UDC 81-139
SRSTI 16.21.47
https://doi.org/10.48371/PHILS.2023.70.3.001

METHODS OF AUTOMATION OF TEXTS IN THE KAZAKH NATIONAL CORPUS

*Amirbekova A.B.[1] , Konyrova A.T.[2], Kaiyrbekova U.S.[3]
[*1]candidate of philological sciences, head of the lexicology department Institute of Linguistics named after A. Baitursynov, Almaty, Kazakhstan, https://orcid.org/0000-0001-8540-8264,
e-mail: marghan01@mail.ru,
[2]candidate of philological sciences, al-Farabi Kazakh National University, Faculty of Pre-Higher Education, Senior Lecturer, Head of the Department of Language training and General Education of Foreigners, Almaty, Kazakhstan, https://orcid.org/0000-0001-6996-6522,
e-mail: akbota74@mail.ru,
[3]candidate of philological sciences, associate professor, head of the department of philology, Peoples' Friendship University named after A. Kuatbekov, Kazakhstan, Shymkent,
https://orcid.org/0000-0002-9466-024X

Abstract. In the era of language globalization, when the lexicographic base becomes fully accessible in the digital system, it becomes possible to optimize language acquisition. The national Corpus of the Kazakh language is a digitized version of the Kazakh word. Since the Kazakh language corpus is a system of linguistic knowledge, on the basis of which it consists of several subcorps, the demand for the corpus is growing day by day. This is due to the fact that Kazakhstan is a multinational state. Therefore, representatives of other nationalities who consume Kazakh culture want to determine the translation equivalent. The corpus will also become an effective linguistic base for language learners.

The purpose of the article is to semanticize the lexicographic base included in the training corpus, especially the words of the lexical layer of the Kazakh language and adapt to automation in accordance with the digital system. The article suggests the difference of the corpus from other subcorpuses, the classification of semantics and methods of interpretation (automation) of semantic groups. This is the scientific significance of the article.

The methods of content analysis, generalization, and description were used when introducing the lexical base into the training corpus. The scientific conclusions presented in the article are of practical importance, contributing to the development of the corpus, the development of electronic applications for language acquisition.

Keywords: National corpus, semantics, automation, language base, translation, lexical layer, educational building, digitalization of the language

Basic provisions

Digital education is one of the mechanisms for the implementation of the national project of the Republic of Kazakhstan "Educated Nation". Therefore, language acquisition and the formation of an electronic resource of linguistic culture and linguistic capital is a requirement of modern society. adapting language learning content to system search browsers requires the skill and skills of linguists. Therefore,

for an easy search for lexicographic information, it is advantageous to create an educational corpus.

Introduction

Training corpus is not only an electronic database of teaching texts according to the level of language learners, but also a linguistic system that shows the characteristics of the lexical layer of each word. Therefore, it simultaneously performs the linguistic and linguistic-didactic functions of the training corpus. A training corpus is an automated vocabulary that serves as an additional tool for language learners, especially for fixing incorrect usage. The reason for the interest in the development of the the learning corpus and its special research arises from the demand of the society. In modern Kazakh society, not only learners of the state language, but also in the Kazakh language environment, there is an effort to write, speak and construct sentences correctly, and use each word according to its meaning. From this point of view, we noticed that not only the community of writers, but the general public is sympathetic to the competent use and development of the Kazakh language. This is because, in the public demand, the explanation of the meanings of words, the translation of new words, the environment and status of the use of literary language and local or spoken languages, the norms of error-free writing and speech, the antonym, synonym, homonym character of words, figurative use and the meaning of the phraseological fund in another language etc. is gaining priority. The language will be an affordable and optimal language tool that meets the linguistic needs of consumers i.e. a training corpus. Therefore, we are dedicating the article to the training corpus, which is presented as an electronic set of linguistic lexical and normative fund for the needs of society. In the article, we focus on the study of the training corpus in foreign and domestic linguistics, its main structural system, ways of developing linguistic content for browsers.

Materials and methods

Standard, bilingual and explanatory dictionaries of the Kazakh language are taken as a material base for the systematic introduction of the lexical layer of the Kazakh language into the training corpus. We classified them into three groups:

1. Material fund - normative fund. The spelling and orthoepic dictionaries are considered as a normative fund. The entire normative dictionary is not included in the training corpus. The words are sorted according to the frequency of errors, that is, the words of the loose norm or the writer's frequent mistakes are selected. To determine this situation, the accumulation method was used. A series of frequently asked words was created due to the error in the question-and-answer section on sites and social networks, which will be a tool for teachers. After that, by interviewing elementary and Kazakh language teachers or using a survey method, frequent mistakes made by students are determined. Summarizing these studies, a chain of difficult-to-spell words is created using an empirical research method. As a result of all the research, the pragmatics of the word sequence collected, that is, the searcher's convenience, is thought out.

2. Material fund - lexical fund. This activity of the training corpus is based on the lexical system: native words/ loanwords; old/new words; literary/non-literary (speech, vulgar word (expression); curse word) words; term/local, professional, historical words;

3. Material fund - semantic fund. This activity is based on the semantic system: homonyms, antonyms, synonyms, polysemy, sequence of eurysmy, sequence of phraseological and figurative words.

4. The material fund is the lexical-semantic minimum of three languages according to the level of language learning. It includes a sequence of words (literary words, terms, phraseological units, proverbs, etc.) and a network of texts.

During the development of this material fund, differentiation, selection method, collection, analysis method, content analysis method, statistical method, philological examination method, marking methods are used.

Literature review. It is true that there were a number of options during the preparation to the development of the training corpus. In order to optimally and fully satisfy the mass demand, there was not enough text network for the training corpus. Therefore, we had to rely on the lexical-grammatical minimum.These conclusions are based on the experience of foreign scientists. The Cambridge Learner Corpus is considered the first training corpus. ICLE (The International Corpus of Learner English), LLC (Longman Learners' Corpus) was subsequently launched.

The more search engines there are in the structure of the training corpus, the higher the ranking, that is, the corpus with more demand for service. In 1990, the construction of the training corpus in Belgium was recognized as the most convenient among the educational buildings serving in Asia and Europe. This is the International Corpus of Learner English (ICLE). New texts, i.e. essays on new knowledge of students written in dominant languages of international relations, were introduced there [1, p. 78].

The compiled later training corpus was based on comparative vocabulary textbooks. That is, bilingual discourse, phrasebook, bilingual dictionary of antonyms, etc. are designed according to the needs of language learners. It includes a wide range S. Grange's training corpus. The Grange corpus consists of a collection of consumer essays in 14 national languages. The Hong Kong Study Corps prioritizes the writing work of English language learners. Educational corpora aimed at language teaching include Michigan Corpus (Michigan Corpus of Academic Spoken English (MICASE)), Tamper Corpus (ELFA (English as Lingua Franca in Academic Settings)). Anna Mauranen (Anna Mauranen, 2009) is the developer of the mentioned electronic language tools [2, p. 16].

According to O.N. Kashmilova's research: “The advantage of the training corpus for teaching a foreign language is not only to find texts and words suitable for the language level and topic. Tools for working with the text corpus (frequency list, concordance, parser) allow not only to identify errors in the lexical layer, but also to create important dictionaries for language learners, to identify actual problems of grammar, and to recognize unresolved language laws. Because the training corpus is a linguistic tool that serves the real process in practice” [3, p. 42].

According to S. Grange and N. Nesselhauf, the criteria for creating a training corpus depend, first, on the research objectives of the compiler, and second, on the limitations of the data set. Their proposed training corpus is divided into 7 cells:

1) learner corpus;
2) corpus of scientific and technical texts;
3) corpus of methodological texts;
4) corpus of artistic texts;
5) oral dialogue corpora;
6) corpus of personal letters;
7) corpus of journal articles [4, p. 117].

According to E. P. Sosnova: “The main source of the organization of the training corpus is a set of materials that are the basis for their analysis to determine the methods and effectiveness of learning the target language” [5, p. 129].

V.V. Rykova says: “The reading corpus is very important to show the mistakes in the written works of language learners, to see the vocabulary, the frequency of frequently used words, to find out which style the vocabulary belongs to” [6, p. 51].

After analyzing the conclusions of the researchers, we came to the conclusion that the training corpus is, firstly, a teaching tool for reading literacy, secondly, an electronic application for language learners, and thirdly, a text base adapted to the language level. Based on these studies, let's discuss ways to adapt the lexical layer of the Kazakh language to the training corpus and present the results.

Results and discussion

First of all, what information should be included in the training corpus of the Kazakh language? What will be the purpose of the corpus of the Kazakh language? What language tools are needed to learn the Kazakh language? As a result of reviewing the issues, we found the following:

- The training corpus of the Kazakh language should implement two goals. The first is the development of a search device that provides an optimal and affordable opportunity for Kazakh-speaking citizens to further improve linguistic literacy, that is, to find the correct word in doubt, to look at linguistic rules. The second is an electronic application for Kazakh language learners. It includes a dictionary, phrasebook, video, audio visuals, texts (topics) depending on the level of the language and topic.

The object of research of the article is the fund for improving lexical literacy for the Kazakh-speaking environment. In the Kazakh-speaking environment, illogical phrases and unsystematic sentences arise due to the change of word meaning and inappropriate usage. For example, in today's Kazakh society, older and younger people often use the phrase “qwıp ketti” (chased out - in the sense of saying whatever it is, deviating from the subject/topic of discussion, talking in vain, slurring of words). Of course, although it is recognized as slang registered among non-literary words, it has become an active unit of the language environment. In its place, questions arise as to what prompted the new concept of “chasing” to be translated into a new concept, even if it is a literary language version.

On the contrary, the fact that "the dog chased" and "chased the knowledge" are more active than the initial meaning of the physical movement shows the displacement of meanings. Whichever meaning is more active, in that society, in that era, that meaning dominates and the original one is forgotten. Therefore, the results of the process of semanticization of each word in the training corpus: initial meanings, synonyms, close synonyms, figurative synonyms, opposite synonyms, homonyms, opposites should be created on the same line, i.e. in such a way that they appear side by side.

Therefore, we need to study the phenomenon of meaning (semanticization) of words in our language in order to differentiate the linguistic information and its content loaded into the training corpus.

Semanticization (polysemay) is the acquisition of several meanings of a word, one of those meanings can become primary, active, variable, symbolic, ancient, new, obsolete, one meaing can be changed accidentally. That means the meaning of words is changed, completed, and destructed. As the word enters the semantization process, its meaning base expands. Due to the semantization base, the types of meanings also begin to increase. The language user, the people, shape the process of change of meaning. The process of semantization is caused by cognitive activities of a person such as thinking, recognizing, sensing, feeling, and imagining. The meaning process includes synonymy, antonymy, hyperonym, hyponym, enantiosemy, eurysemy, and polysemy.

Automation is the introduction of the system of linguistic knowledge into an electronic tool in the Internet space (application, site, search resource, etc.) in a way that is convenient for the public to search, and that is attractive, concise and informative.

Therefore, before automating the language fund in accordance with the demands of the modern society, it is necessary to identify the unresolved problems in the phenomenon of semantization. Because language learners are very important to the exchange of alternatives. The vocabulary of the Kazakh language should be equal to that of other languages. A lexicographic fund should be available to enable this. Therefore, if possible, there is a need to develop an automated bilingual or even trilingual lexicon.

Firstly, the parallel lexicographic fund of antonyms and words with opposite meanings is necessary for language learners. However, it is important first of all to compile a dictionary of antonyms of the Kazakh language. Although A. Zhumabekova's "Dictionary of antonyms of the Kazakh language" [7, p. 3] was published in 2000, a parallel dictionary of antonyms was not compiled. At the same time, the principles of meaning must be strictly observed. This is because antonyms require a gradation contrast that occurs from the lexical layer (not created in an artificial (by suffix) way). For example: alys-zhakyn (far-close), dos-kas (friend-enemy), on-teris (positive-negative), auyr-zhenil (heavy-light), tar-ken (narrow-wide), aura-sau (sick-healthy), etc. If we create a contradictory meaning in the nature of negation of a particular concept, this is the opposite meaning. The opposite meanings are formed from the derived word using the suffixes - syz, -siz, -ma, -me, -ba, -be. For example: akyldy-akylsyz (smart-crazy), kər-kərme (see-don't see),

zhagymdy-zhagymsyz (positive-negative). Distinguishing antonyms in this way in our language is understood as obeying the laws of meaningfulness. After the chain of antonyms and opposite meanings is separated and supplemented, work is carried out to register the equivalent in another language. The dictionary of parallel antonyms will be automated and built into one slot of the educational building.

Also in the Kazakh language, words with opposite meanings are used in the paremiological fund, in verse words for special purposes. Antonyms used with pragmatic interest for the purpose of comparing good and evil can also be uploaded to the training corpus. For example:

- in the form of a phrase: elgezek bala (a nimble boy) – tilazar bala (disobedient child); shaban at (a lazy horse) - zhurdek at (a fast horse); isker adam (business man) – olақ adam (a clumsy man), kisyk agash (a crooked tree) - tuzu agash (a straight tree); tattoo korshi (friendly neighbor) – araz korshi (a neighbor who is in a quarrel).

-in the form of proverbs: ashchy men tyshchyny tatkan biler, alys pen zhakyndy zhortkan biler (those who have tasted know what bitter and sweet is, those who have traveled know what far and near is); az soz - altyn, kөp soz – kөmir (less words-gold, more words-coal).

Secondly, the automated version of synonyms – synonymizer is especially important for language learners. The first version of the synonymizer in the Kazakh language was uploaded to the site Linguistic Search Sozdikqor [8].

Synonyms enrich a person's vocabulary range.

Synonyms are identical in meaning words. But in our language, the type of synonymous words has expanded. For example, the semantic types of the word *sotkar* (a pugnacious child) in the literary language can be distinguished as follows:

- lexical synonym: tentek, buzyk (naughty);
- regional synonyms: sotanaq, taytik, ausar;
- stylistic synonyms: shalduar, sodyr, ұrynshak, zhanzhalkoy, baukespe, akki.
- phraseological synonyms: қанды балақ (bloodthirsty), қызыл көз бәле (red

eye trouble), қитықтан ситық шығарғыш (making a scandal out of nothing), таяқтан саяқ шығарғыш.

Now we think that in the development of the full version of the synonymizer, it is necessary to be guided by a model that is distinguished by synonymic types:

Thirdly, if the illustrative explanatory parallel dictionary of phraseological units was automated, the complete information would be found by the language learner with one browser. Automation of phraseological units in the training corpus is developed according to the principle of visualization and equivalence.

Fourthly, the creation of the "Anonymizer" subcorpus. It is very important to distinguish homonyms from paronyms, puns and homographs. For that, it is very important to determine the way homonyms are created. Currently, there are 4 different ways of creating homonyms:

1) development of ambiguous words from the semantic point of view: Jal (1) - hair on the back of the animal's neck; Jal (2) is a long stretch, a ridge.

2) sound changes of words: Ep1 (yer - old Turkic, abbreviated form from erkek) - the gender is male. Er (2) is a harness. Er (3) is a verb meaning to follow someone.

3) affixation: Alma (1) (apple) is a fruit. Alma (2) is a verb. Baspa (1) is an institution that prepares printed products. Baspa (2) is a sore throat (angina). Baspa (3) is a place (canopy) with only the roof covered and open on all sides. Baspa (4) is a national dish, a type of a dried milk product. Baspa (5) is a verb.

4) words from other languages have the same pronunciation as the original words: Kydyr (1) (Khyzyr in Arabic) is a saint in the form of a person who travels around the country, brings happiness and wealth to people. Kydyr (2) - take a walk, walking.

It is necessary to distinguish homonymous words from multi-meaning words (polysemy) and broad-meaning words (eurysemy). This is because the process of transition from polysemy to homonym is taking place in the development of meanings. According to the results of semantic analysis in Kazakh language, it is said that broad-meaning words develop faster than homonyms and are more numerous. Is the principle of distinguishing homonyms from a polysemous word just that they belong to a different word class? Perhaps this principle is a definition used in explaining the concept of homonyms in elementary school. Among the nouns, there are many synonyms. To distinguish them, it is necessary to determine the legal principles of homonyms. Therefore, in the process of creating a "homonymizer", when words are sorted according to their meaning, the laws of the homonym will also be determined. This would be a consistent expression of the "from practice to theory" method. Texts related to the word "AT" (a horse) were collected using the electronic resource "National Corpus of the Kazakh Language" (https://qazcorpus.kz/) [9] of the Institute of Linguistics named after A. Baitursynuly. M. Belbaeva's "Dictionary of homonyms of the Kazakh language" states that the word "AT is homonymous in two senses [10, p. 19]:

1. AT - horse (racing horse);
2. AT - first name (His name (in Kazakh at) is Samat.);
3. AT - chess piece (term);
4. AT - fame, glory (aty zher zharady – a famous person);
5. AT - shooting from weapons (shoot into the sky, shoot with a gun).

And now it is possible to identify 5 homonymous meanings of the word "IT" (a dog) and to recognize each of them as a separate lexical unit.

These are five individual words. There is no semantic connection between these words, only the names are the same. Each of these words can create a polysemous word in itself. If a sequence of absolute homonyms is included in the educational corpus, their lexical meanings are loaded, and illustrative information is provided, it will undoubtedly be a useful electronic resource for students. In the process of automating the dictionary of homonyms for the learning corpus, it is necessary to take into account the efficiency of language learners. That is, it is quickly remembered with the versatility of visual presentation. For example:

In this direction, it is quite possible to transform the linguistic knowledge in the lexical layer of the Kazakh language into an electronic resource. The most

important principle is that it is important to consider the aspect that is easily accessible to language learners.

Conclusion

Automation of lexical units is carried out naturally with the help of a text. Therefore, it is very important to have a text corpus in the automation process. In education, the use of automated applications of the lexical layer, that is, online services, significantly contributes to the development of lexical and vocabulary skills, to the improvement of interpersonal relations during language learning, the development of emotionality, and creative abilities.

In order to automate lexicological knowledge, first of all, it is required to stabilize the laws in the process of gaining the meaning.

As we have seen, operations that automate lexical analysis models by computer programs require the implementation of a large amount of comprehensive linguistic information. Linguistic programs are formed as a result of automation. Automated linguistic resources to date can be grouped as follows:

1. Dictionaries and thesauruses. All their paper counterparts are electronic.

Forms are much more convenient to work with due to the automation of the search process. At the same time, it remains relevant to create versions of such resources equipped with real search tools.

2. Text conversion programs and text generators. They are supplemented with a linguistic fund that satisfies a specific search interest, and the system of possible errors allows for automatic correction in advance. Programs for automatic abstracting and annotation of texts can be attributed to the same group.

1. Text analysis and linguistic processing programs. This application allows you to check spelling, to accommodate transfers, check the text lexically and stylistically, and analyze the statistics of their errors. At the same time, one of its complex tasks is to computerize the change, transition, and complement of meanings: it is shown through lemmatization of the word (normalization, return to the original form). In addition, the manifestation of language automation is a source of plagiarism. Linguistic programs should include utilities to automatically match and classify texts. They allow to solve problems related to plagiarism detection.

2. Natural language processing systems. This group includes the most complex linguistic software. Of course, it is based on solving the problems described above, but more important goals are set: understanding the meaning of the text and giving a competent meaningful answer. In fact, this is the direction of development of artificial intelligence systems. The main application here is to create a natural language interface for computers.

In conclusion, during automation of lexicological knowledge in the educational corpus, first of all, it is required to permanently introduce the regularity of meaning. It is required to add a user-friendly graphical interface to the application features. It is required to be intuitive for non-IT users and easy for language learners.

The application should be distinguished by its versatility. It is important to refer to possible variants rather than being limited to a single search module. For example: according to the Kazakh spelling rules, is it nemkuraily or nemkuraydy

(indifferent)? The application that finds it must refer to the literary version of the doublet forms in the application. These are among the processing tasks. Therefore, the examples show the results of all the modules of the program.

Within the framework of the article, we will present the definition of the corpus, the criteria for the creation and organization of corpus, its application by foreign and domestic researchers, as well as the goals and possibilities of their application in the process of teaching foreign languages.

The article was investigated within the framework of the project BR 18574183 "Automatic recognition of the Kazakh text: development of linguistic modules and IT solutions".

REFERENCE

[1] Granger S. (ed). Learner English on computer. London: Addison Wesley Longman. 2010. 512.

[2] Mauranen Anna and Elina Ranta. English as a Lingua Franca: Studies and Findins. Newcastle: Cambridge and Scholars, 2009. 562.

[3] Kamshilova O. N. Inostrannye iazyki v distantsionnom obuchenii: materialy III mezhdunarodnoi nauchno-prakticheskoi konferesii, Tom 2 (Foreign languages in distance learning: materials of the III Intern. scientific-practical. conf. Vol. 2). Perm, 23-25, 2009. URL: http://window.edu.ru. – 42-49. [In Rus.]

[4] Granger S. The computer learner Corpus: a versatile new source of data for SLA research // Learner English on Computer. L., 1998. 894.

[5] Sosnina E.P. Prikladnayia lingvistika. O rasrabotke I ispolsovanii russkogo obrasovatelnogo konkursa perevodov (Applied Linguistics. On the development and use of the Russian educational corpus of translations). URL: http://ling.ulstu.ru. 2019. 242. [In Rus.]

[6] Rykov V.V. Korpus tekstov i rechevayia deyatelnost – problema shodstva (Corpus of texts and speech activity - problems of similarity). Corpus linguistics -2006: tr. intl. conf. Oct 10-14 2006, St. Petersburg. SPb., 2006. 347-355. [In Rus.]

[7] Zhumabekova A. Kazakh tilinin antonymder sozdigi (dictionary of Kazakh antonyms). Almaty, 2000. 128. [In Rus.]

[8] https://sozdikqor.kz/

[9] https://qazcorpus.kz

[10] Belbaeva M. Kazakh tilinin omonimder sozdigi (Dictionary of homonyms of the Kazakh language). Almaty, 1988. 193.[In Kaz.]

ҚАЗАҚ ТІЛІНІҢ ҰЛТТЫҚ КОРПУСЫНДАҒЫ МӘТІНДЕРДІ АВТОМАТТАНДЫРУ ӘДІСТЕРІ

*Әмірбекова А.Б.[1], Қоңырова А.Т.[2], Қайырбекова Ұ.С.[3]

*[1]А. Байтұрсынұлы атындағы Тіл білімі институты, Лексикология бөлімінің меңгерушісі, филология ғылымдарының кандидаты, Алматы, Қазақстан,
https://orcid.org/0000-0001-8540-8264,
e-mail:marghan01@mail.ru,

[2]филология ғылымдарының кандидаты, әл-Фараби атындағы Қазақ ұлттық университеті, Жоғары оқу орнына дейінгі білім беру факультеті, Шетелдіктердің тілдік және жалпы білім беру дайындығы кафедрасының аға оқытушысы, меңгерушісі, Алматы, Қазақстан,
e-mail: akbota74@mail.ru,

https://orcid.org/0000-0001-6996-6522,
[3]Академик Ә. Қуатбеков атындағы Халықтар достығы университеті, Филология кафедрасының меңгерушісі, филология ғылымдарының кандидаты, доцент, Қазақстан, Шымкент,
https://orcid.org/0000-0002-9466-024X

Аңдатпа. Тілдің жаһандану дәуірінде, яғни сөздік қор толықтай цифрлық жүйеде қолжетімділікке ие болған кезеңде, тіл меңгертуді оңтайландыруға мүмкіндік туады. Қазақ тілінің ұлттық корпусы – қазақ сөзінің цифрландырылған күрделі нұсқасы. Себебі қазақ тілінің корпусы тілтанымдық қызметтермен қамтамасыз ете алады, соның негізінде бірнеше подкорпустардан тұрады. Оның подкорпустарының ішіне оқу корпусына деген сұраныс күннен күнге артып келеді. Оның себебі Қазақстан көпұлтты мемлекет болғандықта. Сондықтан қазақ мәдениетін тұтынушы өзге ұлт өкілдері өз мәдениетінен пара-пар нұсқасын тану үшін аударма-баламасын анықтағысы келеді. Сондай-ақ оқу корпусы тіл үйренушілерге тиімді тілтанымдық база болары да сөзсіз.

Мақаланың мақсаты – оқу корпусына енетін тілтанымдық қорды, әсіресе қазақ тілінің лексикалық қабатындағы сөздерді тақырыптық, мағыналық, мәдени-семантикалық топтарға іріктеп семантизациялау және цифрлық жүйеге сәйкес автоматтандыруға лайықтау әрі икемдеу. Мақалада оқу корпусының басқа подкорпустан айырмашылығы, семантизациялау жіктелімі және мағыналық топтарды кестелеу (автоматтандыруға лайықтау) жолдары ұсынылды. Мақаланың ғылыми маңыздылығы да осында.

Оқу корпусына лайықты тілдік қорды енгізу барысында контент-талдау әдісі, кестелеу, жинақтау, сипаттама әдістері қолданылды. Мақалада ұсынылған ғылыми тұжырымдар оқу корпусын әзірлеуге, тіл меңгертуге арналған электронды қосымшаларды жасауда септігін тигізетін тәжірибелік маңызы бар.

Мақала ***BR 18574183 «Қазақ мәтінін автоматты тану: лингвистикалық модульдер мен IT -шешімдер әзірлемесі»*** атты жобаның аясында зерттелді.

Тірек сөздер: ұлттық корпус, семантизация, автоматтандыру, тілтанымдық база, аударма, лексикалық қабат, оқу корпусы, тілдің цифрлануы

МЕТОДЫ АВТОМАТИЗАЦИИ ТЕКСТОВ В НАЦИОНАЛЬНОМ КОРПУСЕ КАЗАХСКОГО ЯЗЫКА

*Амирбекова А.Б.[1], Конырова А.Т.[2], Кайырбекова У.С.[3]
*[1]кандидат филологических наук, заведующий Отделом лексикологии Института языкознания имени А.Байтурсынова,
Алматы, Казахстан,
https://orcid.org/0000-0001-8540-8264,
e-mail: marghan01@mail.ru,
[2] кандидат филологических наук, старший преподаватель, заведующий кафедрой языковой и общеобразовательной подготовки иностранцев, факультет довузовской подготовки, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан,
https://orcid.org/0000-0001-6996-6522,
e-mail: akbota74@mail.ru,
[3] кандидат филологических наук, доцент, заведующий кафедрой филологии Университета дружбы народов имени академика А. Куатбекова, Казахстан, Шымкент,
https://orcid.org/0000-0002-9466-024X

Аннотация. В эпоху глобализации языка, когда лексикографическая база становится полностью доступной в цифровой системе, появляется возможность оптимизировать овладение языком. Национальный корпус казахского языка – оцифрованный вариант казахского слова. Так как корпус казахского языка является системой лингвистических знаний, состоящей из нескольких подкорпусов, спрос на учебный корпус растет день ото дня. Это связано с тем, что Казахстан является многонациональным государством. Представители других национальностей, потребляющие казахскую культуру, хотят определить перевод-эквивалент. Учебный корпус также является эффективной лингвистической базой для изучающих язык.

Цель статьи – семантизация лексикографической базы, входящей в учебный корпус, особенно единиц лексического слоя казахского языка, и адаптация к автоматизации в соответствии с цифровой системой. В статье описано отличие учебного корпуса от других подкорпусов, представлены классификация семантизации и способы интерпретации (автоматизации) семантических групп. В этом научная значимость статьи.

При изучении внедрения в учебный корпус лексической базы применялись методы контент-анализа, обобщения, описания. Научные выводы, представленные в статье, имеют практическое значение, способствуя разработке учебного корпуса, разработке электронных приложений для овладения языком.

Статья исследована в рамках проекта BR 18574183 «Автоматическое распознавание казахского текста: разработка лингвистических модулей и IT-решений».

Ключевые слова: национальный корпус, семантизация, автоматизация, языковая база, перевод, лексический слой, учебный корпус, цифровизация языка.

Статья поступила 05.06.2023