IRSTI 16.01.11 doi.org/10.55491/2411-6076-2024-1-189-196

Q.B Slyambekov[1*], A.M. Sadyk[2]
[1]Institute of Linguistics named after A. Baitursynuly, Kazakhstan, Almaty
[2]University of International Business named after K.Sagadiyev, Kazakhstan, Almaty
[*]e-mail: zatpost@gmail.com

THE NATIONAL CORPUS OF THE KAZAKH LANGUAGE: THE SEMANTIC MARKUP OF VERBS

Abstract. Due to the fact that the development of the corpus has become one of the priorities for all languages of the modern world, the improvement of the national corpus of the Kazakh language (NCKL) is also a very relevant issue. One of the types of linguistic information reflecting the meaning of a word in the NCKL database is lexical-semantic markup. The article examines the world experience of lexical-semantic markup and provides an overview of foreign research. Analyzing the national corpus of the Russian language and the Kalmyk language, the peculiarities of the national corpus of the Kazakh language are noted. The methods of dividing verbs into lexical-semantic groups are indicated, on the basis of which the markup of the corpus is formed, i.e. the definition of codes that reveal the meaning of the word.

In the study, lexical-semantic groups were classified according to the method of describing verb meanings and synthesizing based on common meanings, semantic groups of Kazakh verbs were compared with each other and with semantic groups in other languages.

As a result of the research, macro- and microgroups characterizing the meaning of the verb were included in the National Corpus of the Kazakh language. In total, 100 lexical-semantic groups have been formed. The lexical-semantic markup of verbs included six different codes. Lexical-semantic markup was attached to 18200 verbs based on the corpus.

The compiled lexical-semantic markup expands the information about the word in the National Corpus of the Kazakh language, allows the user to easily determine the meaning of the word, sort verbs with a similar meaning and verbs with a positive, negative connotation of meaning. It can be said that lexical-semantic markup is one of the first steps to facilitate the recognition of the semantics of words of Kazakh artificial intelligence, which is expected to be created in the near future.

Keywords: National Corpus , corpus linguistics, verb, lexical-semantic group, lexical-semantic markup.

Қ.Б. Слямбеков[1*], А.М. Садық[2]
[1]А. Байтұрсынұлы атындағы Тіл білімі институты, Қазақстан, г. Алматы
[2]К. Сағадиев атындағы Халықаралық Бизнес университеті, Қазақстан, г. Алматы
[*]e-mail: zatpost@gmail.com

ҚАЗАҚ ТІЛІНІҢ ҰЛТТЫҚ КОРПУСЫ: ЕТІСТІКТЕРДІҢ ЛЕКСИКА-СЕМАНТИКАЛЫҚ БЕЛГІЛЕНІМІ

Аннотация. Корпус әзірлеу қазіргі әлем тілдерінің барлығы үшін басым бағыттардың біріне айналғандықтан, Қазақ тілінің ұлттық корпусын (ҚТҰК) жетілдіру де аса өзекті мәселе. ҚТҰК базасындағы сөз мағынасын танытатын линвгистикалық ақпараттың бір түрі – лексика-семантикалық белгіленім. Мақалада лексика-семантикалық белгіленімге қатысты әлемдік тәжірибе қарастырылып, шетелдік зерттеулерге шолу жасалады. Орыс тілі және қалмақ тілі ұлттық корпустары талдана келе қазақ тілінің ұлттық корпусындағы ерекшеліктер атап өтіледі. Етістіктерді лексика-семантикалық топтарға бөлу, соның негізінде корпус белгіленімін жасақтау, яғни сөз мағынасын ашатын белгі-кодтарды анықтау жолдары көрсетіледі.

Зерттеуде лексика-семантикалық топтар етістіктердің мағыналарын сипаттау және ортақ мағыналары негізінде синтездеу тәсілі арқылы жіктелді, қазақ етістіктерінің мағыналық топтары өзара және басқа тілдердегі семантикалық топтармен салыстырылды.

Зерттеу нәтижесінде Қазақ тілінің ұлттық корпусына етістіктің мағынасын сипаттайтын макро- және микротоптар енгізілді. Барлығы 100 лексика-семантикалық топ жасақталды. Етістіктердің лексикалық-семантикалық белгіленімі алты түрлі белгі-кодты қамтыды. Корпус базасындағы 18200 етістікке лексика-семантикалық белгіленім қойылды.

Жасалған лексика-семантикалық белгіленім Қазақ тілінің ұлттық корпусында сөз туралы ақпаратты кеңейтіп, қолданушының сөз мағынасын оңай анықтауына, мағынасы ұқсас етістіктерді және жағымды, жағымсыз реңкке ие етістіктерді сұрыптауына мүмкіндік береді. Лексика-семантикалық белгіленім алдағы уақытта жасалуы көзделіп отырған қазақ тіліндегі жасанды интеллектінің сөз семантикасын тану қызметін де жеңілдететін алғашқы қадамдардың бірі деуге болады.

Тірек сөздер: Ұлттық корпус, лингвистикалық корпус, етістік, лексика-семантикалық топ, лексика-семантикалық белгіленім.

 © А.Байтұрсынұлы атындағы Тіл білімі институты

К.Б. Слямбеков[1*], А.М. Садык[2]
[1]Институт языкознания им. А.Байтурсынулы, Казахстан, Алматы қ.
[2]Университет международного бизнеса имени К. Сагадиева, Казахстан, Алматы қ.
[*]e-mail: zatpost@gmail.com

НАЦИОНАЛЬНЫЙ КОРПУС КАЗАХСКОГО ЯЗЫКА: ЛЕКСИКО-СЕМАНТИЧЕСКАЯ РАЗМЕТКА ГЛАГОЛОВ

Аннотация. В связи с тем, что разработка корпуса стала одним из приоритетных направлений для всех языков современного мира, совершенствование национального корпуса казахского языка (НККЯ) также является очень актуальным вопросом. Одним из видов лингвистической информации, отражающей значение слова в базе данных НККЯ, является лексико-семантическая разметка. В статье рассматривается мировой опыт лексико-семантической разметки и дается обзор зарубежных исследований. После анализа национальных корпусов русского и калмыцкого языков, отмечаются особенности национального корпуса казахского языка. Указываются способы деления глаголов на лексико-семантические группы на основе которых формируется разметка корпуса, т. е. определение кодов, раскрывающих значение слова.

В ходе исследования лексико-семантические группы классифицировались по способу описания и синтеза значений глаголов на основе их общих значений, семантические группы казахских глаголов сравнивались между собой и с семантическими группами в других языках.

В результате исследования в Национальный корпус казахского языка были введены макро и микрогруппы, описывающие значение глагола. Всего сформировано 100 лексико-семантических групп. Лексико-семантическая разметка глаголов включала шесть различных кодов. 18200 глаголов в базе данных корпуса были отмечены лексико-семантической разметкой.

Составленная лексико-семантическая разметка расширяет информацию о слове в Национальном корпусе казахского языка, позволяет пользователю легко определять значение слова, сортировать глаголы со схожим значением и глаголы с положительным, отрицательным оттенком. Можно сказать, что лексико-семантическая разметка является одним из первых шагов, облегчающих деятельность по распознаванию семантики слов искусственного интеллекта в казахском языке, который предполагается создать в ближайшем будущем.

Ключевые слова: Национальный корпус, лингвистический корпус, глагол, лексико-семантическая группа, лексико-семантическая разметка.

Introduction

The modern path of science's development is directly related to the concept of "Innovation". The main goal of state programs in Kazakhstan is the introduction of innovative technologies in all spheres of science and production. In linguistics, the concept of innovation is applied primarily to the effective use of information and communication and computer technology, the development of popular science products, the obtaining of quantitative, qualitative, and objective data using modern technology in research work, and the development of an alternative empirical base.

Innovative research in Kazakh linguistics began with the All-Union scientific meeting "Statistical and Informational Research in the Turkic Languages", held back in the Soviet era at the Institute of Linguistics of the Academy of Sciences of the Kazakh SSR in 1969. In 1970, the scientific group "Linguistic Statistics and Automation" was created at the Institute under the guidance of the renowned mathematician K. Bektaev, a specialist in mathematical statistics. This scientific group was initially engaged in the creation of frequency dictionaries for various styles of the Kazakh language. Thus, in the 1990s of the last century, a 20-volume collection of M.O. Auezov's works was included in the computer memory, and in 1995, "Frequency Dictionaries of M.O. Auezov's Works in 20 Volumes" was included.

Corpus linguistics studies the compilation and use of national corpora, which summarize texts in different languages of a country into a computer database and put them into a program management system. Corpus studies in world linguistics began in the 1960s. In many other languages, national corpora have now been created. The creation of a corpus is a modern and innovative tool in the age of modern information technology. In today's environment, with countless electronic services installed on mobile phones, automating research tools, making them more efficient and accessible to young researchers, and preserving our spiritual heritage of written texts in various styles accumulated over the centuries, valuable works, etc., in computer memory is a topical problem. Work on the creation of a National Corpus of the Kazakh language is currently in full swing at the Institute of Linguistics.

While world research on the practice of corpus creation has been considering the problem of

 © А.Байтұрсынұлы атындағы Тіл білімі институты

individual languages for a century, the problem of the National Corpus of the Kazakh language has been put forward only in the last 20 years. Currently, for the NCKL, which has lexical, morphological, phonetic and phonological markings, lexical-semantic markup is an innovation. Therefore, despite the fact that there have been studies on the theory of lexical-semantic markup of the Kazakh word, there are no works devoted to practice yet.

Materials and methods

For a better understanding of corpus linguistics, corpora in the languages of the world were taken into account, and the main directions and system of functioning of each corpus were identified. These works were carried out on the basis of comparison and analysis. The construction of the National Corpus of the Kazakh language was explained on the basis of the descriptive, i.e., synchronic method. The theoretical information and practical analysis concerning the problem in question – lexical-semantic markup of verbs – were studied and presented in the article. Each conclusion has been drawn on the basis of linguistic methods and supported by examples. The rich lexical stock of the Kazakh language, especially verbs in the Kazakh language, has been taken into account in the study.

Literature review

General studies of corpus linguistics are presented in the world of linguistics in the works of E. Finegan, "Language: its structure and use" (Finegan, 2004), McEnery and E. Wilson, "Corpus Linguistics" (McEnery, 2001), and in the textbooks by V. Zakharov, "Corpus Linguistics" (Zakharov, 2020), and A. N. Baranov, "Corpus Linguistics: Introduction to Applied Linguistics' (Baranov, 2003). As well as separate analyses about national corpora in each language, they are studied in works by Z.A. Sirazitdinov "Modeling Bashkir Grammar" (Sirazitdinov, 2006), V.Z. Demjanikov "English-Russian terms in applied linguistics and automatic text processing", Ch.F. Meyer, "English Corpus Linguistics" (Meyer, 2002), M. Kyto, and M.A. Rissanen, "A Language in Transition: The Helsinki Corpus of English texts" (Kyto, 1992).

The study of computer linguistics in Kazakh linguistics begins with a doctoral monograph by A.Zhubanov defended in 2002, which is a significant contribution to Kazakh applied linguistics, titled "Basic Principles of Formalizing Content of Kazakh Text". Also, the problem of corpus linguistics was analyzed in such works as "Problems of Applied Linguistics" by A.Zhubanov, "Statistics of Kazakh Language" by A. Zhubanov, "Scientific and Practical Basics of Morphological Markup of Kazakh Language Texts", "Scientific and Practical Significance of Annotated Corpora", "Scientific and Research Potential of Language Corpus" by A. Zhanabekova, "The Problem of Homonyms in the Corpus of Kazakh Texts" by B. Karbozova, and "On the Electronic Base of the Corpus Linguistics" by S. Kulmanov.

Results and discussions

The word "разметка" in the Russian language is known in Kazakh linguistics as "белгіленім" (A.Zhubanov invented the term "белгіленім"). The case is not just a collection of texts stored in computer memory, but also a tool with an interface that carries out numerous linguistic analysis automatically. The aim is to automatically analyze each level of language in order to subordinate it to the computer in the Corps. For automatic analysis, glutinative languages are extremely useful. This is the first morphological notation tool with automatic root and suffix analysis, and it is used in all national buildings since the root and appendix of such languages are readily visible. The Kazakh language's morphological parser (Analyzer) can automatically separate word forms even in the National Corpus.

Lexical-semantic markup is an automatic program that reflects a thematic or lexical-semantic group of words. In the world's corpora, lexical-semantic markup is adopted later than morphological markup. This is due to the complexity of the annotation. Nevertheless, creating the semantic partitioning in the National Corpus of the Russian language was not very difficult. This is due to the fact that the Lexicographer database of 10,000 words, based on the vocabulary of the Russian language, has been in existence since 1992. This system was developed under the guidance of E.V. Paducheva and has a very rich verb base. The advantage of a lexicographer is that it allows one to analyze words according to their semantic specificity. The main direction of the system is defined by the authors of the project as follows: The Lexicograph system is designed as a high-tech tool for linguistic research. Its main goal is to obtain a list of words according to predetermined semantic parameters (primarily taxonomic) and to

 © А.Байтұрсынұлы атындағы Тіл білімі институты

study the linguistic character of the obtained lexical classes, in particular, combinability, types of control models, and diathetic movement based on systematic correlation in the lexicon" (Kustova, Lyashevskaya, Paducheva, Rakhilina, 2005: 156). The Lexicographer database, which began its work at the end of the twentieth century, has been of great help to corpus compilers in creating lexical-semantic markup. As a result, lexico-semantic search in the National Corpus of the Russian language has been launched since October 2004: "Despite the small volume, theoretically we can say that the Lexicographer database of experimental research data plays a major role in creating the lexico-semantic corpus and forms the complete basis of the corpus lexico-semantic dictionary" (Kustova, Lyashevskaya, Paducheva, Rakhilina, 2005: 157).

Three key characteristics are used to categorize words while creating the lexical-semantic markup for the Russian National Corpus. These are: 1. by lexico-grammatical categories: qualitative, relative (adjective), factual, abstract (noun), etc.; 2. by derivational character: "diminutive", "augmentative", "attenuative", "nomen agentis", "nomen femininum", etc. Thematic (taxonomic) class, "evaluation," causation (verb and verb), etc. are examples of distinct semantic characteristics. Depending on how detailed each word class is, different sorts of semantic labels and the linguistic information they provide are communicated. For instance, it classifies verbs based on four characteristics:

1. *Lexico-semantic groups*, also known as movement, physical impact, creation, destruction, emotion, and speech, are categorized in accordance with semantic groups. Each lexical-semantic group is identified using an English name. For instance:

t: move: move (run, jerk, throw, carry)

t: move:body: change body position or body part (bend, bend over)

t: put: placement of an object (put, enclose, hide)

impact: physical action (hit, prick, wipe).

2. *Causation.* Below is a sign indicating whether it is causative or non-causative:

ca: causative-ccausative verbs (show, twirl)

noncausative verbs (see, twirl)

3. *Description of auxiliary verbs.* They are divided into phase and causative auxiliary verbs:

aux:phase: phasic (start, continue, stop).

aux:causative, auxiliary causative (to cause, to lead).

4. *In the derivational denotation,* verbs are divided into prefixive verbs, semifixive verbs, and secondary imperfective verbs:

d: pref: prefix verbs (run in, look around)

d: semelf: semelfactives (to nod, to sneeze, to wince, to sway).

d: impf: secondary imperfectives (-iva-, -va-, -a-) (to drink, to beat).

The National Corpus of the Russian Language has a lexical-semantic breakdown of six classes of words: noun, adjective, numeral, verb, pronoun, and adverb. The noun is not given as a separate word but as an actual, abstract, or proper, internally divided into three types. The thematic group is called a 'taxonomy'. The parallel markup of several lexical-semantic groups in the selection of semantic filters allows different semantic tones in one word to be taken into account. And verbs are sorted into four main features: taxonomy, causation, auxiliary verbs, and word formation, and classified into 18 large and 9 small lexical-semantic groups.

The National Corpus of the Kalmyk language was created by the Kalmyk Institute for Humanitarian Research of the Russian Academy of Sciences, following the lead of the National Corpus of the Russian language. There are two Kalmyk language corps in existence right now: the first is the "Kalmyk Corps," which was created by E. Vankayeva, a graduate student at RSUU, and the second is the "National Corps of the Kalmyk Language". The National Corpus of the Kalmyk language's lexico-semantic markup provides a metasemantic categorization in four main areas: lexico-grammatical, lexico-thematic, evaluative, and derivational.

The transfer of theme clusters common to all Word classes distinguishes the Kalmyk Corps from the Russian Corps. The following is how the corpus compilers define this transfer's essential elements: "We also concur that all Word classrooms use topic groups. Word semantics, in our perspective, are a constant in the language. Sema, for instance, is a lexical unit that is shared by terms from several word

 © А.Байтұрсынұлы атындағы Тіл білімі институты

classes and appears in all languages. It is related to the issue of animals. Additionally, the Kalmyk language uses isafet structures (modn Ger, "wooden house") frequently. When two nouns are combined, the first one is an adjective according to the word class. Modn "tree" can refer to either. The classification of all word classes into common thematic groups simplifies technical work by constructing everything in a standard, ready-made model that does not require separate study, but in the "standard grouping", it is difficult to reveal word meanings and shades of meaning. For this purpose, the lexical and semantic markup of each word class should be considered separately, and the classifications should be made separately" (Kukanova, 2005: 187).

The basic requirements for developing markups in the Kazakh language have been established, and a system of markups has been devised that can adequately explain the lexical and semantic aspects of verbs in the Kazakh language by differentiating markups in the National Corps of foreign languages. The lexical and semantic markup of verbs in the Kazakh language identifies the general and specific semantics of the verbs in the language, provides a lexical and grammatical description, establishes the connotation tone, classifies the verbs into semantic groups in accordance with the Kazakh language's structure, and combines small semantic groups into large classes for practical use. The most understandable terms and names are employed, concepts are utilised, and consideration is given when coding the markups in the corpus. Taking into account the homonymy of verbs, a separate markup is made for each meaning.

The National Corpus of the Kazakh language is created in six areas in accordance with the aforementioned criteria for lexical and semantic markups:

1. *Due to the way words are formed*, they are *singular/complex* and *main/derivative*. The markup's composition includes the expression of the word's personality and composition, which helps you understand the word's meaning more clearly. A verb's grammatical structure, which suggests that it is a derived verb deriving from the root verb or another word, or a complicated verb made up of numerous words, is used specifically to determine a verb's derived meaning. For instance, the verb "ызыңда" is derived from the imitator "ызың" and as such, its primary meaning is to represent the sound made by insects in nature. On the basis of this meaning, one can determine its lexico-semantic group (Table 1).

Table 1 – Word formation description

Ызыңда	Етістік	*Дара*	*Туынды ызың+да*

The corpus uses conventional terminology to describe the word-forming sign code, namely, that it may be intelligibly divided into basic, derivative, singular, and complex categories. This isn't because experts don't understand or use new terminology; rather, it's done to make use as simple as possible by giving the user ideas they're already familiar with. This is due to the fact that linguists understand English words like "augmentation" and "attenuation" but other types of experts might not.

2. *Depending on the lexico-grammatical meaning:* positive or negative, transitive and non-transitive verbs. This transfer is closely related to the fact that the listed categories are lexico-grammatical categories that are both semantic and grammatical in nature. In other words, they provide the term additional semantic and grammatical value, which indirectly affects its meaning. The scientist M. Orazov defines the transition/non-transition category as a semantic category in his work "Etistik" (“Verb”) (Orazov, 1991; 158).

3. *Division by minor semantics.* The ultimate differentiating semantics of each word are established based on the individual (minor) semantics, and the meaning is clearly understood. They served as the foundation for the development of a classification that included 75 sub-semantic groups.

4. *Division by General (large) semantics.* Combining tiny semantic groups into large classes based on shared semantics is known as division by general (large) semantics. This seeks to make searching easier. The transmission of 5–6 large groups in the Corps, in accordance with the Russian Corps' experience, is advantageous for the user's visual perception. 10 of these major classes are offered by the Kazakh language corps.

5. *Depending on the connotative Nature*. Three indicators are used: "positive," "negative," and

 © А.Байтұрсынұлы атындағы Тіл білімі институты

"neutral," depending on the connotative character. Each verb has an own emotional tone, and this emotional tone has an impact on its syntax, usage in context, and meaning. Only the emotional undertone can distinguish some synonymous words. And depending on the situation, some verbs can have a variety of shades. This is necessary not only to distinguish between negative and positive but also to show the possibilities of the Kazakh language by identifying verbs that can be used in both meanings in our language.

6. *Exemplification.* Illustrative examples were included to the classification and labeling of verbs into lexical and semantic groups in order to contextualize the markup's meaning.

Examples of lexical-semantic markup of verbs can be seen in (Table 2).

Table 2 – Lexical-semantic markup of verbs

Белгі-кодтар (Markings)	*Қарай*	*Аққөзден*
Сөз табы	етістік	Етістік
Сөздің морфемалық құрамы	қара+й	ақ+көз+ден
Сөзжасамдық тәсілі	синтетикалық	аналитика-синтетикалық
Тұлғасы	туынды	туынды
Құрамы	дара	күрделі
Іс-әрекеттің болымды/болымсыздығы	болымды	болымды
Салттылық/сабақтылық	салт	салт
Макротоп	Бір түстен екінші түске ауысу, басқа бір күйге ену	Адамның жай-күйі, қалып етістіктері
Микротоп	бір түстен екінші бір түске ауысу, басқа бір қалыпқа ену	Адамның бір күйден екінші күйге ауысуы, біреудің екінші біреуді әлдебір күйге түсіруі, күй-қалыптың ауысу
Субъектіге қатысы	–	Орындаушы – субъектінің өзі
Коннотация	бейтарап	жағымсыз
Иллюстрация	Үйдің төбесіне жапқан сырғауыл қ а р а й ғ а н, түгел ыстанған (М. Әуезов, Таңд. шығ.)	Қолымызға сәл-пәл билік тисе ақ болғаны, дүниені ұмытып, а қ к ө з д е н і п кетеміз (Қ.Жұмаділов, Соңғы көш)

The verb is the second-largest grammatical category after nouns. Any root in the Turkic language, according to Turkologists who study the homonymy of nouns and verbs in Turkic languages, is syncretic, emphasizing both an object and a gesture. Because of this, renowned Turkish linguist A.M. Shcherbak asserts that gesture names are the source of verb forms. In the process of identifying phenomena and processes using linguistic methods, the foundation for communicating the speaker's relationship to the action's performer and the action itself in real speech started to be laid during the time when gesture names were noun-verb syncretism. As a result, a semantic plurality of verb forms was created through gesture names (Momynova, 2012). Kazakh linguists paid special attention to verbs and conducted extensive research. In Kazakh linguistics, the verb is considered in structural grammar from the point of view of traditional principles and in functional grammar from the point of view of aspect. In this case, it is necessary to take into account the grammatical, lexical, and semantic aspects of verbs.

Conclusion

There were several issues with the practical side as lexical-semantic markups developed. The need to define figurative verbs, differentiate between homonymous and polysemous verbs in context, incorporate verbs with various shades of meaning in one semantic group, etc. The article's fundamental properties are actively transferred during the initial stage. Future research will focus on developing semantic-syntactic models as the best way to address the aforementioned issues.

 © А.Байтұрсынұлы атындағы Тіл білімі институты

The creation of semantic and syntactic models, which will eventually play a significant part in the process of text analysis, will begin with the creation of lexica-semantic markup. We are able to examine the language from various perspectives and thoroughly research the Kazakh language's structure and system thanks to the inclusion of a semantic analyzer on a parallel electronic resource.

The National Corps of the Kazakh language, which is expanding and developing, is now the primary digital base of the Kazakh language thanks to the dedication of researchers and journalists. This base shows the potential of the language, makes it easier to understand and use, and speeds up research.

The article was written within the framework of the research project BR18574183 "Automatic recognition of the kazakh text: development of linguistic modules and it solutions".

References

Baranov A.N. (2003) Korpysnaıa lıngvıstıka. Vvedenıe v prıkladnyıy lıngvıstıky. – M.: Editorial URSS, 2003. – 114 s. [Baranov A.N. (2003) Corpus linguistics. Introduction to applied linguistics. – M.: Editorial URSS, 2003. – 114 p.] (in Russian)

Finegan E. (2004) Language: its structure and use. – N.Y.: Harcourt Brace College Publishers, 2004. – 607 p. (in English)

Isaev S. (1998). Qazirgi qazaq tilindegi sozderdin grammatıkalyq sıpaty. – Almaty: Rayan, 1998. – 304 b. [Isayev S. (1998) Grammatical nature of words in the modern Kazakh language. – Almaty: Rauan, 1998. – 304 p.] (in Kazakh)

Jybanov A. K., Janabekova A. (2017). Korpystyq lıngvıstıka. – Almaty: Qazaq tili, 2017. – 336 b. [Zhubanov A. K., Zhanabekova A. (2017) Corpus Linguistics. – Almaty: Qazaq tili, 2017. – 336 p.]. (in Kazakh)

Kykanova V. (2005) Prıncıpy semantıcheskoı razmetkı Nacıonal'nogo korpysa Kalmykskogo ıazyka. – Semantıcheskaıa razmetka leksıkı v Nacıonal'nom korpyse rysskogo ıazyka: prıncıpy, problemy, perspektıvy. – Nacıonalnyı korpys rysskogo ıazyka: 2003-2005 Rezyltaty ı prospekty. – S. 187-192. [Kukanova V. (2005) Principles of semantic markup of the National Corpus of the Kalmyk language. National Corpus of the Russian language: 2003-2005. Results and prospects. – M., 2005. – P. 187-192.] (in Russian)

Kystova G. I., Lyashevskaıa O. N., Padycheva E. V., Rakhılına E. V. (2005) Semanticheskaıa razmetka leksıkı v Nacıonal'nom korpyse rysskogo ıazyka: prıncıpy, problemy, perspektıvy. – Nacıonalnyı korpys rysskogo ıazyka: 2003-2005 Rezyltaty ı prospekty. – S. 155-174. [Kustova G. I., Lyashevskaya O. N., Paducheva E. V., Rakhilina E. V. (2005). Semantic markup of vocabulary in the National Corpus of the Russian language: principles, problems, prospects. – National Corpus of the Russian language: 2003-2005 Results and prospects. – M., 2005. – P. 155-174.] (in Russian)

Kyto M., Rissanen M. (2012) A Language in transition: The Helsinki Corpus of English texts. – Helsinki: University of Helsinki, 2012. – 680 p. (in English)

McEnery T., Wilson A. (2001) Corpus Linguistics. – Edinburgh: Edinburgh University Press, 2001. – 235 p. (in English)

Meyer Ch. F. (2002) English Corpus Linguistics & An Introduction. – Cambridge: Cambridge University Press, 2002. –168 p. (in English)

Momynova B. (2012) Qazaq tilinin morfologıasy. – Almaty: Arys, 2012. – 239 b. [Momynova B. (2012) Morphology of the Kazakh language. – Almaty: Arys, 2012. – 239 p.] (in Kazakh)

Orazov M. (1991). Qazaq tilinin semantıkasy. – Almaty: Rayan, 1991. – 216 b. [Orazov M. (1991) Semantics of the Kazakh language. – Almaty: Rauan, 1991. – 216 p.] (in Kazakh)

Sırazıtdınov Z.A. (2006). Modelırovanıe grammatıkı bashkırskogo ıazyka. – Yfa: Gılem, 2006. – 160 s. [Sirazitdinov Z.A. (2006) Modeling of the grammar of the Bashkir language. – Ufa: Gilem, 2006. – 160 p.] (in Russian)

Zakharov V.P. (2020) Korpysnaıa lıngvıstıka. – SPb: Izdatelstvo Sankt-Peterbyrgskogo ynıversıteta, 2020. – 234 s. [Zakharov V.P. (2020). Corpus linguistics. – SPb: St. Petersburg University Press, 2020. – 234 p.] (in Russian)

Zybov A.V., Zybova I.I. (2004) Informacıonnye tehnologıı v lıngvıstıke. – M., 2004. – 208 s. [Zubov A.V., Zubova I.I. (2004). Information technologies in linguistics. – M., 2004. – 208 p.] (in Russian)

Әдебиеттер

Баранов А.Н. (2003) Корпусная лингвистика. Введение в прикладную лингвистику. – М.: Едиториал УРСС, 2003. – 114 с.

Жұбанов А.Қ, Жаңабекова А.Ә. (2017) Корпустық лингвистика. – Алматы: Қазақ тілі, 2017. – 336 б.

Захаров В.П. (2020) Корпусная лингвистика. – СПб.: Изд. Санкт-Петербургского университета, 2020. – 234 с.

Зубов А.В., Зубова И.И. (2004) Информационные технологии в лингвистике. – М., 2004. – 208 с.

Исаев С. (1998) Қазіргі қазақ тіліндегі сөздердің грамматикалық сипаты. – Алматы: Рауан, 1998. – 304 б.

Куканова В. (2005). Принципы семантической разметки Национального корпуса Калмыцкого языка // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. – М., 2005. – С. 187-192.

Кустова Г. И., Ляшевская О. Н., Падучева Е. В., Рахилина Е. В. (2005) Семантическая разметка лексики в Национальном корпусе русского языка: принципы, проблемы, перспективы // Национальный корпус русского языка:

 © А.Байтұрсынұлы атындағы Тіл білімі институты

2003-2005. Результаты и перспективы. – М., 2005, – С. 155-174.

Kyto M., Rissanen M. (2012) A Language in transition: The Helsinki Corpus of English texts. – Helsinki: University of Helsinki, 2012. – 680 p.

McEnery T., Wilson A. (2001) Corpus Linguistics. – Edinburgh: Edinburgh University Press, 2001. – 235 p.

Meyer Ch. F. (2002) English Corpus Linguistics & An Introduction. – Cambridge: Cambridge University Press, 2002. – 168 p.

Момынова Б. (2012) Қазақ тілінің морфологиясы. – Алматы: Арыс, 2012. – 239 б.

Оразов М. (1991) Қазақ тілініц семантикасы. – Алматы: Рауан, 1991. – 216 б.

Сиразитдинов З.А. (2006) Моделирование грамматики башкирского языка. – Уфа: Гилем, 2006. – 160 с.

Finegan E. (2004). Language: its structure and use. – N.Y.: Harcourt Brace College Publishers, 2004. – 607 p.

 © А.Байтұрсынұлы атындағы Тіл білімі институты