

**THE REPUBLIC OF AZERBAIJAN**

*On the Rights of a Manuscript*

**THE OPTIMAL STRUCTURE OF THE VOCABULARY  
BLOCK IN THE NATIONAL CORPUS OF THE  
AZERBAIJANI LANGUAGE**

Specialty: 5704.01 – Language theory

Science: Philology

Applicant: **Rena Mahmudova Huseyn kizi**

**THESIS**

of the dissertation submitted  
for the degree of Doctor of Philosophy in Philology

Baku-2023

The dissertation work was carried out in the department of Computer Linguistics of the Linguistics Institute named after Nasimi of ANAS.

Scientific adviser: doctor of philological sciences, professor  
**Masud Ahmad oglu Mahmudov**

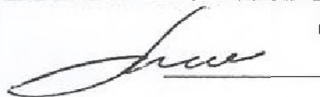
Official opponents: doctor of philological sciences, professor  
**Ilham Mikayil oglu Tahirov**

doctor of philosophy in philology,  
assosiate professor  
**Cavanshir Khankishi oglu Muradov**

doctor of philosophy in philology  
**Sevinj Salim kizi Mammadzade**

Dissertation Council ED 1.06 operating under the Linguistics Institute named after Nasimi of the Azerbaijan National Academy of Sciences of the Higher Attestation Commission under the President of the Azerbaijan Republic

Chairman of Dissertation Council:



doctor of philological sciences, professor  
**Nadir Balaoglan oglu Mammadli**

Scientific Secretary of Dissertation Council:



doctor of philosophy in philology, associate professor  
**Sevinj Salim kizi Mammadova**

Chairman of the scientific seminar:

doctor of philological sciences, associate professor  
**Gulsum Israfil kizi Huseynova**



## INTRODUCTION

**The relevance and degree of development of the topic.** The age we live in is the age of information technologies, the age of the Internet. All the countries of the world prefer to use electronic means to get information about any event, investigate problems and get the necessary results. The development of society, which is changing day by day, is accelerated through information technology. This rapid development has an important role in the development of linguistics as well as in all fields of science. Therefore, the representation of the language in the electronic environment, the more effective effect of scientific researches by the electronic method, necessitates the application of information technologies to the language.

For a long time, the special attention has been paid to the ideology of the information society as an object of scientific and theoretical research. E. Toffler, D. Bell, M.Castells and other well-known scientists played a special role in the historical development of this ideology. They adjoined this new qualitative phase of human history from different perspectives and defined it differently. Therefore, there are different approaches to the ideology of the information society. The main goal of the information society is to solve the problems facing humanity and ensure the development of society by creating an abundance of information. Against the background of these issues, fundamental studies are being conducted in the direction of investigating the place and role of language in the visual field<sup>1</sup>.

*“If we look the concept of information society from a technological point of view, we will see that here the computer and the Internet perform the collection, storage, processing, transmission and other important functions of information as the main technical*

---

<sup>1</sup> Habibova, K.A., Jafarov, Y.M. Language policy in the virtual space // – Atlantis press, – 2019. – Vol.: 81, – p.789-792. URL: <https://www.atlantispress.com/proceedings/mtde-19/125908941>

*means*"<sup>2</sup>.

Lately, a new and very promising field - corpus linguistics - has emerged within the framework of computational linguistics. The research object and directions of corpus linguistics are so wide and comprehensive that many specialists do not hesitate to consider it as an independent field of science. Corpus linguistics studies the creation of information-search systems aimed at the systematic collection of texts representing any specific language in an electronic format according to predetermined rules and their use in the research process. In the national corpus of a specific language, that language fully includes all genres, literary language, styles, grammar, vocabulary, etc.<sup>3</sup>

The national corpus of any language can also be considered a model that imitates that language. This model provides an opportunity to explore many complex issues related to language learning.

It should be noted that language corpora have brought new essence and ideas to linguistics both on the theoretical level and on practical issues. Linguistic corpora have created a foundation for the formation of new attitudes to linguistic phenomena of a new perspective at all levels of language. Lexicography is considered to be the area most affected by language corpora. The point is that lexicography is a very hard and labor-intensive field of activity. In traditional dictionaries, it takes a lot of time to collect, systematize, explain and arrange words according to the alphabet, and when the work of compiling the dictionary is finished, the dictionary is already considered outdated. As Agamusa Akhundov noted, "The explanatory dictionary of the Azerbaijani language has lived a life of 60 years of compilation and 21 years of printing"<sup>4</sup>. Undoubtedly, 60

---

<sup>2</sup> Əliquliyev, R.M. İnternet-jurnalistikanın formalaşmasının bəzi aspektləri / R.M.Əliquliyev, C.F.Valehov, R.Ş.Mahmudov. Ekspress-informasiya. İnformasiya cəmiyyəti seriyası. – Bakı: İnformasiya cəmiyyəti, – 2008, – 32 s.

<sup>3</sup> Mahmudov, M.Ə. Türk dillərinin milli korpusu / M.Ə.Mahmudov. – Bakı: Elm və təhsil, – 2018, – 392 s.

<sup>4</sup> Azərbaycan dilinin izahlı lüğəti: [4 cilddə]. – Bakı: Şərq-Qərb, – 2006, I cild, – 744 s.

years of design and 21 years of print life is a long time. During this period, great changes took place in the vocabulary of the Azerbaijani language, many new words entered the vocabulary of the language, and a certain amount of words became archaic and passed into the passive background. Therefore, the dictionary can be considered out of date until it is compiled and printed. Of course, the material collected in the 40s cannot be considered new for the 90s. That is why corpus linguistics and corpus lexicography can be considered as the newest directions of linguistics. It is no coincidence that this new direction of lexicography is given great importance in Turkology. All this can be considered as one of the factors determining the relevance of the research work submitted to the defense.

One of the most important, leading components of national language corpora is the block of dictionaries. The dictionary block can also be characterized as the section that computer users turn to for various purposes. That is why creating a block of dictionaries in national language corpora and determining its optimal structure is a very important and urgent issue. The recent start of scientific research on corpus linguistics in Azerbaijan republic and the recent works in this field increase the relevance of the topic once more.

As mentioned, corpus linguistics is considered a promising field that has emerged recently and is rapidly developing. Nevertheless, certain researches related to corpus linguistics have already been carried out, a number of works have been done and are being successfully continued. The first information about Corpus Linguistics in Azerbaijani linguistics was given in the monographs "Computer Linguistics" and "National Corpus of Turkic Languages" by M. Mahmudov.

In relation to corpus linguistics, the project called "Development of the dictionary composition of the Azerbaijani language, public monitoring of compliance with language norms, and creation of an integrated electronic system and dictionary for the preparation of the corpus of the language" can be characterized as the first theoretical and experimental success in the field of creation of the national corpus of the Azerbaijani language.

Another study related to the development of the topic is the dissertation work of S.S. Mammadzadeh on the topic "National language corpora and principles of their formation".

In addition to these, reports were made at many scientific conferences related to the dissertation topic, and articles and materials were published in scientific journals. It is possible to find information about those works in the used literature section of the dissertation.

**The object and subject of the research.** The subject of the dissertation work is the methods and ways of compiling national corpora of world languages. The subject of the research work is to determine the optimal structure of the ways of placing dictionaries in the block of dictionaries in the national corpus of the Azerbaijani language.

**The research goals and aims.** The main goal of the research is to study the issues of creating the optimal structure of the dictionary block in the national corpus of the Azerbaijani language, to facilitate the use of dictionaries of various purposes in the research or translation process. For this purpose, the solution of following tasks are primarily taken into consideration:

- to explain the essence, goals and tasks, spheres of use of corpus linguistics;
- to clarify the place and linguistic status of corpus linguistics in computational linguistics;
- study of world experience in creating national language corpora;
- to determine similar and different features of national language corpora;
- to analyze structural differences and peculiarities of national language corpora;
- to analyze the structure and features of national corpora on Germanic and Romance languages;
- to determine the features of corpora prepared in the materials of Slavic languages;
- Research the possibilities of creating and using the national

corpus of the Azerbaijani language;

- to analyze the issues of placement of all types of dictionaries in the dictionary block;

- analysis of the structural features of the block of dictionaries of national language corpora;

- to determine the optimal structure of the block of dictionaries in the national corpus of the Azerbaijani language;

- to study the functions of monolingual, bilingual, translation, sub-corpora of the statistical dictionaries placed in the dictionaries block;

- to develop the methods of placement and use of other dictionaries (orthography, orthoepy, phraseology, dialectology, abbreviations, emphasis, etc.) in the block of dictionaries;

- to interpret the essence of concordances, their place and functions in the block of dictionaries;

- to explain the importance of concordances on the basis of the works of individual writers in the study of the literary language as a whole;

- to explain the necessity of establishing interaction with national corpora of languages of non-kinfred systems.

**The research methods.** Descriptive, comparative-contrast and system-structural approach methods were used in the dissertation work.

**The main defended arguments include:**

- the block of dictionaries is one of the most important components in national language corpora. Unlike other components, it is the dictionaries block that users frequently refer. The main issue is to arrange the optimal forms of communication and relations between the dictionaries represented in the national corpus;

- The electronic dictionaries related to the Azerbaijani language in the electronic space cannot be included in the block of dictionaries of the national corpus of the Azerbaijani language;

- The block of dictionaries of the national corpus of the Azerbaijani language is an open system. If necessary, electronic versions of new dictionaries can be included there, or electronic

dictionaries already placed there can be revised and improved;

- The handing out of each vocabulary unit in the dictionary block with its context is one of the important conditions;

- the corpus of bilingual and multilingual texts should be included in the vocabulary block of the national corpus. The availability of bilingual and multilingual corpora can greatly facilitate the users' work;

- information related to any word (vocabulary unit) searched for in the dictionary block should be provided to the user by taking and adding it from all representative dictionaries. Some of this information may not be of interest to the user. This is considered acceptable in terms of completeness and standard of information;

- the concordances of the most common, well-known historical monuments, works of individual writers should be placed in the dictionary block;

- vocabulary units in the dictionary block of the national corpus should be linked with the corpus of texts representing the literary language of Azerbaijan. Users should be able to benefit from the result of this connection. The user should be able to obtain the contexts that characterize any vocabulary item linguistically. This is important in terms of the full understanding of the meaning of the word;

- The block of dictionaries in the national corpus of the Azerbaijani language should have the possibility of communication and mutual exchange with the national language corpora of other related and unrelated languages.

**The scientific novelty of the research.** For the first time, the issues of creating the national corpus of the Azerbaijani language were investigated and the optimal structure of its vocabulary block was developed in the dissertation. In the research process, the structure of the national corpora of languages of non-kindred systems and their blocks of dictionaries was investigated, and similar and different features were analyzed by studying the experience of preparing national language corpora in advanced science centers of the world. As a result, it was considered appropriate to apply the



most optimal options for placing dictionaries in the dictionary block of the national corpora of German, Slavic, Turkic and other world languages to the dictionary block of the national corpus of the Azerbaijani language.

**The theoretical and practical significance of research.** The issues of creating national language corpora can be characterized as both theoretical and practical research. For the first time, the theoretical foundations and concept of national language corpora are prepared, then the practical construction of that system is carried out. However, this process interacts with theoretical and practical issues and accomplishes each other. Aspects of the theoretical concept that do not justify themselves in practical use are analyzed and revised. Therefore, the creation of national language corpora as a whole can be considered the result of mutual relations at the theoretical and practical level. The main theoretically important issue of the research is the development of the theoretical foundations of the problem of creating a block of dictionaries of national language corpora, determining its optimal structure, and systematizing it. The analysis of the theoretical aspects of the work shows that corpus creation is of great importance for the development of modern linguistic science. Theoretical observations and analyzes in this field find their application in practical systems.

The practical importance of the research is that the results of the research can be applied in the process of improving the existing systems related to the creation of national corpora in Azerbaijani and other Turkic languages.

The study was mainly conducted in a synchronous aspect. Theoretical literature and other materials related to Azerbaijani, Turkish, Russian and English linguistics were used as sources in the dissertation. The experience of preparing and using national language corpora in the world's leading science centers is taken as an example. Structural features of national language corpora placed on the Internet and popular with users have been the focus of attention.

The issues of creation and use of national language corpora are neoteric for Azerbaijani linguistics, and the teaching of this new field

of linguistics in higher schools is extremely important and useful. For this purpose, the results of the research can be used in the preparation of textbooks and teaching aids.

**The research approbation and implementation.** The main content of the research is reflected in the articles published in various scientific collections. The results of the research were reported at scientific conferences and seminars. The published articles fully cover the content of the dissertation. 21 scientific articles (including 5 abroad) were published on the subject.

**The name of the institution where the dissertation was carried out.** The dissertation work was carried out at the Department of Computer Linguistics of the Linguistics Institute named after Nasimi of the Azerbaijan National Academy of Sciences.

**The structure of the research work.** The dissertation consists of introduction, three chapters, conclusion, list of used literature and list of abbreviations.

**The volume of the structural sections of the dissertation separately and the total volume with a sign.** Introduction is 9 pages, I chapter is 54 pages, II chapter is 21 pages, III chapter is 27 pages, conclusion is 3 pages, list of used literature is 18 pages, list of abbreviations used in the dissertation is 2 pages. The total volume of the work is 136 pages - the number of signs is 206,511.

## **THE MAIN CONTENT OF THE RESEARCH**

The relevance and the development degree of the topic, the object and subject, goals and objectives, methods of the research, the defense provisions are defined, the scientific innovation, theoretical and practical significance of the research, the approval and application of the research work, the name of the organization where the dissertation work is performed, the structure of the dissertation are defined in the introduction of the dissertation. The information about the volume of the sections separately and the total volume with a sign is presented.

The first chapter of the dissertation is called "*World experience in creating national language corpora*". This chapter consists of 6 sub-chapters.

In I subchapter of this chapter, "*Essence and scope of use of national language corpora*" was analyzed. Corpus is an information-retrieval system consisting of a collection of texts, usually in natural language, collected and stored electronically during the study of language based on research and studies. The corpus includes texts of various genres and types, such as news texts, examples of fiction, scientific articles, etc. Corpora are used to determine the frequency of word processing based on certain language samples, analyze grammatical structures, examine stylistic features, machine translation, speech recognition, etc. It is also used for developing and training natural language processing computer systems and many other purposes.

Currently there are a large number of corpora in electronic format covering many world languages. For example, it is shown that there are a large number of corpora used in linguistic research in the Russian language. From this point of view, any existing collection of texts related to a certain topic is considered a corpus. Corpus is a collection of texts collected, marked and annotated for certain stages of linguistic analysis according to certain principles. This explanation about the corpus can also be applied to the texts collected in the framework of the machine fund of national languages.

According to Russian linguist V.P. Zakharov, "*a corpus is a word of German origin and is a collection of texts used for the purpose of linguistic analysis*"<sup>5</sup>... The contents of these texts consist of thousands and millions of words and are stored in the computer's memory. Most modern corpora are systematized. This means that the texts are systematized according to their characteristics, that is, according to genres and dialects. A corpus is a collection of

---

<sup>5</sup> Захаров, В.П. Корпусная лингвистика. Учебно-методическое пособие // В.П.Захаров. – Санкт-Петербург: Санкт-Петербургский государственный университет, – 2005, – 48 с.

systematized, computerized texts. The corpus is created for the purpose of learning any language. Other sets of the language are used for a different purpose. The composition of the corpus should support the purpose of language learning.

E. Finegan explained the meaning of corpus in his textbook as follows: "*Corpus is a collection of texts in machine format containing information obtained under any circumstances*"<sup>6</sup>.

According to E. Wilson and T. McEnery, "*a corpus is a set of language fragments that includes all criteria relevant to any selected language*"<sup>7</sup>.

According to M. Mahmudov, "*a corpus is a set of natural language texts arranged in a certain order, providing prompt and accurate information to researchers about various linguistic events and facts stored in an electronic medium*"<sup>8</sup>.

The main users of national corpora are linguists-researchers of various profiles. However, users of the corpus should not be limited to only tens. Literary experts, historians, other representatives of the humanities, and ordinary users can also be interested in reliable information about the stylistic features of the language of a certain period or author.

II Subchapter of I chapter provides the information on the structure of "*The national language corpora on Germanic and Romance languages*". It was decided to create the first linguistic corpus obtained in the process of applying computer technologies to language in 1960. The first electronic corpus - The Brown Corpus included about 500 texts from American newspapers, magazines and books. Each text included in Brown's corpus included 500 texts consisting of 2000 words, and a total of one

---

<sup>6</sup> Finegan, E. Language: its structure and use / E. Finegan. – Los Angeles: West Group, – 2004, – 607 p.

<sup>7</sup> McEnery T., Wilson A. Corpus linguistics. – Edinburg: – Edinburg University Press, – 2001, – 235 p.

<sup>8</sup> Mahmudov, M.Ə. Türk dillərinin milli korpusu / M.Ə. Mahmudov. – Bakı: – Elm və təhsil, – 2018, – 392 s.

million words<sup>9</sup>. The British National Corpus is a corpus consisting of more than 100 million word-forms collected from numerous sources, written and spoken language samples, designed to represent a large part of the English language since the end of the XX century<sup>10</sup>.

The Corpus of Historical American English COHA contains more than 400 million words of text covering the years 1810-2000. It is 50-100 times larger than other comparable historical corpora of the English language.

The purpose of creating separate corpora for American and British English is not to confuse the differences in these languages and to study the materials of each language separately.

The national corpus of French language, a member of the Romance language family, Frantext is a database of 5,415 references or 254 million words, developed at ACPFL (Analysis and Computer Processing of the French Language) and placed online since 1998. This corpus provides forms, lemmas, and grammatical categories, simple and complex searches, and displays results in 700 character contexts<sup>11</sup>.

Spanish, Italian, Romanian, and Portuguese languages belonging to the Romance language family also have corpora, but the corpora of these languages repeat each other systematically and structurally.

As can be seen from the examples, national corpora of Romance languages do not have as rich structure as national corpora of Germanic languages. This can be attributed to the fact that Germanic languages are more widespread and have a high degree of functionality.

III Subchapter of I chapter deals with "*The national language corpora of Slavic languages*". Until recently, a corpus that could be

---

<sup>9</sup> Захаров В.П. Корпусная лингвистика. Учебно-методическое пособие. Санкт-Петербург, Санкт-Петербургский государственный университет, 2005, 48 с.

<sup>10</sup> <https://www.english-corpora.org/bnc/>

<sup>11</sup> <https://www.frantext.fr/>

used by researchers all over the world did not exist in the Russian language. However, the creation of a corpus in Russian linguistics expanded the research area and boundaries of linguistic science. At the same time, radical changes took place in the methodology and practice of language research.

The great interest and support for the creation of the national corpus of the Russian language arose from the demands placed on linguists from a modern point of view. It is not surprising that during those times, not only in Russia, but also in many other countries of the world, work was started in the field of creating the corpus of the Russian language. For example, in 1999, the corpus of the Russian language began to be created at the University of Tübingen in Germany. The Tübingen corpus is based on the Upsal corpus. Therefore, the corpus can be used online. Its volume consists of one million words and is composed of 60 different text fragments. Artistic and journalistic style is based on those text fragments. The literary texts are taken from works written by 40 authors in 1960-1988. Publicistics includes newspaper texts related to various thematic areas in 1985-1988. The second subcorpus in TK includes interview texts<sup>12</sup>.

In the IV subchapter of I chapter, the information is provided on "*The national language corpora of Turkic languages*". In the issue of the compilation, structure and placement of electronic dictionaries, which are one of the important components of national language corpora, different approaches are manifested not only in non-kindred system languages, but also in the same language families. As it is known, back in 1988, the idea of creating a machine fund of Turkish languages in the territory of the former USSR was put forward and the main directions of its creation were determined. At that time, the Institute of Linguistics of the Kazakhstan Academy of Sciences was recommended as the center for the creation of a machine fund of Turkic languages<sup>13</sup>. In this

---

<sup>12</sup> <http://www.sfb441.uni-tuebingen.de/bl/rus/korpora.html>

<sup>13</sup> Жубанов, А.К. Казахское языкознание: прикладная лингвистика / А.К.Жубанов. – Алматы: «КИЕ», - 2012, - 696 с.

subchapter, the processes and structure of the national corpus of Turkic, Kazakh, Bashkir, and Tuva languages, which belong to the family of Turkic languages, have been analyzed.

V subchapter I chapter deals with "*The works relating to the national corpus of the Azerbaijani language*". The issue of using mathematical-statistical methods and new technological tools in the study of the Azerbaijani language began in the second half of the last century. At that time, in many leading centers of the world, the interesting researches were conducted in the field of language learning with mathematical and statistical methods with the help of electronic computers. In many scientific centers of the former USSR, including Moscow, Leningrad (now St. Petersburg), Kiev, Minsk, etc. There was great interest in such studies in their cities. The first researchers working in the field of language research with new methods and technologies in Azerbaijan were experts who received post-graduate studies or had scientific experience in the aforementioned science centers.

The meeting of different fields of science, the creation of a new scientific direction and the emergence of new fields of science based on one science are the leading aspects of the development process of sciences in modern times. Many researchers point out that there is a mathematical order and regularity in the structure of Turkic languages, including the Azerbaijani language, in the mechanism of joining grammatical and lexical suffixes to word roots.

The emergence of new fields of science as a result of the convergence of different fields of science or the division of the same field of science in Azerbaijani linguistics can be considered as a legitimate result of the influence of similar events in world science. As an example of what we said, A.A. Akhundov's article "From the experience of statistical analysis of Azerbaijani language vowels" published in Scientific Works of ASU (now BSU) in 1963 can be cited.

For the first time in Azerbaijani linguistics, the author of the article informs the readers about the results of his research in the

field of analysis of Azerbaijani vowels using statistical methods. It was an attempt by the researcher to apply statistical methods in the study of the vowel system of the Azerbaijani language and an attempt to study vowels in a different way, which we are not used to until now, from a new point of view. Later, A.A. Akhundov expanded his research in this direction and tried to arouse interest in this new field of linguistics with the book "Mathematical Linguistics".

The first studies on corpus linguistics in the Azerbaijani language belong to J. Rahmanov. The author dealt with the problem of Machine fund adopted in the former USSR. The researcher studies the corpus problem in a slightly different way. In his works related to the establishment of the Turkic language machine fund, the issues of building different corpora within the fund are raised. It shows that one of the initial tasks of the process of creating a machine fund of Turkic languages is the inclusion of monuments and various dictionaries related to Turkish languages into the fund. From this point of view, terminological dictionaries include separate field terms and they include a larger number of lexical units<sup>14</sup>. The preparation work in the field of the national corpus of the Azerbaijani language was carried out within the framework of the national corpus of Turkic languages. These works can be considered the initial stage of the creation of the national corpus of the Azerbaijani language.

The last subchapter of I chapter I is called "*The use of language corpora in teaching*". Modern society is characterized by the transition from information to knowledge. It is necessary to respond to the demands of the information age. Such requirement is the general computerization of humanitarian knowledge and the use of computer technology in linguistics. Teachers and students often have to deal with a large amount of literature related to computers and information technologies. Currently, the most interesting is the educational system characterized by information management.

---

<sup>14</sup>Rəhmanov C. Türk dillərindən bir-birinə avtomatik tərcümənin ümumi məsələləri haqqında // Terminologiya məsələləri. Bakı: Elm, 2012, s.160-168.



The computer linguistics is already being taught in higher education institutions of many developed countries. In most countries of the world, the corpus is used not only for conducting research, but also in the teaching process.

Accordingly the national corpus of the Azerbaijani language should be further improved based on the world experience, and it should be started to be used in order to give an effective result in the teaching process in a short time.

Chapter II of the dissertation is called "*Structural differences and peculiarities of national language corpora*". I subchapter of this chapter analyzes "*General issues of the national language corpus*".

As it is known, language corpora are created with certain goals. Language corpora include not only linguistic issues, but also literary studies, logic, stylistics, linguistic statistics, educational issues, etc. targets can be selected as research objects. There are corpora that are clearly created for educational purposes, and the search systems built in them are educationally directed. As it is known, corpus linguistics includes issues of studying the languages of individual writers and poets from a linguistic-statistical and stylistic point of view. For example, concordances can be characterized as one of the components of national language corpora. In our research, the field of corpus linguistics related to lexicography was taken as the basis. The research examines the placement of dictionaries in national language corpora and their effective use.

In II subchapter "*Structural features of English-language corpora*" are analyzed. As we know, there are different variants of English language. The peculiarities of the English-language corpora are that the corpora of each version of the English language have been created and studied separately.

Representing British English, the British National Corpus is a monolingual corpus. That corpus covers modern British English, not other languages spoken in Britain. However, non-British English and foreign words can also be found in the corpus. It

includes texts written in a variety of styles and is not restricted to any subject area, genre or register. COCA is a corpus that represents the characteristics of modern American English.

Oral and written language patterns were also included in the British national corpus. The written part of BNC includes academic books and popular works of art covering all ages and interests, published and unpublished letters and memoranda, selected materials from regional and national newspapers, special periodicals and magazines related to teaching, and many other types of texts. The oral part includes interviews (taking into account different age levels and regions), demographically balanced materials from social classes and spoken language collected in different contexts. Complete technical documentation covering all aspects of BNC, including its design, development and content, is provided by the Reference Guide to BNC (XML Edition version).

II Sub-chapter of II chapter is called "*The leading aspects of national corpora created on Slavic languages*". A special information research system designed to work with corpora was developed by Russian linguists. That system allows working with different types of marked corpora as well as concordances. With the help of such a system, it is possible to communicate with individual components of the query system related to the corpus. This system is primarily intended for philologists, but those who want to readily use the results of the study of this or that corpus of texts can also benefit from it. That system is designed to conduct independent analysis and obtain new information. These data include the information about the alphabet and frequency dictionaries of this or that work, their units, words separated by different signs, etc. The first corpus material prepared with the help of this system is the text corpus "A.S.Pushkin's dramaturgy and poetry". All poetic and dramatic works of Pushkin are included in this corpus. Those works were prepared on the basis of the poet's academic publication<sup>15</sup>. One of such data research systems is the grammatical-semantic frequency dictionary of the language of A.P.Chekhov's literary

---

<sup>15</sup> [http://www.philol.msu.ru/~lex/pdfs/kiisa\\_buklet.pdf](http://www.philol.msu.ru/~lex/pdfs/kiisa_buklet.pdf)

works. The dictionary was prepared on the basis of 600 texts, which includes the analysis of all completed prose and dramatic works of the writer. 17 of them are plays and 583 are prose works. Therefore, this dictionary can be considered the first lexicographic work reflecting the language and lexical composition of Anton Pavlovich Chekhov's literary works. The volume of the dictionary consists of more than 36 thousand lexical units. "Electronic corpus of A.P. Chekhov's artistic works" was taken as the basis for writing the dictionary in the laboratory. That corpus was prepared on the basis of the complete collection of A.P. Chekhov's works and letters<sup>16</sup>.

The statistical dictionary of F.M. Dostoyevsky's language was created by V.M. Andryushenko, A. Y. Shaykevich, N.A. Rebechkaya at the Russian Language Institute named after V.V. Vinogradov of the Russian Academy of Sciences. The dictionary consists of 94 tables. The numbers of the tables correspond to corpus texts or corpus subcorpora (fiction, criticism and journalism, letters). On the left side of the tables, linguistic objects - word-lemma, graphic word, grammatical form, word-correcting elements are sequentially located, and on the right side there are statistical indicators. Statistical indicators, as a rule, show absolute frequency. Some tables show the relative frequency<sup>17</sup>.

The national corpus of the Russian language primarily covers the period from the beginning of the XIX century to the beginning of the XXI century. At that time, the language was represented in various sociolinguistic variants - literary, colloquial, dialect, folk talk. The corpus includes original versions of fiction (translated versions are not considered), prose, drama, poetry. It should be kept in mind that these works should be interesting from the point of view of language and should be of great cultural importance. In addition to fiction, other writing patterns (at the present time, oral speech patterns) can be included in the corpus. Other examples of

---

<sup>16</sup>[http://www.philol.msu.ru/~lex/pdfs/slovar\\_yazyka\\_proizvedenij\\_chehova2012.pdf](http://www.philol.msu.ru/~lex/pdfs/slovar_yazyka_proizvedenij_chehova2012.pdf)

<sup>17</sup> [http://cfrl.ruslang.ru/dost\\_cd0/descrip.htm](http://cfrl.ruslang.ru/dost_cd0/descrip.htm)

writing are memoirs, essays, journalism, public and scientific speeches, personal correspondence, diaries, documents, etc.

IV subchapter of II chapter deals with "*The differences and similarities of national corpora created for Turkish languages from other language corpora*". The National Corpus of the Turkish language includes the corpora of the Turkish language from nine different areas, the years between 1990 and 2010 the all types of "books, periodicals, various published texts, various unpublished texts"; "social sciences, arts, commerce and finance, thought and faith, world problems, applied sciences, natural and basic sciences, etc."; "gender of the author (female, male)", "author, type of authors (many, organizational, individual)", readership (children, youth, everyone) etc. The users can limit or expand the search according to the features they want using the options that contain the information <sup>18</sup>.

"The formation of a special vocabulary block in the national corpus of the Kazakh language is set as a goal. This block includes a statistical dictionary made on the basis of M. Auezov's 20 volumes and a 10 volume explanatory dictionary, as well as a grammatical dictionary, frequency dictionaries of various genres and other types of dictionaries. It is planned to place electronic versions of all existing terminological dictionaries.

During the compilation of the "Frequency Dictionary of the Kazakh Language", the word-form was taken as a linguistic unit. Taking this language unit as a basis will give Kazakh researchers ample opportunities to follow the process of studying the graphic system and changing words.

The general dictionary contains more than 36 thousand lexical units and is divided into 3 parts. In the first dictionary, the frequency of word-forms is arranged in alphabetical order, their color, absolute frequency and the number of texts in which they are found are signified <sup>19</sup>.

---

<sup>18</sup> <http://www.dam.org.tr/index.php/tr/derlemeler/66-soezlue-tuerkce-derlemi>

<sup>19</sup> <https://tbi.kz/ru/frequency-dictionary>

The machine fund of the Bashkir language consists of subfunds that include seven databases. Furthermore, the corpus folklore forms a separate block in the national corpus of the Bashkir language. Moreover, in addition to the corpus of the Bashkir language, the machine fund of the Bashkir language includes subfunds (prose works, journalism, folklore). Dialectological base, lexicographical base, grammatical base, experimental phonetic base are placed in the subfunds. Epics are included in the folklore corpus of the Bashkir language, too.

The last III chapter of the dissertation is called "*Dictionary block of national language corpora*". I subchapter of this chapter defines and analyzes "*Block of dictionaries as a leading component in national language corpora*". The history of the lexicology science has a deep story. The ways of expression of language, which is a means of communication, are words, the richer they are, the clearer and more precise the idea to be expressed. Lexicography has an important role in creating interlinguistic and intercultural communication.

The creation of dictionaries is not an easy process. In computer dictionary, these processes are solved in a short time. Compilation of dictionaries is carried out under the name of computer lexicography, computational lexicography, machine lexicography, automatic lexicography terms.

The author of the book "Basics of Computer Linguistics" Y.N.Marchuk considers all of these terms to be the subject of computer linguistics, states that they are actually synonymous terms with the same meaning<sup>20</sup>.

Based on this opinion of Y.N. Marchuk, we can say that the object of research of these listed terms are computer dictionaries.

*"Computer dictionaries have several advantages:*

- *using dictionaries is simpler and faster;*
- *to understand the exact meaning of the word, it is possible to refer to several dictionaries at the same time;*

---

<sup>20</sup> Марчук, Ю.Н. Основы компьютерной лингвистики // Ю.Н.Марчук. - М.: МГОУ, - 2002, -234 с.

- it is easier to highlight any language and compare it with other dictionaries;

- if it was necessary to publish a book again to make changes in ordinary dictionaries, it is possible to add, remove or make corrections to new words in computer dictionaries.

*Computer dictionaries even have additional information - a picture of the object represented by the word, a sound variant, etc. it is possible to place*<sup>21</sup>.

In our opinion, the dictionaries form the core of corpus linguistics. Therefore, the corpus shows the meaning of the words in the texts placed in it, which part of speech they belong to, etc. and as a result of automatically marking the words in that text for learning, it is given the opportunity to get all the information about that word. The most effective way to do this is to apply concordances to linguistic corpora.

II subchapter of III chapter is called "*Dictionaries in the national corpus of the Azerbaijani language*".

The idea of creating a national corpus of the Azerbaijani language has a long development path. The researches conducted in the field of studying the Azerbaijani language with precise mathematical-statistical methods and modern technologies can be considered the first rudiments of this. Conducting scientific researches in this field in the world once again proves the necessity of creating a national corpus.

First of all, it would be important to establish the function and activity of the vocabulary subcorpus in the national corpus of the Azerbaijani language. For this, electronic versions of each of the dictionaries belonging to the Azerbaijani language should be prepared. Along with placing those dictionaries in the lexicography subcorpus, establishing a relationship between them is also an important condition.

---

<sup>21</sup> Вадяев, С.Е. Электронная лексикография и корпусная лингвистика // Аспекты становления и функционирования западногерманских языков, - Самара: Изд-во «Самарский университет», - 2003. - с.83-92.

The corpus should include texts that contain the language style features of the Azerbaijani language. The more texts included in the corpus, the more comprehensive and accurate the results obtained during the search. Obviously, it would be more useful to use optimal placement methods here. The subcorpus forms a specific request to a specific text array and creates an opportunity to receive information from it.

If any national language wants to prove its presence, exist and gain prestige in the globalized world, it must be represented on the Internet. Now the world uses and benefits from the Internet more and more as a source of knowledge.

The Azerbaijani language, which is the state language, must also prove its existence, availability, and reputation through wide and active representation on the Internet.

III subchapter of III chapter is called "*Monolithic subcorpora*". The purpose of monolingual dictionaries is to learn one or another language in depth. When we say monolingual corpora, we mean any artistic work, pattern of oral folk literature, etc. The difference is that a monolingual corpus must be able to use that text for linguistic research. A corpus is conceived as a reduced analogue, model of a language or sublanguage. That is, corpus text is not just a collection of texts, but texts collected in an electronic version with certain regularity (certain principles).

Monolingual corpora are especially important in terms of checking the correctness of translated sentences, the possibility of being observed in the communication process. In the process of speech recognition, monolingual corpora are very useful as a language model in specifying words that are not recognized correctly due to incorrect pronunciation. It should be taken into account that the processing frequency of the incorrectly pronounced words will be less than the correct variant of the sounds found in the monolingual corpus. For example, if the word "result" was mistakenly given as "result" or "consequence" in the sentence "You will have a result", the correct version of this word can be found with the help of a monolingual corpus. Monolingual corpora also

allow to correctly identify syntactic units in translated sentences. The corpus of monolingual texts is indispensable in the process of accurately defining and explaining common and special nouns in a sentence and correctly placing punctuation marks. The task of corpus creators is to collect as many texts as possible about the language to be studied.

IV subchapter of the III chapter is called "*Bilingual subcorpora*". Developed within the framework of the "Dilmanc" project in Azerbaijan, bilingual parallel text corpora covering all styles of the language and large-scale monolingual corpora representing any specific language should be noted, Bilingual parallel text corpora have been launched both as dictionaries and as translation software.

Bilingual corpora development has many peculiarities. As is known, bilingual corpora are mainly formed on the basis of electronic resources translated into another language (books, magazines, newspapers, official documents, news portals and other websites).

V subchapter of III chapter is called "*The subcorpus of translation dictionaries*".

A parallel corpus consists of source texts and their translations. In scientific literature, they are also called translation corpus. They are most often used in translational and comparative studies. According to the direction of translation, these corpora can be unidirectional, bidirectional (also called reciprocal) and multidirectional. A unidirectional corpus contains only translations of texts from one language to another, not vice versa.

In the bibliography of Azerbaijani language dictionaries on the Internet, in addition to "Dilmanc", Polyglot, AzerDict, Google Translate, etc. electronic translation dictionaries can also be found.

The "Polyglot" dictionary system, which has won the deep sympathy of users, was developed within the framework of the project for the development of Information and Communication technologies in the Azerbaijan Republic.



Sub-chapter VI of III chapter is called "*The subcorpus of statistical vocabularies*". Statistical dictionaries play a fundamental role in studying the historical development path of the language and the changes that have occurred. By using statistical dictionaries, the frequency of use of the searched word in the work, the work of linguists who want to study the historical changes it has undergone, becomes quite easy.

Azerbaijani linguistics has accumulated rich experience in the field of preparation of statistical dictionaries. In the 1970s, for the first time in Azerbaijan, experiments were carried out in the field of automation of linguist-researcher activities, the word-form frequency and alphabet-frequency lists based on J. Jabbarli's works was prepared.

New technological tools are successfully applied in the study of ancient written monuments of Azerbaijan. A lot of work has been done in this field in Azerbaijan, and research in this direction is continuing and expanding. Work in this area was based on the experience previously gained in the compilation of frequency and counter dictionaries.

In the last subchapter of III chapter, "*The ways of placing other vocabularies (spelling, spelling, phraseology, dialectology, abbreviations, emphasis, etc.) in the corpus*" are shown. Monolingual, bilingual, multilingual, explanatory, frequency, terminological, phraseological, synonym, antonym, homonym, paronym, onomastic, dialectological, orthographic, orthoepic, historical, etymological and other dictionaries on the Internet is no longer considered a novelty for world science. Many works have been done in Azerbaijan in the field of placement and use of separate dictionaries in the electronic space, and work in this direction is ongoing.

As new electronic dictionaries are included in the dictionary block of the national corpus of the Azerbaijani language, as the scope expands accordingly, as the structure is improved, it will be possible to get more complete and characteristic information about any word (phrase), lexical unit, lemma. The block of dictionaries of the

national corpus is supposed to function as an open system. This means that electronic versions of new dictionaries can be added there at any time, or dictionaries already placed there can be revised and improved.

The results obtained in the research process carried out in the dissertation can be grouped as follows:

1. The block of dictionaries is one of the most important and leading components in the national corpus of the Azerbaijani language. Unlike other components, it is the dictionaries block that users refer to the most.

2. From the research, it is possible to come to the conclusion that the dictionaries should be represented in all their varieties in the dictionary block. A system should be developed so that all the lexical-grammatical, lexicographic features of the searched word can be fully presented to the user.

3. The block of dictionaries of the national corpus of the Azerbaijani language is an open system. If necessary, electronic versions of newly compiled dictionaries can be added there, or electronic dictionaries already placed there can be revised and improved.

4. All electronic dictionaries that make up the block of dictionaries must function as parts of a single system, have a common search system and rules of use.

5. Ensuring that each vocabulary unit in the vocabulary block is provided with its context is one of the important conditions. The multiplicity (richness) of the context examples of vocabulary units creates the possibility of a more correct perception and understanding of the meaning of the word.

6. The corpus of bilingual and multilingual texts should be included in the dictionary block of the national corpus of the Azerbaijani language. The presence of bilingual and multilingual corpora can greatly facilitate the work of users.

7. We believe that information related to any word (vocabulary unit) searched for in the dictionary block should be provided to the user by taking and adding it from all the dictionaries represented in

the corpus. This can be considered acceptable in terms of completeness and standard of information.

8. One of the most important components of national language corpora should be concordances. Concordances of the most common, well-known historical monuments, works of individual writers should be placed in the dictionary block.

9. The concordances placed in the dictionary block should be presented to users both separately and in a combined version. In the future, after concordances covering individual works and the language of writers are combined, they should be able to represent the literary language of Azerbaijan as a whole.

10. It would be very important and effective to have the possibility of communication and mutual exchange of the block of dictionaries in the national corpus of the Azerbaijani language with the national language corpora of other kindred and non-kindred languages, establishing relations at the level of experts, organizing online discussions and seminars.

**The main content of the research work is reflected in papers and articles published below:**

1. Kompüter lüğətçiliyi leksikoqrafiyada yeni mərhələ kimi // Ümummilli lider H.Əliyevin anadan olmasının 93-cü ildönümünə həsr olunmuş Gənc tədqiqatçıların IV Beynəlxalq Elmi konfransı, – Bakı: QU, – 29-30 aprel, – 2016, – s.970-972.

2. Azərbaycan dilinin tədqiqində yeni istiqamətlər // Akademik T.Hacıyevin anadan olmasının 80 illik yubileyinə həsr olunmuş “Azərbaycan Filologiyası: inkişafın yeni mərhələsi” mövzusunda Respublika elmi konfransı, – Bakı: BDU, – 2 noyabr, – 2016, – s.62–64.

3. Azərbaycan dilinin riyazi-statistik metodlar və yeni texnoloji vasitələrlə öyrənilməsi məsələləri: problemlər, perspektivlər // – Bakı: Tədqiqələr, – 2016. №1, – s.18-28.

4. Milli korpusun yaradılmasında dünya təcrübəsi // Akademik A.Axundovun 85 illiyinə həsr olunmuş “Ağamusa Axundov və

Azərbaycan filologiyası” mövzusunda beynəlxalq elmi konfrans, – Bakı: AMEA, – 24-25 aprel, – 2017, – s.448-450.

5. Türkiyə milli korpusu // Ümummilli lider H.Əliyevin anadan olmasının 95-ci ildönümünə həsr olunmuş tələbə və gənc tədqiqatçıların “Gənclər və elmi innovasiyalar” mövzusunda respublika elmi-texniki konfransı, –Bakı: AzTU, I hissə. – 3-5 may, – 2018, – s.501–502.

6. Türk dilləri üzrə milli korpuslar // “Müstəqillik illərində üsulların inkişafı və dilin lüğət tərkibinin zənginləşmə istiqamətləri” mövzusunda respublika elmi konfransı, – Bakı: – 25-26 dekabr, – 2018, – s.224–230.

7. Azərbaycan dilinin elektron lüğətləri // “Müstəqillik illərində üsulların inkişafı və dilin lüğət tərkibinin zənginləşmə istiqamətləri” mövzusunda respublika elmi konfransı, – Bakı: – 25-26 dekabr, – 2018, – s.231–235.

8. Dictionary block of the national corpuses of the Turkic languages // – Тольятти: Балтийский гуманитарный журнал, – 2019. Том 8, №1 (26), – с. 103-107.

9. Dil korpuslarında elektron lüğətlərin verilməsi üsulları // – Bakı: Filologiya məsələləri, – 2019. № 4, – s.168-179.

10. Türk dilinin milli korpusunda leksikoqrafik bölüm // – Bakı: Filologiya məsələləri, – 2019. № 5, – s.148-154.

11. The issues on the optimal structure of lexicography block in national language and processing of its software // – Paris: Advances in Economics Business and Management Research (Materials of “1st International Scientific Conference “Modern Management Trends and the Digital Economy: from Regional Development to Global Economic Growth” MTDE 2019”, Yekaterinburg: Russia, –14-15 April, – 2019), – 2019. Vol. 81, – p. 802-805.

12. Afad Qurbanovun tətbiqi dilçiliyin tədrisində rolu // Tanınmış dilçi alim, türkoloq, ictimai xadim, professor Afad Qurbanovun anadan olmasının 90 illiyi münasibəti ilə respublika elmi konfransı, – Bakı: – 24 may, – 2019, – s.138-142.

13. Konkordanslar korpusu dilçiliyinin mühüm sahəsi kimi //

Tanınmış türkoloq alim Məhəbbət Mirzəliyevanın 70 illiyinə həsr olunmuş “Azərbaycan dilçiliyinin aktual problemləri” mövzusunda respublika elmi konfransı, – Bakı: Unicopy, – 28 oktyabr, – 2019, – s.188-191.

14. Azərbaycan dilinin milli korpusunda altkorpuslar // – Bakı: Journal of Baku Engineering University. Philology and pedagogy, – 2019. Vol.3, № 1, – s.32-43.

15. Dictionaries block as a major component of national language corpus // Актуальные проблемы гуманитарных и общественных наук: сборник статей V Всероссийской научно–практической конференции, – Пенза: РИО ПГАУ, – 2019. – s.80-83.

16. Subcorpus of statistical dictionaries in national language corpuses // Trends in the development of modern linguistics in the age of globalization: materials of the V international scientific conference, – Prague: Vědecko vydavatelské centrum «Sociosféra-CZ», – on October 17-18, – 2019, – p. 20-23.

17. The concordances in the researches of the language // – Bishkek: Alatau academic studies, – 2019. №3, – p.46-56.

18. Dil korpuslarının tədrisdə istifadəsi // Bakı: Bakı Avrasiya Universiteti, “Sivilizasiya” elmi-nəzəri jurnal, – 2019. Cild 8, № 4, – s. 145-154.

19. Milli dil korpusları və dil siyasəti //Second international scientific conference of young scientists and specialist, – Bakı: ANAS, – 3-6 March, – 2020, – s.438-439.

20. Milli dil korpuslarının dilin inkişafında rolu // – Bakı: Bakı Qızlar Universiteti, Elmi əsərlər, – 2020. Cild 11, № 4, – s.49-53.

21. Azərbaycan dilinin milli korpusunda birdilli və ikidilli altkorpuslar // “Çağdaş dövəmdə türk dünyasının aktual problemləri” adlı onlayn respublika elmi konfransı, – Lənkəran: LDU, – 24 dekabr, – 2021, – s.176-178.

The defense of the dissertation will be held on 19 september in 2023 at 11:00 at the meeting of the Dissertation Council ED 1.06 operating under the Linguistics Institute named after Nasimi of the Azerbaijan National Academy of Sciences.

Address: AZ 1143, H.Javid Avenue 115, V floor, Linguistics Institute named after Nasimi of ANAS.

The dissertation is available in the library of the Linguistics Institute named after Nasimi of the Azerbaijan National Academy of Sciences.

Electronic versions of the dissertation and abstract are posted on the official website of the Linguistics Institute named after Nasimi of the Azerbaijan National Academy of Sciences.

The abstract was sent to the necessary addresses on 16 june in 2023.

Signed: 12.06.2023  
Paper format: 60x84 16<sup>1</sup>  
Capacity: 40, 854 characters  
Printing: 20 copies