesgudov, S. Altayev, D. Tachmuradova, M. Ismailova, S. Omuraliyeva, N. Sufyanova, M. Ravshanov, J.M. Guzev regarding the practice of lexicographic research, the scientific works of Turkologist, Kazakh lexicographers-specialists such as S. Zhienbaev, I. Kenesbaev, K. Akhanov, A. Bolganbaev, B. Suleimenova, B. Kaliev, S. Bizakov and M. Malbakov on issues of lexicography.

Results

- a description of the concept of notation in a corpus, types of notation;
- defined the theoretical and practical foundations of the creation of a linguistic corpus;
- the development of the implementation of linguistic notation of the world's largest national corpus is analyzed;
- the basics of creating linguistic directories, which are the basis for creating an algorithm of automatic recognition of words in the corpus, are mentioned;

Expected socio-economic effectiveness:

- introduction of a computer program service of linguistic designations into the internet system for the general public;
- introduction of various dictionary databases into the Internet system for the general public;
- learners (pupils, students, masters, doctoral students), educators (school teachers, methodologists, textbook authors, university professors), journalists, scientists, etc. accessibility of educational language materials, etc.

Application of the scientific results obtained:

- The implementation of the automatic linguistic text analysis program primarily contributes to solving many linguistic phenomena in the field of linguistics, facilitating linguistic research, knowing the language from a new angle, creating effective linguistic practical tools;
- the implementation of the automatic linguistic analysis of the text has a major impact on the production of various teaching aids, textbooks, and tutorials, creating lexical minimums and compiling frequent dictionaries, increasing the efficiency of linguistic analysis;
- since the implementation of an automatic text processing program is closely linked to the field of computer science, it allows for the development of computer programming.

- The implementation of the automatic text processing program has its function in the public social sphere, as it opens the way for the activation of all kinds of social activities carried out through the Kazakh language.

The target consumers of the obtained results are students and teachers of higher educational institutions, graduate and postgraduate students, school teachers, researchers, lexicographers (dictionaries), programmers, etc.

Discussion

The use of corpus to study linguistic phenomena is considered an empirical working methodology based on the use of evidence, samples of language, and speech use. A corpus of data is what is usually understood as a corpus in the broadest sense of the word. For example, the Dictionary of the Royal Spanish Academy, Spain's main "linguistic" institute, defines a corpus as follows: "the most extended and ordered set of data or set of technical, literary and other texts that can serve as a basis for research". [4]. The Explanatory Dictionary of the Russian Language (ed. by D.N. Ushakov) defines a corpus as a coherent set of texts [7].

The use of computer technology for the collection, organization and processing of data was a success factor that gave the task of creating corpora a modern look, transforming individual experiments into a coherent scientific methodology and discipline, the so-called corpus linguistics.

Let us outline some of the milestones that laid the foundation for corpus linguistics and contributed to its development and consolidation as a scientific field. Thus, the Spanish researcher M. Villandre Llamares [8, p. 329-349] notes that until the XIX century, corpus in linguistics was defined as:

a) a set of written texts (data);

b) an object studied in terms of dead languages (Latin, Sanskrit);

c) an object which could only be approached by linguists through the corpus method, as it was impossible to collect linguistic data through living speakers.

In the nineteenth century and until the middle of the twentieth century, this research methodology continued to be applied and was based on the collection of large numbers of texts to:

1) explaining the process of language acquisition by children (transcription of child-parent interaction);

2) establishing orthographic norms;

3) compiling vocabulary lists for second language learning;

4) conducting comparative studies of languages;

5) developing descriptive grammar.

These lines of research have been noted by such authors as McEnery [9, p. 448-463]; McEnery, Wilson [10, p. 13-176]; McEnery, Wilson [11, p. 103-106]; McEnery, Xiao, Tono [12].

In the first half of the 20th century, American structural linguistics laid the foundations for corpus linguistics as an empirical methodology based on observation of language data, although the term corpus linguistics itself came much later, in the early 1980s. Researchers of that period believed that the corpus was the only tool suitable for the study of languages, arguing that the corpus itself could provide the

necessary data for an exhaustive description of a particular language. This new conception of the corpus, the so-called 'structural corpus', was characterized by the following features:

1) a set of oral samples or written transcriptions (data);

2) the definition of the purpose – the study of living languages, but those not previously recorded in written form (American Indian languages);

3) necessity, as collecting oral speech samples was the only way to "access" these languages;

4) focusing the work on the phonetic and (morpho-)phonological aspects as those levels where an "inventory" of all elements can be made, taking into account their complete nature;

5) ignoring the factor of representativeness of the results: as the data analysis was carried out visually and manually, it was not possible to operate with a large amount of data (which is why this methodology was criticized and considered biased) [8, p. 333].

The work on the creation of the corpus, which began in Kazakhstan at the beginning of the 21st century, is mainly based on the experience of creating a national corpus of the Russian language. At present in the country, it has been created at the Artificial Intelligence Center of L.N. Gumilev Eurasian National University, at the Department of General Linguistics and European Languages at Al-Farabi Kazakh National University, and the A.Baitursynuly Institute of Linguistics.

In Kazakh linguistic education, scientific directions related to our independence have begun to spread their wings. In the field of linguistics, the study of language from the human factor perspective has raised the possibilities of language to new horizons. Along with cognitive linguistics, psycholinguistics, and functional linguistics, the field of applied/computer linguistics, which studies language in continuity with the fields of mathematics and computer science, began to develop in the anthropolinguistic paradigm.

The field of statolinguistics in Kazakh linguistics was developed in the 1970s under the leadership of the famous mathematician Kaldybai Bektaev and Professors Askar Zhubanov, Amankesh Zekenova, Almasbek Belbotaev, and others. It starts with the experience of our statisticians in creating frequency dictionaries.

Professor Askar Kudaibergenuly Zhubanov, who for many years headed the Applied Linguistics Department at our institute, conducting research in the field of applied linguistics, such as statolinguistics, formal modeling, computer linguistics, and lecturing at universities in these areas, also stressed the need for a national corpus of the Kazakh language. he was able to discern from an early age.

The scientist A. Zhubanov explains it: "The nature of the development of the global corpus linguistics requires taking national complete texts as a special object of research. stylistic, structural, semantic, functional, etc of the Kazakh texts. Therefore, the creation of a computer database of the automated corpus of texts in the Kazakh language is also a very valuable issue from a scientific and practical point of view.

The Department of Applied Linguistics has focused its research on this problem.

The first steps in creating a national corpus of the Kazakh language have been made by placing small texts in the database and supporting morphological features in the text. In particular, between *2009* and *2011*:

- the text units (*150,000 words*) were "supported" by morphological markup (manual markup) of selected texts (fiction style) from the works of famous Kazakh writers A. Kunanbaev, M. Auezov, A. Kekilbaev, M. Magauin, M. Makataev, entered into the corpus database.

Based on this small database, a morphological notation was made for the other texts included in the corpus.

However, due to the small size of the database with "maintainable" morphological labels, it became necessary to create a morphological analyzer capable of automatically assigning morphological labels to the texts. This morphological analyzer had to automatically divide the texts included in the corpus into root and suffix, describe their relationship to the word class, and the transformation of word forms (grammatical/morphological).

At the same time, the topic *"Annotated National Corpus of Kazakh Literary Language*" (2012-2014) was presented. Professor Zhubanov A.K.:

- register words of all volumes of the 15-volume Dictionary of Kazakh Literary Language were supplemented with words from the one-volume Dictionary of Kazakh language, grammatical marks indicating their relation to the word class were checked and corrected, electronic dictionary database (register) of basic words of Kazakh language was created;

- a computer database of word-formation (word-formation) applications for each word class characteristic of the Kazakh language was created.

In addition, the structure of the word complements was predefined at the grammatical level in the database. The notation symbol was written in abbreviated form for each word complement, i.e. grammatical designations were designated by an abbreviation for word complements. A list of the word-change affixes of individual word classes was compiled according to tables compiled by A. Zhubanov according to the endings of the final sounds of the word bases (vowel, consonant, strict, shy, etc.);

- word-form database, created based on the mentioned base word base and morphological denotation database of conversion applications;

- a morphological analyzer was created based on the database of word forms created from the base words (register) and the conversion table, automatically assigning grammatical labels to corpus texts or selected texts of any volume.

- ehe dictionaries of anthroponyms and toponyms of the Kazakh language, compiled by the staff of the Onomastics Department, were included in the corpus and the software was able to show whether the onyms occurring in the texts of the corpus are anthroponyms or toponyms according to this list.

A corpus is first and foremost a large volume of texts for linguistic research. Research work is known to follow a certain system. Since language is a means of communication, a complex system with a very wide scope, language research is

based on the works of famous authors or monuments of art, print, or history published in a certain period, as well as on a certain genre, style, topic or issue. It is subdivided into branches. If such widely scattered materials are not simply introduced into corpus memory but are subordinated to a certain system, this will greatly facilitate effective research work. At the same time, the Department of Applied Linguistics, given the need to provide meta-text (extra-linguistic) information to the texts included in the corpus, launched a research project *"Metatext Label Positions in the National Corpus of the Kazakh Language"* (2015-2017) in order to study the practice of introducing metatext information in global corpora and introducing it into the corpus of Kazakh language. Collected from 5 styles of poetry and the Kazakh language.

a) fiction prose (*1 million words* from the works of M. Auezov, A. Kekilbayev, J. Aimautov, O. Bokeev, G. Musirepov, B. Mailin);

b) poetry (M. Makataev, A. Baitursynuly, J. Moldagaliev, K. Akhmetov, M. Otemisuly, M. Zhumabaev, M. Shakhanov, T. Moldagaliev, T. Aibergenov, H. Ergaliev, Sh. Kudaiberdiev, K. *1 million* words by Amanzholov, etc)

c) scientific-humanitarian texts (from Tiltanym magazine published in 2012-2014, texts of collections of scientific conferences published in recent years (2010-2015) in the field of philology, scientific collections published in Uly dala vysyasty series ("K. Zhubanov", "I. Kenesbaev", "A. Baitursynuly", etc. *1 million words*)

d) publicist texts ("Native Language", "Egemen Kazakhstan", "Young Alash", "Aikyn", "Kazakh Literature", "Zhetisu", "Astana Money", "Field and City", "Akikat", etc. *1 million words* from newspapers);

e) official (business) style texts (business documents from Internet sites, *1 million words*);

f) speech style (*1 million words* from interview texts taken from Internet sites), etc;

- studied the theoretical and practical methods of staging meta tags in the world linguistics and determined effective methods of their introduction in the national corpora of the Kazakh language; as a result, a semi-automatic computer program (with 9 parameters) for staging meta tags was created;

- meta tags (text author; text title; time and place of text writing; text size; text type; text source, etc.) were added to the texts selected from the above 5 styles;

- a 6 million word corpus of Kazakh language texts (89.250.84.132) from the 5 styles of the Kazakh language was created by the Department of Applied Linguistics;

- a monograph (ISBN 978-601-7293-43-7) / textbook on research experience "Corpus Linguistics" (by A. Zhubanov, A. Zhanabekova) was published.

Many linguists, in particular N. Chomsky [13, p. 53-58; 6, p. 41; 153; 171], who professed rationalism in the study of language, promoted that the empirical methodology of the American corpus structuralism in the 1960-the 70s gave way to the other approach: the so-called rationalism, according to N. Chomsky, "intuition of a linguist". [13, p. 558]. Chomsky criticized corpus linguistics from the theoretical point of view. He thought that the only resource of a linguist in the study of language was his intuition, which was the only meaningful criterion [13].

The difference between the concepts of the initial period of corpus linguistics and N. Chomsky's theories can be presented in the following sequence:

1 – The difference between the concepts

Corpus linguistics:	N. Chomsky:
1) focuses on phonetics and phonology	1) focuses on syntax
2) language is seen as a complete phenomenon	2) language is an entity without boundaries
3) the corpus is capable of explaining all phenomena contained within itself	3) the linguist's intuition is the only way of descriptor
4) the corpus is complete and perfect	4) the linguist's intuition is the only way of descriptive

Along with a critical analysis of Chomsky's theoretical ideas, several researchers have noted the practical problems of the first experiments in corpus linguistics. Data processing was extremely slow, expensive, and often erroneous. Thus, D.Abercrombie called corpus research a "pseudo-technique" contributing to the unreliability of analysis [14, p. 22].

It was the emergence of computer technology that gave a new impetus to this scientific trend. Some researchers refer to this period as the corpus linguistics of the new generation [8, p. 340]. The main characteristics of this period of corpus research (60-the 70s) were:

a) availability of computers: only in this period did computers become powerful enough to analyze data (although already in the late 40s R.Busa carries out the first experiments in computer processing of corpus data referring to T.McEnery) [9, p. 451];

b) representativeness of the data: most of the projects were aimed at collecting written texts whose analysis would allow us to characterize the state of the language in a given period.

In the 50s A.Joyland, based on the work of T. McEnery [9, p. 459], established framework criteria for language samples:

a) representativeness and balance;

b) tendency not to use spoken language samples due to technical difficulties and difficulties in transcribing, so written text corpora prevail;

c) size: million words [15, p. 320-339].

Justifying the need for a representative corpus, D. Barber points out that if the concept of "common language" is an abstract category and language functions as a system of different genres and/or styles, then the reference corpus should include all styles and genres of speech as well as territorial dialects. Speaking of the social representation of language, D. Barber argues that corpora should include dialects, sociolects, and professional languages, or languages for special purposes. Barber specifies [15, p. 209-213] that a language is to be presented from a historical perspective, i.e., include the texts of all known historical epochs. Thus, D. Barber thinks that the representativeness of corpus is connected with the balance, proportional representation of language genres and styles of all strata of society, which corresponds to the existing reality [15, p. 243-227]. In general, representativeness is seen by D. Barber as the representation of a wide range of functional styles and genres in a corpus of texts. P. Baker writes that the concept of representativeness is closely related to the concept of validity or the correspondence of the received data to the real state of language in the given sphere of use [6], here, as researchers believe, full representativeness in corpora is unattainable and

impossible. As E.A. Krasina notes, the verbal image, and complex image structures are governed by deep associative links, which are found at different levels of the dynamic structure of the fiction text [8, p. 337].

Recall that the first significant corpus in English appeared in the 1960s. In 1959 in the UK, R. Kirk (University College, London, UK) laid the theoretical foundations for the Survey of English Usage Corpus (SEU), the first European corpus project set up for descriptive and analytical analysis of language. The corpus consisted of 200 texts of 5000 words each, and the collection of material began in 1961. It was an attempt at a systematic description of British English, 1955-1985, based on transcriptions of spoken and written texts. This project defined the basic norms and procedures of future corpus linguistics.

Corpus linguistics as an independent discipline finally took shape in the 1990s. It was during this period that electronic corpora became an indispensable resource for language research, creating linguistic hypotheses, and building natural language processing systems. The revival of corpus linguistics (CL), in our opinion, was greatly influenced by several scientists, among whom we would like to mention J.Leach [10, p.105-122], who began a polemic with the criticism of the theoretical and practical ideas of N. Chomsky and D. Abercrombie (see above). If in the 1960s this criticism was partly objective, now, according to J. Leach, due to the evolution of computer technologies, the following main arguments are made in defense of corpus creation

1) Corpus is a scientific methodology and therefore has an undeniable advantage over intuition since it can be controlled and disregard patterns invented by linguists interested in the predicted outcome. Furthermore, in the area of quantitative data such as frequency, intuition is an unacceptable tool - our perception of frequency is subjective;

2) the grammatical nature of corpus texts, so that corpus reflects linguistic competence (with N. Chomsky arguing that because the corpus is patterns of speech manifestations and patterns of language use, they do not reflect linguistic competence. However, the work of V. Labov [6, p. 1-44] has proved a high percentage of grammatical sequences in corpora);

3) the importance of quantitative data: corpora are an incomparable source of obtaining this kind of data;

4) if the structure of a corpus is scientifically rigorous, then data relating to the frequency of use will be representative of the language as a whole;

5) the use of a computer disproves the claim that "pseudoscientific" methods are used;

6) computer processing of large volumes of information at low cost and high speed avoids subjective human errors [1, p. 111-115].

The term "corpus linguistics" entered scientific usage after 1984 when J. Aarts and W. Meijs published the paper "Corpus Linguistics In Recent Developments in the Use of Computer Corpora" [12]. From this moment the term begins to be used in its modern meaning. In our opinion, the following factors contributed to the formation of this scientific trend:

1) the flowering of applied linguistics in general and computer linguistics in particular, which made obvious the need to collect and study data on the use of language means in speech activities by both native and non-native speakers. This need is explained by the fact that, on the one hand, corpora reflect the variability of language and, on the other hand, can capture new structures or those constructions which do not correspond to theoretical descriptions. Moreover, in situations with non-native speakers, corpora are a true model of possible uses of the language in speech;

2) the eclecticism and ambiguity of the concept and its application: the use of corpus in the modern understanding of the concept does not contradict the analytical opinion of the linguist; in itself, neither corpus (American structuralist position) nor the intuition of the ideal speaker (listener) (according to N. Chomsky) is self-sufficient for explaining linguistic phenomena. It is now recognized that studying a corpus as a set of texts is impossible without the intuition and interpretative ability of the scholarly analyst who uses his knowledge of the language (as a native speaker or competent non-native speaker), and also without his knowledge of the linguistic structure (as a linguist);

3) the considerable availability of electronic corpora thanks to the Internet;

4) the development of new technologies of informatization of texts, such as optical character recognition, automatic dictation, etc.;

5) the importance of quantitative data in the study of certain linguistic aspects;

6) The need for more extensive glossaries and dictionaries to support computer systems that can deal with texts of all kinds, sub-languages, jargon, and varieties of language for special purposes (topics) (e.g. medical or legal texts) [2].

This period of the 1990s also includes the creation of scientific conceptual apparatus. J. Sinclair defines a corpus as a set of texts in a natural language chosen to characterize a variety of languages [14, p. 171]. The above definition emphasizes the main criterion for the creation of corpus: natural, that is, the pristine unprocessed text in oral or written form, the natural speech manifestation of the language form. Later, the notion of a corpus is expanded. M. Stubbs sees a corpus as a collection of texts intended for some purpose, usually research or teaching. A corpus is not something the speaker does or knows, but something constructed by the researcher. It is a record of the cumulative activity of a significant number of language users, structured for study and created to reveal characteristics of the most typical language use [14, p. 239-240]. M. Stubbs thought that computer research of large corpora might show the way out of the paradoxes of the dualism of language. Several researchers discussed whether corpus linguistics should be regarded as a theory or methodology only [10, p. 25-26]. Thus, R. Simpson and J. Swales call corpus linguistics a technique or technology of corpus creation and analysis [9, p. 449]. Most researchers conclude that the discipline of linguistics we are considering is the implementation of an empirical approach to observable data stored in the form of electronic corpora, a kind of methodological tool for the study of languages, providing innovative opportunities for language descriptions, analysis, and teaching. Also, corpus linguistics is an empirical basis for the creation of learning and teaching materials such as grammar, dictionaries, etc., both based on general discourses and

specialized discourses with an oral or written record. Thus, the Chilean scholar G.Parodi [12] supposes that corpus linguistics is a set of methodological principles for the study of any linguistic field, serving the purposes of language research in its usus based on the material of linguistic corpora and relying on computer technologies and ad hoc programs. Therefore, corpus linguistics cannot be treated as a field of linguistics similar to phonology, semantics, or syntax, but as a research method applicable to all disciplines of linguistics, at all levels of language, and in terms of various theoretical approaches. Let us give several other definitions of corpus linguistics as "the field of linguistics which specializes in obtaining results from the study of corpora" [12, p. 63-66]; "the study of language based on a text corpus". [12, p. 1]; "the use of a vast collection of accessible texts in the computer-processed form" [2, p. 7]; "a set of texts which are supposed to be representative of a given language, dialect or another dialect of the language to be used for linguistic analysis" [2, p. 13]; "a set of selected and ordered fragments of a language according to explicit linguistic criteria to be used as a sample language". [2, p. 14]; "a set of machine-readable texts of finite size, selected for maximum representativeness of the considered linguistic diversity". [2, p. 15]; as: "a corpus is a sample of language which is built based on the selection of texts made according to deterministic criteria and purpose of the research" [2, p. 151]; "a corpus is a sample of language which is built based on the selection of texts made according to deterministic criteria and purpose of the research". [2, p. 151]. But the term "corpus" should be properly applied only to the well-organized collection of data collected within a sampling frame which is intended to study a particular linguistic characteristic (or set of characteristics) through the collected data" [1, p. 49]; "A corpus is a collection of natural language texts collected in a homogeneous electronic format, selected and ordered according to explicit criteria, serving as a model of the contemporary or diachronic state or level of a particular language for scientific research" [1, p. 49]. A corpus, as a term used in modern linguistics, can best be defined as a collection of selected texts, whether written or spoken, in machine-readable form, which can be annotated with various forms of linguistic information. [10, p. 4].

As for the usuality of the language, according to E.A. Krasina [1], when Latin is more expanded (utterance-quote), it acquires usual meanings. On the contrary, with the dehumanized content of Latinism, it gravitates to the index signs and performs the functions of textual fasteners and discursive markers.

If we talk about scholars representing domestic linguistics, here we also see several non-contradictory definitions of the concept of "corpus". For example, V.P. Zakharov believes that a corpus is a large, electronically represented, structured, and marked, philologically representative array of linguistic data intended to solve specific linguistic problems [1, p. 13]. N.V. Kozlova defines a corpus as a collection of texts of one or more languages connected by certain parameters, as a collection of written and spoken utterances. In this case, the components of the corpus, texts, consist of data, and possibly metadata describing this data, and the linguistic annotations that organize this data [1, p. 83].

So, summarising the different definitions, some general criteria can be highlighted:

a) the texts should be contained in electronic media or on the Internet;

b) the size of corpora should be progressively larger, to reach 100 million words, although smaller corpora can be created for special purposes (note that previously there was a view that the larger the corpus, the more opportunities a researcher has to reflect the actual functioning of the language in all its variability, nowadays the specific focus of researchers is (the Cervantes Corpus, etc.);

c) openness: the corpus is constantly being updated, the so-called monitor corpus;

d) data authenticity: the texts must be real examples of the use of the target language;

e) selection criteria: the texts should not be chosen arbitrarily but according to the linguistic and/or extra-linguistic objectives the corpus is pursuing. This is the main criterion that distinguishes a corpus from other collections of texts, such as archives or digital libraries;

f) representativeness: the selection of texts must meet statistical parameters which will ensure that the text represents the kind of language which is the object of study (the representative sample). This type of language can refer to the work of a particular author, to a particular historical period, to a particular genre, etc.;

g) commercial orientation: corpora are not the domain and prerogative of research centers alone, many projects can be carried out within the framework of commercial entities, such as publishing concerns;

h) expanding the repertoire of languages for which corpora are being developed, as well as creating multilingual corpora;

i) the widening of the scope of the study of different linguistic aspects, from grammatical to discursive, with wide consideration of historical, psycholinguistics, and culturological factors;

j) the presented linguistic data should have a certain marking for the linguistic analysis (by marking, V.P. Zakharov understands the attributing of special marks to the texts and their components: external, extralinguistic, structural, and linguistic proper, describing the lexical, grammatical and other characteristics of the text elements [2, p. 37];

k) the analysis performed implies the possibility of classifying the obtained material, taking into account the subject matter of the text, the degree of specialization, genre characteristics, etc.

One of the most important criteria is the availability of the corpus electronically. The second important point is, of course, its representativeness. According to A.E. Kibrik, representativeness can be assessed "by the change in the relative frequency of the phenomenon in question as the sample increases. If the relative frequency of a phenomenon changes less and less with each subsequent text fragment, it means that the corpus as a whole is representative" [1, p. 21]. We believe that the representativeness of a corpus is a prerequisite that allows us to give the set of different texts the status of a corpus of texts which allows us to carry out a linguistic analysis. However, the question of unconditional representativeness of a corpus should be recognized as still open, since the linguistic activity of the human society is characterized by extreme diversity, which makes it difficult to objectively

reflect all language variants, national and cultural variants of the language in a corpus. This, in our view, explains some of the subjectivity of the results of the analysis of corpora of texts.

Let us turn to examples of the best-known and most significant corpora today:

1. 'The Bank of English' – more than 524 million words, samples of speech usage in oral and written form from different national variants of the English language: British, American, Canadian, and Australian). The texts are marked according to grammatical categories and more than 200 million words have been analyzed in terms of syntax. An important feature is the continuous updating of the corpus. At present, the project is called Project COBUILD and is carried out at the University of Birmingham under the leadership of the already mentioned J. Sinclair in collaboration with Collins COBUILD. The corpus was initiated in 1991, but since 1980 COBUILD has been collecting electronic texts for its dictionaries. All the data is publicly available on the 'Collins Word Web', a database of over two and a half billion words, to which 35 million are added each month. It is the most extensive resource of its kind in the world.

Another important corpus resource for the English language is the 'British National Corpus (BNC), with 300 million words of modern British English in written and spoken form. The project is being carried out under the auspices of a scientific and industrial consortium led by Oxford University Press, together with other publishers specializing in dictionaries, Lancaster University, Oxford University, and the British Library. The corpus was created between 1991 and 1994 and is an example of a closed corpus to research late 20th-century British English to develop reference materials (consisting of 90% written texts and 10% oral texts).

2. 'Corpus de Referencia del Español Contemporáneo' (CREA), a databank of modern Spanish (from 1975 to the present) developed by the Royal Spanish Academy (la Real Academia Española). It contains over 200 million words. 90% of the samples are written texts, the rest are oral. The corpus takes into account geographical, thematic, and chronological criteria as well as the source from which the text was obtained. The data bank is considered a monitor corpus - new texts are periodically added to increase the representativeness of the corpus. It is the most significant corpus of Spanish, serving both for academic research and for the creation of commercial products.

3. Corpus Diacrónico del Español (CORDE), a databank of Spanish from a diachronic perspective, a collection of texts from the formation of Spanish until 1975. This corpus is the historical complement of the CREA, with a total of over half a million words.

An important aspect in the theory of the study of linguistic corpora can also be considered their different types. N.V. Kozlova believes that all the existing multitude of corpora of texts can be divided into three categories: 1) freely accessible; 2) partially accessible and 3) commercial [10, p. 76-89]. J. Sinclair and J. Torruella and J. Listeri [6, p. 45-77] have developed parameters for classifying texts (note that in practice this typology is not always explicit)

1) the kind of language;

2) the number of languages to which the texts belong;

3) the boundaries of the corpus (a corpus whose purpose is to describe a sublanguage (legal, informatics language, etc.) may be of limited size);

4) the general or specialized nature of the texts;

5) the temporal cross-section the texts cover;

6) the data analysis and processing techniques applied to the corpus.

In most cases, these criteria are determined by the purpose for which the corpus is designed: the study of an author's work (e.g. Abai's poems) or a literary work of a specific historical period, a description of the national language in general (modern Kazakh) or a specific variety of language, a territorial dialect, a language for a specific purpose, or a specific linguistic aspect (e.g. Kazakh cultural norms, medical texts, etc.), a particular commercial product (e.g. tourist dissemination, a travel agency, etc.), and the acquisition of knowledge.

Depending on the period covered by the texts, corpora are divided into:

A) diachronic, or historical corpus: it includes texts from different periods, allowing us to analyze the evolution of language over a long period and to investigate the historical development of a language phenomenon or of the whole language system, which distinguishes them from the monitor corpus, which does not cover such long periods (the CORDE corpus we mention);

B) synchronous corpus: the presentation of textual material to consider the state of a language as a system at a certain point in time (British National Corpus).

It is also possible to classify corpus according to their existing mark-up, namely as unmarked and marked-up. An unlabelled corpus is an array of texts that contain a certain number of mentions of the item searched for. However, the search results provided by unlabelled corpora can be used in linguistic research, but only from a purely statistical point of view. The corpora marked in terms of morphological, syntactic, prosodic, and other characteristics provide much more opportunities for linguistic analysis.

Of course, this list of possible typological criteria listed is neither closed nor does it claim to be a strict boundary for delineating corpus types.

Conclusion

Thus, a corpus, which is a reflection of a language, is, like a language itself, a dynamically evolving system and implies all new criteria and approaches to describing and analysing linguistic material and developing new methodological procedures. A corpus can provide detailed information about a particular language, but it is impossible to collect a corpus covering an entire language because it is impossible to collect all the samples of that language use, so it should always be assumed that a corpus is just some finite collection of samples of the infinite universe of a language.

We can conclude that a corpus is an electronically represented, usually marked for linguistic analysis, provided with a relatively easy-to-use search engine, representative array of unedited texts representing the maximum number of variants of a language. Whereas in the infancy of corpus linguistics researchers pointed out that linguistic variation could be neglected, with the advent of electronic corpora the diversity of language forms has become more evident and the possibilities of

language data research have expanded. The extensive typology of corpora created with different criteria and their diversity allows both the linguist and the lay user to choose the one that suits the aims and objectives of a particular and specific scientific study.

The corpus is now a unique resource for any linguistic research in general and computer linguistics in particular. Its main advantages lie in the fact that it consists of real language samples, ensures the objectivity of the obtained results and conclusions, and makes it quite easy to verify the validity of this or that theory, its merits and demerits. Thanks to the introduction of computers with ever-increasing storage capacity and data processing speed, access to language and speech samples has become fast and reliable, as has data extraction, processing and analysis. On the other hand, corpora provide statistical and quantitative data that would otherwise be inaccessible due to high costs or impossible due to unreliable results obtained by "manual" processing, given the large size of individual corpora. Thanks to the development of corpus linguistics methodology and techniques, descriptive studies of languages are available to researchers and any other user, supported by corpora at any of the linguistic levels: phonetic-phonological, grammatical, semantic, pragmatic and others. The exceptional value of the use of corpora as a source of data is revealed in the teaching of mother tongues and foreign languages or in the development of didactic materials, dictionaries, grammars, other products related to machine translation or speech technology, etc.

Having studied some of the conditions of the "background" of corpus linguistics, as well as the current state of this scientific discipline, having analysed the data available today, as well as the continued creation and replenishment of corpora, we can conclude that the evolution, the dynamics of this promising scientific field is relevant to linguistic theory and practice and, according to G. Parody, is reaching boiling point, but the mechanisms of computer linguistics undergo a process of constant changes and adjustments, allowing to significantly enrich and make.

This research has been funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Grant No. AP15473441)

REFERENCES

[1] Mel'nikov G.P. Sistemnaja tipologija jazykov: Principy, metody, modeli (System typology of languages: Principles, methods, models.). M.: Nauka, 2003. 395 s.[in Rus.]

[2] Jūbanov A.Q. Korpustyq lingvistika (Corpus linguistics). Almaty: «Qazaq tılı» baspasy, 2017. 336 b. [In Kaz.]

[3] Moure T., Llisterri J. Lenguaje y nuevas tecnologías: el campo de la lingüística computacional" // M. Fernández Pérez (coord.) Avances en Lingüística aplicada. Santiago de Compostela: Universidade de Santiago de Compostela, Servicio de Publicacións e Intercambio Científico, 1996. P. 147-227.

[4] Real Academia Española. Diccionario de la lengua española. Madrid: Espasa, 2001. Access mode: https://dle.rae.es/corpus (date of application: 10.11.2020).

[5] Aijmer K., Altenberg B. (eds.) English Corpus Linguistics. Studies in Honour of Jan Svartvik. London: Longman, 1991.

[6] Chomsky N. Language and mind. Cambridge, 2006.

[7] Ushakov D.N. Tolkovyj slovar' russkogo jazyka (Explanatory dictionary of the Russian language.) Rezhim dostupa: https://gufo.me/dict/ ushakov/korpus (data obrashhenija: 29.10.2020).

[8] Villayandre Llamazares M. Lingüística con corpus // Estudios humanísticos. Filología. 2008. № 30. P. 329-349.

[9] McEner T. Corpus Linguistics // The Oxford Handbook of Computational Linguistics / R. Mitkov (ed.). Oxford: Oxford University Press, 2003. P. 448-463.

[10] McEnery T., Wilson A. Corpus Linguistics. Edinburg: Edinburgh University Press, 1996.

[11] McEnery T., Wilson A. Corpus Linguistics. Edinburgh: Edinburgh University Press, 2001.

[12] McEnery T., Xiao R., Tono Y. Corpus-Based Language Studies. An advanced resource book. London – New York: Routledge, 2006. Access mode: https://www.lancaster.ac.uk/fass/projects/corpus/ZJU/xCBLS/chapters/B03.pdf (date of application: 29.10.2020).

[13] Chomsky N. Quine's empirical assumptions // Words and objections. Essay on the Work of W.V Quine / D. Davidson & J. Hintikka (eds.). Dordrecth: D. Reidel, 1969. P. 53-68.

[14] Abercrombie D. Studies in Phonetics and Linguistics, London: Oxford University Press, 1965. Access mode: http://www.davidcrystal.com/Files/BooksAndArticles/-4896.pdf (date of application: 06.11.2020).

[15] Juilland A.G., Brodin D.R., Davidovitch C. Frequency dictionary of French words. Hague – Paris: Mouton, 1970.

ӘЛЕМДІК ҰЛТТЫҚ ТІЛДІК КОРПУСТАРДЫ ҚҰРУ ТЕОРИЯСЫ МЕН ӘДІСТЕМЕСІ

*Пірманова К.Қ.[1], Тureханова А.[2], Ашимова Н.[3]

[*1] әл-Фараби атындағы Қазақ ұлттық университетінің PhD постдокторанты, Алматы, Қазақстан, e-mail: kunsulu.pirmanova@mail.ru,

[2]магистр, Абай атындағы Қазақ ұлттық педагогикалық университетінің аға оқытушысы, Алматы, Қазақстан, e-mail: akmaral.turekhanova@mail.ru,

[3]магистр, Абай атындағы Қазақ ұлттық педагогикалық университетінің аға оқытушысы, Алматы, Қазақстан

e-mail: 05nurg@gmail.com

[1]https://orcid.org/0000-0003-3783-3199

[2]https://orcid.org/0000-0002-9925-227X

[3]https://orcid.org/0000-0002-1524-9429

Аңдатпа. Корпустық лингвистиканың қалыптасуы мен даму кезеңдеріне кешенді зерттеу жүргізілді. Мақаланың мақсаты қарастырылып отырған лингвистикалық пәннің ғылыми маңыздылығы мәселесіне ғылыми көзқарастарды талдау, сондай-ақ осы бағыттың негізін құрайтын ұғымдар мен критерийлер кешенін анықтау болып табылады. Ұсынылған мақаланың өзектілігі лингвистикалық корпустарда үлкен әлеует бар екендігімен анықталады, оны ғылыми қауымдастық әлі толық түсінбейді, тек мәтін — корпустық лингвистиканың негізгі объектісі — оны жүзеге асырудың әртүрлі формаларында тіл жүйесінің негізгі компоненттерінің бірі және қазіргі ана тілінің сөйлеу әрекеті. Осы зерттеу нәтижелерінің түбегейлі жаңалығы белгілі бір бекітілген корпусқа қатысты әзірленген және тексерілген жаңа буынның корпустық сөздіктері мен корпустық грамматикаларын құрудың заңдылығы туралы айтуға мүмкіндік береді. Жүргізілген талдаудың жаңалығы корпустық зерттеулердің орындылығы лингвистикалық шындықтың жаңа сапасымен байланысты және қазіргі қоғамның қажеттіліктеріне жауап беретін уақыттың маңызды талабы ретінде расталғандығында. Мақалада корпустық лингвистиканың ғылыми бағыт ретінде қалыптасуының негізгі кезеңдері қарастырылады, осы кезеңдердің әрқайсысына тән

ғылыми идеялар мен тәсілдер сипатталады, отандық және шетелдік лингвистика шеңберіндегі корпустық лингвистиканың негізгі тұжырымдамалық ережелеріне шолу жасалады. Біз әртүрлі ғылыми бағыттардың өкілдері арасындағы қайшылықтарды егжей-тегжейлі талдаймыз және зерттелетін ғылыми бағыттың қалыптасуының әртүрлі тарихи кезеңдеріндегі корпустарды зерттеу тәсілдерінің ұқсастықтары мен айырмашылықтарын қадағалай отырып, белгілі бір тәсілдің артықшылықтарын анықтаймыз.

Тірек сөздер: корпустық лингвистика, ұлттық корпус, әдістеме, репрезентативтілік, лингвистикалық талдау, метаақпарат, классификация, критерийлер.

ТЕОРИЯ И МЕТОДОЛОГИЯ СОЗДАНИЯ МИРОВЫХ НАЦИОНАЛЬНЫХ КОРПУСОВ ЯЗЫКА

*Пирманова К.К.[1], Туреханова А.[2], Ашимова Н.[3]

[*1] PhD постдокторант Казахского национального университета имени аль-Фараби, Алматы, Казахстан, e-mail: kunsulu.pirmanova@mail.ru,

[2]магистр, старший лектор Казахского национального педагогического университета имени Абая, Алматы, Казахстан,
e-mail: akmaral.turekhanova@mail.ru,

[3]магистр, старший лектор Казахского национального педагогического университета имени Абая, Алматы, Казахстан,
e-mail: 05nurg@gmail.com

[1]https://orcid.org/0000-0003-3783-3199
[2]https://orcid.org/0000-0002-9925-227X
[3]https://orcid.org/0000-0002-1524-9429

Аннотация. Проведено комплексное исследование этапов становления и развития корпусной лингвистики. Целью статьи является анализ научных подходов к вопросу научной значимости рассматриваемой лингвистической дисциплины, а также выявление комплекса понятий и критериев, составляющих фундамент данного направления. Актуальность представляемой статьи определяется тем, что в лингвистических корпусах заложен огромный потенциал, который еще не в полной мере осмыслен научным сообществом, хотя бы в силу того, что текст — основной объект корпусной лингвистики — в различных формах своей реализации представляет собой одну из главных составляющих системы языка и речемыслительной деятельности современного носителя языка. Принципиальная новизна результатов данного исследования позволяет говорить о правомерности создания корпусных словарей и корпусных грамматик нового поколения, разработанных и верифицированных по отношению к конкретному фиксированному корпусу. Новизна проведенного анализа заключается в том, что подтверждена целесообразность корпусных исследований как сущностное требование времени, связанное с новым качеством лингвистической реальности и отвечающее потребностям современного общества. В статье рассматриваются основные этапы становления корпусной лингвистики как научного направления, характеризуются научные представления и подходы, присущие каждому из этих этапов, представляется обзор основных понятийных положений корпусной лингвистики в рамках отечественного и зарубежного языкознания. Мы подробно анализируем полемику между представителями различных научных направлений и выявляем преимущества того или иного подхода, прослеживая сходства и различия между подходами к изучению корпусов на различных исторических этапах становления изучаемого научного направления.

Ключевые слова: корпусная лингвистика, национальный корпус, методология, репрезентативность, лингвистический анализ, метаинформация, классификация, критерии.

Статья поступила 07.11.2022