ISSN 2788-7979 (online)

ШОЛУЛАР, СЫН ЖӘНЕ БИБЛИОГРАФИЯ

ОБЗОРЫ, КРИТИКА И БИБЛИОГРАФИЯ

REVIEWS, CRITICISM AND BIBLIOGRAPHY

IRNTI 16.21.33 DOI: 10.59102/kufil/2024/iss3pp291-304

G. Madiyeva[1], L. Alimtayeva[1], Zh. Satkenova[1], K. Pirmanova[1]
[1]Al-Farabi Kazakh National University, Almaty, 050000, Kazakhstan

ANNOTATION OF ANALYTICAL STRUCTURES IN LANGUAGE CORPORA

This study analyzes the morphological and word-forming features of analytical structures in language markups by reviewing the annotations in Kazakh, Turkic, and Russian corpora. The findings reveal a significant degree of similarity among the corpora of leading Turkic languages, including Kazakh, Tatar, and Bashkir.

The analysis shows that there are no substantial issues with annotating combined and paired words that machines recognize as single compound units. Consequently, these can be searched and annotated similarly to other compounds, with morphological, word-forming, and lexical annotations applied to the entire unit. However, phrases written with spaces cannot be searched as a single lemma, resulting in each item being annotated individually rather than as a unified compound. This limitation negatively impacts the functionality of the corpora. Conversely, the annotation of phraseological units within the national corpus of the Kazakh language as a cohesive whole is a notable advantage of this corpus.

Improving the annotation of analytical forms and formants of nouns (degrees), verbs, and auxiliary words in accordance with the language structure will enhance corpus functionality. Although this task is complex, an analysis of the current lexical resources in the National Corpora of the Kazakh language, along with ongoing fundamental research, indicates that the gradual automation of analytical structure annotation is imminent.

Key words: corpus of the Kazakh language, nominative verb constructions, analytical form, morphological markup, word-formation markup.

MAIN PROVISIONS

The National Corpus of the Kazakh language represents the latest development in the field of linguistics, which is undergoing constant evolution. The establishment of the national corpus of the Kazakh language is inextricably linked to the name of the esteemed mathematician scientist Askar Zhubanov. The Corpus website contains an electronic corpus of the Kazakh language. "The corpus contains 40 million words of text. The texts have been categorised into five stylistic categories: artistic, scientific, journalistic, formal and colloquial. A search of the corpus can be conducted by word, wordform (word transformation). The results will display a list of sentences in which the word in question was used, along with the source. Information regarding the occurrence of a word or word form, or any word in the examples, is provided for all levels of the language. The Corpus can be utilized by Kazakh speakers and Kazakh language learners alike» [1].The principal authors of the scientific and academic works related to the automation of the Kazakh language are A. Zhubanov, A. Zhanabekova etc.

In general linguistics, the typological classification of languages is associated with the names of F. Schlegel, A. Schlegel, G. Girard, and W. Humboldt. Based on G. Girard's research, Nicole Boze, in her article "Language," distinguishes languages into only two types: 1) analogical (analytical); 2)

ISSN 2788-7979 (online)

transpositive (synthetic). She also stated that no language can be purely analytical or purely synthetic [2, 264].

French scholars, when categorizing languages into certain types, based their classification not on the morphological level but on the syntactic level (word order). In this regard, the authors considered language not in isolation but in connection with the human thinking system. The order of words in speech is a phenomenon closely related to the process of development of thought. Starting with the Port-Royal Grammar, French authors observed the reflection of the logical basis of the thinking process in language. According to N. Boze, word order is the result of the analysis of thought, and it can serve as a basis for discourse analysis. This idea was supported by Voltaire and Antoine de Rivarol. For instance, in his essay "On the Universality of the French Language," Antoine de Rivarol writes that the analytical structure of the French language is coherent, precise, and logical: "The French language first names the subject, then the verb expressing the action, and after the verb the object of that action [3 , 49].

In his 1774 treatise "On the Origin of Languages," Adam Smith categorized world languages as "compounded" and "uncompounded." Similar to the French scholars, Smith's classification is based on whether languages have an analytical or synthetic structure. In the first type of languages, words convey grammatical meaning "within themselves," whereas in the second type, meaning is conveyed through auxiliary words, such as prepositions and auxiliary verbs [4, 471].

According to Adam Smith's reasoning, analytical languages emerged from the development of human cognitive abilities, striving for complexity and abstract thinking. Evidence of this hypothesis can be found in the analytical structures within our language. For example, in Kazakh, conjunctions, auxiliary words, and complex structures may not convey complete meaning on their own but form a complete meaning when combined into a complex construction. The verb "kulau" (to fall) in Kazakh has various meanings. Using analytical methods, the speaker aims to convey their thoughts precisely.

According to the typological classification of languages, the Kazakh language is agglutinative, which means that it is a synthetic language that primarily has a synthetic structure. However, this does not imply that analytical structures are absent or very rare in our language. On the contrary, the number of analytical structures in Kazakh is increasing over time. Therefore, determining their usage frequency, comparing them with synthetic structures, conducting statistical analysis, and examining their forms in dictionaries are pressing issues in contemporary Kazakh linguistics.

In analyzing the morphological (degree of adjectives, analytical forms of verbs) and word formation (compound words formed through analytical methods) characteristics in the corpora of the analytical structures, which are the subject of our research, we relied on the works of scholars such as K. Zhubanov, I.E. Mamanov, K. Akhanov, A. Yskakov, M. Orazov, and others.

Russian researchers A.V. Ventsov and E.V. Grudeva posit that the standards for the Corpus should be as rigorous as those employed in dictionaries and encyclopedias:

“It appears that products of this nature, such as the national corpus of the language, should be subject to the same requirements as dictionaries, encyclopedias, textbooks, and so forth. In other words, the presence of factual errors is unacceptable in principle. A more time-consuming approach, but one that is also more effective in terms of the desired result, is the technology of semi-automatic text markup. This assumes the presence of an operator that is able to resolve issues related to homonymy in the broadest sense of the word” [5, 77].

Currently, under the leadership of the director of the A. Baitursynov Institute of Linguistics A. Fazylzhanova and the head of the Department of Applied Linguistics Professor A. Zhanabekova, the researchers of the Department of Applied Linguistics, grammar, terminology (B. Momynova, S. Kulmanov, B. Karbozova, etc.) conduct a lot of research within the framework of several programme-targeted financing projects and make a significant contribution to the development of the Kazakh linguistics Corpus. Complex problems in the development of various subcorpora, the results achieved are published and presented to the public in scientific conferences and journals.

ISSN 2788-7979 (online)

INTRODUCTION

At present, all branches of science are developing and reaching new heights through the use of innovative technologies. In the era of globalization, automation and digitization of all branches of science and education are becoming pressing concerns. In particular, the field of linguistics is not lagging behind and is seeking to gain new insights into language. The reason of the development and formation of a new branch of linguistics, namely "computational linguistics" can be attributed to the growing necessity for the automation of language materials. In this regard, the field of computational linguistics, which is concerned with the automation of language, is undergoing rapid development. In particular, over the past 10–15 years, a significant number of educational programmes have been developed in this field, not only at the level of individual research centres and institutes, but also within the philology faculties of universities across the country. Concurrently, the government offers assistance and provides substantial financial resources. At present, the most effective means of enhancing the use of the language in accordance with contemporary standards is to improve the national corpora of the Kazakh language to the greatest extent possible, supplement it with additional language materials, and develop it further. As a consequence, language corpora constitute the foundation for the development and enhancement of computer systems that process natural language, including automatic text analysis, machine translation, and word recognition.

MATERIALS AND METHODS

The corpus linguistics of the Turkic languages only commenced a period of accelerated development in the 1990s, and thus the creation of a publicly accessible corpus of the Turkic languages is an especially pertinent endeavour. Currently, the number of representative corpora for Turkic languages is on the rise. As a case in point, we may cite the following examples:

1) The Turkish national corpus that comprises a collection of texts from various genres.

2) The Bashkir poetic corpus, comprising a volume of over 1.8 million words. It is the second largest poetic corpus in the world. The corpus is notable for its focus on the works of Bashkir poets from the 20th and early 21st centuries.

3) The corpus of written Tatar, comprising a volume of over 116 million words.

The experience of developing the Turkic language corpus had a positive impact on the development of the Kazakh language corpus. Nevertheless, the issues of development and replenishment of the national Kazakh corpus remain pertinent, as this direction is a challenging and intricate aspect that is subject to ongoing enhancement. The existing and currently functioning projects of the corpus of the Kazakh language are comprised of the "National Corpus of the Kazakh language" of the Institute of Linguistics named after A. Baitursynov, which includes the main Corpus and 11 subcorpora. These subcorpora are as follows: cultural and representative, parallel, writers, spoken, proverbs, onomastics, A. Baitursynov, dialectal, advertising, and phraseology (qazcorpus.kz). The subcorpora of the National Corpus of the Kazakh language, developed by the National Scientific and Practical Center "Til-Kazyna named after Sh. Shayakhmetov", include corpora of publicistic texts and spoken language. Additionally, the "Almaty Corpus of the Kazakh Language" is worthy of mention. (Web-corpora). The principal objective of this study is to differentiate between the morphological and word-forming markups of analytical forms, formants in the corpora and subcorpora of the Kazakh language.

In line with the study's profile and orientation, analytical methods, including analysis, compilation, systematization, and description of language units, and methods of comparison of various corpus data are employed.

ISSN 2788-7979 (online)

RESULTS

During the research, the main corpora of the Kazakh language – national and minor corpora – were analyzed. The morphological and word formation characteristics of analytical structures in the Kazakh language corpora were compared with those of complex units in the corpora of other Turkic languages.

It is known that analytical structures in the Kazakh language are considered from two aspects: word formation and word transformation. From this perspective, in the corpus-based texts, complex structures should be annotated as one lemma in terms of word formation, and analytical formants or auxiliary words should be marked as a one gramema in terms of word transformation.

There is no significant problem in annotating compound and reduplicated words that are recognized as a single complex unit by machine in terms of word formation. These units can be searched in their entirety, and their morphological-word formation-lexical characteristics correspond to a single complex unit.

The main problem concerns analytical units written with spaces. It was found that it is not possible to search these phraseological complex units as a single lemma and that their morphological-word formation-lexical characteristics are not annotated as a whole complex unit but as separate parts. This issue is known to hinder the functionality of corpora (language teaching, linguodidactics, artificial intelligence, etc.).

The problems encountered in annotating analytical structures in the Turkic language corpora are common to the corpora of Turkic languages. We can say that the annotation of analytical structures in the corpora of Tatar and Bashkir, which have a rich vocabulary, is comparable to that of the Kazakh language corpora. However, although it is not possible to search for analytical structures as whole units in the Tatar and Bashkir language corpora as in the Kazakh language corpora, they are highlighted in bold in the text, indicating that they represent a single concept.

In the field of corpus linguistics, Russian language corpora are among the dominant ones, characterized by their polyfunctionality and extensive vocabulary. In these corpora, it is possible to search for such units as a single complex unit, and they are highlighted in bold to indicate that they represent a single concept, as in the Tatar and Bashkir corpora. However, their lexical-grammatical-word formation characteristics are shown based on individual components. The annotation of the syntactic functions of words can be considered an advantage of these corpora.

In the National Corpus of the Kazakh Language named after A. Baitursynuly, the annotation of phraseological units as whole units within the paradigm of complex units is recognized as a significant advantage of this corpus.

It is clear that properly annotating nominal (degree) and verbal analytical forms and formants, as well as auxiliary words, according to the structure of the language will improve the functionality of the corpus. The word formation characteristics of adjectives with intensifying syllables written with a hyphen are correctly indicated. Intensifying adverbs and basic adjectives are annotated separately. In the paradigm of nominal analytical structures in the Kazakh language, adverbial clauses also play a role. Since these do not change the lexical meaning of the main word, they should be recognized as a single lemma. However, some adverbial phrases are annotated as phraseological units.

The annotation of compound verbs and verbs with analytical forms is one of the most problematic and complex issues, as auxiliary verbs are semantically multi-functional and grammatically potential linguistic units. There are still many controversial issues from the perspective of linguistic theory regarding which aspect (word transformation or word formation) these two complex units belong to. However, it is also true that these complex units form the core of many functional-semantic categories (such as temporality, aspectuality, modality, etc.) in the Kazakh language. Therefore, the proper annotation of these linguistic units undoubtedly increases the practical value of language corpora. Currently, large projects are continuously being conducted to

ISSN 2788-7979 (online)

address these problems. This gives confidence that analytical structures will be automatically recognized and properly annotated.

DISCUSSION

The Kazakh language employs analytical structures, which are divided into two groups: complex nouns and complex verbs. It is of great importance to present the markup of complex words of analytical form in the corpus, to study their correctness, and to examine how they are reflected throughout the text. This is because the linguistic corpus is a resource that is compiled into an electronic language database, which provides all the information about the language. The distinctive feature of the Kazakh language national corpus is the manner in which annotations are provided. This implies that the language markup is attributed to each of the language units in the Corpus.

In the corpus, the noun is divided from the lexico-morphological point of view as follows:

1. Compound noun:

a) Combined word

b) Paired word

c) Compound word

d) Shortened word

In accordance with the established principles of word-forming markup, the composition and method should be clearly defined.A review of several complex words from various corpora reveals a range of characteristics: the national corpus of the Kazakh language presents the word *u'kendi-ki'shi'li* (*small-big*) (the morphological markup: adjective; semantic: complex, paired, relative; lexical - mixed old and little, different), *syn'g'yr-syn'g'yr* (*chik-chik*) - (morphological markup: onomatopoeic; semantic: 0; word formation: derivative, compound, analytical, paired word; lexical: dyn'g'yr-syn'g'yr, etc.) [6]. It is evident that the corpus demonstrates a certain degree of word formation and morphological markup of paired words as a single linguistic unit. This corpus reveals that complex paired words are formed according to the derivational, complex, and analytical methods. They are observed to occur as paired words within compound words and as a repeated paired type among paired words. Nevertheless, there is a certain degree of confusion regarding the signs according to the types of markups.

An examination of the subcorpora of the National Corpus of the Kazakh language, developed by the National Scientific and Practical Center "Til-Kazyna named after Sh. Shayakhmetov", reveals that the morphological markup of the paired word *u'lkendi'-ki'shi'li'* (*small- big*) is not provided. However, the links to the Sozdikqor and Termincom websites are indicated, where the morphological markup is not activated. Furthermore, the markup of the paired word *syn'g'yr-syn'g'yr* (*chik-chik*) was not identified [8].

In the Tatar National Corpus "Tugan Tel", the word *'u'kendi'-ki'shi'li'* (*small- big*) is not recognised as a complex word; rather, it is translated as homogeneous parts. The adverb *alg'a-artqa* (*up-down*) is presented as a single lexical item, yet without any indication of its word forming features, it is solely identified as an adverb on the basis of its morphology [8].

One of the key challenges in the annotation of language corpora is the manner in which language units written with spaces are treated. The implementation of their recognition as a single lemma is a pressing issue in the present day. In examining the corpus for compound numerals formed in this manner, it becomes evident that they are not recognized as a compound unit. Additionally, the morphological and word-forming markups are observed as a standalone unit. To illustrate, the numeral *zhiyrma bes* is identified as two separate numerals, *zhiyrma* and *bes*, in both of the corpora that were previously analysed. The fact that they represent a single concept from the perspective of word formation has yet to be revealed.

In the Almaty Corpus of the Kazakh language, the word *zhiyrma bes* is attributed to the part of speech "numeral" and has a Russian translation [9]. In the National Corpus of the Tatar language "Tugan Tel", the compound numeral *otyz bes* is designated as separate lemmas in the form "egerme+ Num", "Bish+ Num", which have a morphological markup as a part of speech "numeral" [8]. In the

ISSN 2788-7979 (online)

Bashkir poetic corpus, as in the Almaty Corpus of the Kazakh language, the compound numeral is designated as a part of speech "numeral" and the Russian translation is marked. For example, *qyrk ike* is marked as *qyrk+Num* and *ike+Num* [10]. In the electronic corpus of the Khakass language, no grammatical markup is placed on compound words. The meaning of each word is only listed at the lexical-semantic level. In accordance with the analytical structure introduced in the linguistic corpus of the Crimean-Tatar language, the variants found exclusively in the texts are listed. No linguistic markups were identified in the texts.

As can be observed, in the majority of linguistic corpora belonging to Turkic-speaking communities, grammatical marking is employed solely in relation to the part of speech. Some corpora are not even annotated (no linguistic markup applied). This indicates that the development of the Corpora is progressing at a slow pace.

In the context of linguistic corpora, there are no issues with the lexical and grammatical marking of compound numerals in large chronologically and in size English corpora. These include the British National Corpus, the Bank of English, and the Brown Corpus. This is due to the fact that in the English language, these structures are written with a hyphen (-) in spelling: *twenty-five, sixty-eight, fifty-four,* and so forth.

A significant number of the analytical structures that underpin the Kazakh language are the forms of degree that fall within the category of degree.

The markup of the analytical forms of adjectives in the corpus can be considered in the category of degree (the superlative form), as well as in the context of word formation. The superlative form of an adjective can be formed in two ways.

1. By the addition of the intensifying syllable (*y'p-y'lken, qap-qara, qyp-qyzyl, sap-sary, ap-ashchy, zhap-zhan'a, etc.).*

2. By the addition of intensifying adverbs (*en', o'te, tym, asa, nag'yz, ti'pti'*).

In the National Corpus of the Kazakh language, the markup of the superlative form of adjective, formed by adding an intensifying syllable, is correctly identified. To illustrate, the word-forming markup of the word *sap-sary* is as follows: derived, complex, analytical approach, paired word, reduplicated.

However, in the superlative form of an adjective created by combining intensifying words such as *o'te, tym, asa, nag'yz and ti'pti' (very, real, even)*, the markups are assigned to each component of a complex word. For example, in the analytical structure of *en' qalauly (the dearest), en' (the most)* is an adverb, single word, root, intensifier, and *qalauly (the dearest),* is an adjective, "ly" is indicated only as an adjective-generating suffix. The phrases *en' keregi' (the most necessary*) and *en' basty (the most necessary)* are presented as phraseological units [6]. Even in subcorpora, the lexical and morphological markups are applied to the intensifying adverb *en* as a distinct lemma.

The subsequent analytical structure that generates grammatical meaning is the analytical forms of case. In the A. Baitursynov National Corpus, case postpositions are classified as function words, belonging to the singular, root word, case category; the meanings are: lexical - 1. *boiymen, zholymen* (*along*). 2. *arqasynda, ko'megi'men* (*by means of*). 3. *na'tizhesi'nde* (*in the result of*) [6]. In the second corpus under consideration, the function word is annotated as follows: semantics – case postposition; lexical – it occurs after the case word and is used to connect the word. The case postpositions *sol siyaqty, deii'n, sheii'n* are marked as follows: semantics – case postposition; lexical – it comes after a case word and is used to link a word [6]. The first corpus indicates the meanings of place and time in the context. In the majority of instances, if the phrase is identified as a discrete unit, the key phrase *zhyl saiyn* (*every year*) is indicated as a phraseological unit. If this phrase is annotated as an analytical form of the locative case (which is associated with the suffix *da*), it would be possible to ascertain both its function and its morphological nature. On this site, the phrase *erten'nen keshke deii'n* (*from tomorrow to evening*) is correctly identified as a phraseological unit. The lexical meaning is correctly annotated as a whole day, or a long day, until it is late. Although it is presented as a discrete unit, separated from the main word in phrases with identical meaning, such as *ko'pke deii'n, balag'a deii'n and ta'uba deii'n*, each is correctly identified with illustrative

ISSN 2788-7979 (online)

examples where contextual meaning is conveyed, including the temporal limits, considerations, assertions and perceptions [6].

The next category of contentious analytical forms of the nominal is the auxiliary nominal. In both instances, auxiliary nominals are indicated as seperate units. In addition to its status as a standalone unit, the qazcorpus.kz website also recognizes its function as a phraseological unit, recognized as a complex unit [6]. One of the advantages of the site https://qazcorpora.kz is that it often provides a more comprehensive annotation of morphological and word-forming features [7].

Even in the Russian national corpus, which is somewhat supplemented, the markings are clarified, as in the examples above, where forms are grammatically distinguished. Nevertheless, a distinction between grammatical, semantic, and syntactic relationships is displayed.

In regard to this matter, the "Corpus of Standard Written Russian" has been found to contain a significant number of complex structures, particularly those involving a noun and a function word. To illustrate, the aforementioned corpus presents the following usages as one-word forms: *v obnimku, drug druga and drug k drugu*. However, the objective of this corpus is to ascertain the frequency of occurrence of specific linguistic forms within the context of the genres represented in the included texts, rather than the linguistic markups. For this reason, only numerical indicators are marked there. The term "v obnimku" is associated with dramaturgy - 1, fiction - 6, journalism - 0, and scientific and popular literature 0 [11].

In the national corpus of the Kazakh language, the "Subcorpus of Phraseology" is designed to examine complex structures as a unified whole based on lexical-semantic markup. It is very convinient to use it as an electronic phraseological dictionary . On the site qazcorpora.kz, the lexical meanings of the unit *ku'i* are annotated as a separate word in various wordforms like *ti'stelegen ku'ii, a'bden ku'ii ketken, ko'n'i'l ku'ii'*.

The issue of emphasis in the corpus is related to compound verb forms. When discussing the phenomenon of analytism in the Kazakh language, the analytical forms and analytical formants of verbs immediately come to mind. From the perspective of the transformation system, verbs are the most susceptible to modification, followed by nouns and then adjectives. However, it is notable that synthetic forms of adjectives are often differentiated, and that analytical forms (such as intensifying syllables and intensifying adverbs) are not commonly discussed in academic grammars. Nevertheless, even in the markups of analytical forms of verbs, the marking of each component, analogous to that of the adjective, is provided separately. It is important to note that the analytical forms of verbs carry not only grammatical meaning, but also, to a certain extent, lexical and semantic meaning. To illustrate, in each of the variants, such as *ki'ri'p keldi', ki'ri'p shyqty (tez, lezde), ki'reii'n dedi', ki'rgi'si' keldi', ki'ri'p qala zhazdady (ki'rmedi')*, not only grammatical meaning is observed, but also semantic meaning.

It is similarly important to study and emphasise the auxiliary verbs in the Kazakh language in corpus linguistics.

Auxiliary verbs indicate the state of the action from the point of view of its semantic nature. A. Altayeva highlights that M. Orazov, a scientist, divides verbs into several groups from the lexical-semantic point of view, with a particular focus on the auxiliary-quality verbs, while T. Kordabayev considers the verbs *otyr, zhy'r, zhatyr, tu'r* to constitute a distinct category with present tense meaning, I. Kenesbayev emphasises that these verbs are used in the predicate function, functioning independently of other verbs [12, 33].

The studies on Kazakh grammar pay particular attention to the three main morphological features that distinguish auxiliary verbs from other types, emphasising the indescribable characteristic of auxiliary verbs (*otyr, zhy'r, zhatyr, tu'r*) for other languages. In particular, 1. the affix –yr is added to the root zhat in the present continuous tense, while the other three remain in their root form and are conjugated; 2. without changing the form they give the meaning of the present tense, while standing in the same position; 3. It is used in two different persons, two different tenses, and two different moods being in the same form [13, 18].

In the construction of a national corpus, it is essential to consider not only the grammatical meaning of verbs, but also their lexical-grammatical meaning. The group utilizing the corpus may

ISSN 2788-7979 (online)

be comprised of Kazakh-speaking individuals as well as foreign languages speakers. Although the meaning of auxiliary verbs in the context of Kazakh and related Turkic languages is readily apparent, it can be incomprehensible to non-related languages. For instance, the verb *bara zhatyr' is translated as 'be going', 'oqyp zhy'r' as 'be reading', 'aityp zhatyr' as 'be telling', 'oilanyp tu'r' as 'be thinking', and 'qorqyp ty'r' as 'is afraid*. It is not uncommon for an individual who is not proficient in the Kazakh language to perceive these structures as *bara zhatyr* – be lying and going, *oqyp zhy'r* – be reading and going, *aityp zhatyr* – be telling and lying, *oilanyp tu'r* – be thinking and standing, *qorqyp ty'r* – be afraid and standing. Consequently, it is imperative to compile and encode not only a single word, but also a sequence of phrases that are currently in use in relation to the markup provided in the Corpus.

A comparison with the Russian national corpus reveals that the marking of the series of compound words is given separately for each component in this corpus. A.V. Ventsov and E.V. Grudeva identified the rationale behind this approach: With regard to the analytical (morphological) forms in the NRC, it can be observed that the traditional analytical form is recognised as consisting of two words. As every lexical-grammatical word in the corpus is subject to morphological description, both of its components are marked as part of the analytical form. To illustrate, the analytical form *budu chitat'* comprises two words. The auxiliary verb *budu* is described as having the meanings of person, number, and so forth, while the infinitive *chitat'* is similarly defined. The same is true of the analytical forms of the comparative (*bolee sil'nyi*), superlative (*samyi bol'shoi*), subjunctive mood (*vzyal by*), passive (*byl sdelan*). The words *bolee, samyi* are assigned the descriptors of function words, while the word *by* is traditionally classified as a particle, the word *byl* is described as an auxiliary verb, the words sil'nyi, *bol'shoi, vzyal* and *sdelan* are described as ordinary adjectives, a subjunctive verb and a short passive participle. [5, 79-80].

In their article, "The issues of morphological marking the words in texts included in the corpus and their introduction into a computer program," S. Kulmanov, A. Zhanabekova, and several group members present a sample of morphological markups of compound nouns, adjectives, numerals, compound pronouns, adverbs, and verbs. It is asserted that all units are recognised as a single language unit and indicated as one wordform that are divided by the use of an underscore and a bold letter between them. However, it is stated that each of the compound units, with the exception of the paired words, is annotated individually. Additionally, the authors highlight that the stem and the grammatical meaning should be presented separately from the adverbial units [14, 110].

Adherence to this principle would ensure the differentiation and appropriate annotation of compound and analytical verbs.

The Kazakh National Corpus is one of the corpora that is currently undergoing development. It is also to be expected that the corpus will contain errors. The markup code is a construct that is created by the power of the hand. However, the requirements for a language corpus necessitate a high degree of discipline, accuracy, and precision, akin to that observed in the creation of dictionaries and encyclopedias. The National Corpus is an electronic database of information about the Kazakh language. It is clear that data pertaining to the style, theme, and author of the texts within the corpus are of significant importance. However, if the information at the structural level of the language is incorrect, the future of such a corpus is likely to be bleak. It is of paramount importance to correctly and fully set the markups related to the different layers of the language. The sign-code should be set in full without delay in order to forestall a recurrence of these adverse consequences.

In Kazakh, the lexical and semantic meaning of many compound structures is derived from the combination of multiple words, rather than a single unit. However, the phenomenon of homonymy in the language presents challenges in determining the function of certain words in context, that is, the features of their use, assigning them grammatical markup and giving them lexical and semantic meaning as a compound structure. A solution to this problem has been identified in the "Corpus of Standard Written Russian".

According to the developers of the Corpus, a joint analysis of complex structures helps to determine the frequency of their use and to make a correct morphological analysis.

ISSN 2788-7979 (online)

In the case of the analytical form of the future tense, the verb *by+t'* is marked as an auxiliary (VAX) with an appropriate morphological description, while the other component, bearing a lexical meaning and formally coinciding with the infinitive, is described as an ordinary infinitive. Consequently, the grammatical meaning of the verbal analytical form is presented solely at the level of the auxiliary verb.

A similar approach is applied to the analytical forms of a passive such as *by +l sde+lan*. In this case, the form from *by +t'* is marked as an auxiliary verb, while the second component (in this instance, sde+lan) is classified as a short participle. Analytical forms of the comparative degree of adjectives, adverbs and predicatives are also described according to the same principle of disconnection of components. The auxiliary component (sa+myi, bo+lee, me+nee) is described as a function word (AUX), and the second component is described as an ordinary adjective, adverb or predicative [5, 81].

It is established that the verb, a part of speech in the Kazakh language, is the most complex and voluminous. The subject of analytical forms of verbs also requires special analysis. The research conducted by Nurzhamal Oralbayeva on this topic is of significant value to the field of Kazakh linguistics. In the classification proposed by the scientist, the semantic aspects of analytical formants, which are particularly evident in our language, are also revealed. At the present time, all of the algorithms that have been introduced by humans are accepted by computer languages. It would represent a significant advance in the field of linguistics if a new branch were to consider the analytical unit in the National Corpus not as a standalone entity, but rather as part of a unified whole. To this end, we propose the issue of coding analytical formants. To illustrate, in our language, it is essential to elucidate the nature of the language from all perspectives in the National Corpus by coding a series of analytical formants that express the positive/negative category, the tense category, the mood category, and so forth. In this context, the works of N. Oralbayeva can be taken as a point of reference. The scientist also defined the number of compound variants of the analytical formants of verbs [15, 53]. It is our contention that the organisation of work on the automatic identification of these formants can be linked with the author's research. As compound analytical formants retain their full morpheme composition, they do not include other morphemes.

Even in the doctoral dissertation of the founder of the corpus A. Zhubanov, "Basic principles of formalizing the content of the Kazakh text," a classification of verbs was presented, distinguishing between lexical and semantic groups. The aforementioned verbs are grouped into the following macrogroups:

1. existential verb;
2. physical action verb;
3. movement verbs;
4. movement verbs that express the relations between people, animals and inanimate nature;
5. mental action verbs;
6. verbs related to the physical state and norm
7. verbs related to the mental state;
8. phase and modal verbs;
9. causative verbs [16, 78-88].

It would be beneficial for the language corpora if such lexical and semantic meanings of verbs were differentiated. Nevertheless, the article by Prof. B. Momynova, M. Imagazina, and U. Anesova, "Lexical-semantic development of verbs in the national corpus of the Kazakh language: world experience, classification, marking in the corpus," may serve as an evidence for the assertion that these markups will not be overlooked, and researchers are also working in this direction [17, 132].

The significance of the positioning of language markups within the corpus, as well as their quality, is also emphasised by Russian linguists. In their view, the morphological marking in the corpus is approached from two distinct perspectives. The first is formal-morphological, the second is profound semantic. The particular characteristics of morphological markups within the corpus are elucidated as follows: “The initial approach, which may be designated as "formal-morphological," postulates that each distinct word form identified within the text, which exhibits a divergence in

ISSN 2788-7979 (online)

appearance from other word forms, is assigned a specific label, irrespective of the underlying grammatical-semantic or syntactical-semantic information. The second approach, which may be termed "profound semantic," is designed to retrieve the most comprehensive semantic information associated with a given word form." [5, 76].

The initial linguistic studies conducted on corpora were constrained to the calculation of the frequency of occurrence of various language units (elements). In essence, the fundamental unit of analysis is a word, wordform and, in other instances, graphemes, morphemes and phrases. It is evident that when a corpus of voluminous texts is in use, the most straightforward action is to calculate the frequencies of different language units. Statistical methods are employed to identify solutions to complex linguistic problems. For example, tools for machine translation, automatic recognition of language elements and speech synthesis, and tools for checking spelling and grammar. It is thus possible to obtain information, such as the frequency of use of analytical structures over a certain period of time, the number of analytical structures that have appeared in the language, and whether this number has increased or decreased, from a qualitative linguistic corpus.

CONCLUSION

It is beyond question that the proper annotation of analytical forms of nouns and verbs in the Kazakh language in language corpora represents a significant challenge.

It is evident that the majority of auxiliary words in the paradigmatic series of word groups in morphology share the same lemma as the corresponding word. It would therefore be preferable to differentiate the grammatical functions of auxiliary words in corpora and to categorise them as grammes. This is, of course, a highly complex issue, akin to attempting to dig a well with a needle. From a linguistic perspective, referring to intricate analyses of morphological and word formation processes in this context yields insights. First and foremost, it establishes the number of morphological objects present in the sentence. Subsequently, the lexical meaning of the linguistic unit is distinguished, and then it is determined whether it is a root or a stem. Following this, the general grammatical meaning and categorical grammatical meaning are distinguished, after which the lexical-grammatical meaning is elucidated. Finally, the relative syntactic meaning is clarified. In accordance with the model of analysis, noun and verb analytical structures are regarded as a single morphological object.

However, when considering the extensive, fundamental research that has been conducted and is currently underway in this field (National Corpus of the Kazakh language), it is reasonable to conclude that the annotation of analytical structures, particularly analytical verb forms in accordance with the language structure, will eventually be automated and will become the most prominent corpus of Kazakh language corpora within the Turkic linguistic community.

«This research has been/was/ is funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. *BR21882334.* «Kazakh poetic corpus development: morphological and poetic designation of Abai's poems»),

REFERENCES

1 Zhubanov, A.K. and Zhanabekova, A.A. (2016), *Korpustyq lingvistika,* [*Corpus Linguistics*], [in Kazakh].

2 Beauzée, N. (1765), *Encyclopédie, ou Dictionnaire raisonné des sciences, des arts et des métiers par une Société des gens de lettres*. T.9, Chez Samuel Faulche et Compagnie, Libraires et Imprimeurs, Paris.

3 Rivarol, A. de. (1784), *Discours de l'Universalité de la langue Française,* Pierre Belfond, Paris.

4 Smith, A. *The Theory of Moral Sentiments or an Essay Towards an Analysis of the Principles by Which Men Naturally Judge Concerning the Conduct and Character, First of Their Neighbours, and Afterwards of Themselves, to Which is Added, a Dissertation on the Origin of Languages,* (1774), London.

5 Ventsov, A.V. and Grudeva E.V. *Analytical forms in Corpus of Standard Written Russian*, [Electronic resource], [in Russian], available at: https://events.spbu.ru/eventsContent/files/corpling (corpora 2006),
6 National corpus of the Kazakh language, [Electronic resource], [in Kazakh], available at: https://qazcorpus.kz/about/1/.
7 Subcorpora of National corpus of the Kazakh language, [Electronic resource], [in Kazakh], available at: https://qazcorpora.kz/search].
8 National Corpus Tatar Language "Tugan Tel", [Electronic resource], [in Tatar], available at: https://tugantel.tatar.
9 Almaty Corpus of the Kazakh Language, [Electronic resource], [in Kazakh], available at: http://web-corpora.net /].
10 Bashkir poetic corpus, [Electronic resource], [in Bashkir], available at: http://web-corpora.net/bashcorpus].
11 Corpus of Standard Written Russian, [Electronic resource], [in Russian], available at: https://narusco.ru/search.
12 Altayeva, A. (2006), *Komekshi etistikterdin' semantikasy*, [*Semantics of auxiliary verbs*], Almaty, [in Kazakh].
13 Zhubanov K. (2010), *Qazaq tili zho'nindegi zertteuler,* [*Research on the Kazakh language*], [in Kazakh].
14 Kulmanov S. Zhanabekova A., Ashimbayeva N., Bisengali A., Shulenbayev N. and Kordabai B. *Problems of morphological markup of words in corpus texts, and their inclusion in a computer program*, Bulletin of the Gumilyov National University, Philology series, vol. 140, n. 3, P. 103-117, [in Kazakh].
15 Oralbai N. (2007), *Qazirgi qazaq tilinin' morfologiyasy*, [*Morphology of the modern Kazakh language*.], Almaty, 390 p. [in Kazakh].
16 Zhubanov A.K. (2015), *"National corpus of the Kazakh language and the problem of metamarking"*, Journal "Tiltanym" of A. Baitursynov Institute of Linguistics, vol. 57, n. 1, P 23-33[in Kazakh].
17 Momynova B., Imangazina M., and Anesova U. (2022), *Lexical-semantic development of verbs in the national corpus of the Kazakh language: world experience, classification, marking in the corpus"* The Kazakh Ablai Khan University of International Relations and World Languages, Philology series, vol. 66, n. 3, P.128-146.

Received: 25.07.2024

Аналитикалық құрылымдардың тіл корпустарында аннотациялануы

Г.Б. Мадиева[1], Л.Т. Алимтаева[1], Ж.Б. Саткенова[1], К.Қ. Пірманова[1]
[1]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, 050000, Қазақстан Республикасы

Тіл корпустарындағы аналитикалық құрылымдардың морфологиялық және сөзжасамдық белгіленімдерін саралау мақсатымен қазақ, түркі, орыс тілдері корпустарындағы белгіленімдерге шолу жасалды. Жетекші түркі тілдерінің корпустары біршама бір-бірімен (қазақ, татар, башқұрт) деңгейлес екендігі анықталды.

Сөзжасам тұрғысынан машинамен бір күрделі бірлік ретінде танылатын біріккен және қосарланған сөздердің аннотациялануында айтарлықтай проблема жоқ екендігіне, яғни сол қалпымен іздеу нысаны бола алатындығы және морфологиялық, сөзжасам, лексикалық белгіленімдерінің тұтас бір күрделі бірлікке сәйкес берілетіндігі айқындалды.

Бос аралықпен жазылатын тіркесті күрделі бірліктерді сол қалпымен бір лемма ретінде іздеу мүмкіншілігінің жоқ екендігі сонымен қатар олардың морфология-сөзжасам-лексикалық белгіленімдері тұтас күрделі бір бірлік ретінде емес, әр сыңары жеке аннотацияланатындығы анықталды. Бұл мәселе корпустардың функционалдығын

ISSN 2788-7979 (online)

тежейтіндігі белгілі. Дегенмен қазақ тілінің ұлттық корпусында фразеологизмдердің бір бүтін бірлік ретінде аннотациялануы, аталған корпустың артықшылығы деп танылды.

Есімді (шырай), етістікті аналитикалық формалар мен форманттардың, көмекші сөздердің де барынша тіл құрылысына сәйкес аннотациялануы корпус қызметін жақсартатыны сөзсіз. Бұл бағыттағы жұмыстар күрделілігмен ерекшеленеді. Дегенмен қазақ тілінің ұлттық корпустарының қол жеткізген қазіргі сөздік қорына және зерттеушілердің болашаққа бағдарланған ауқымды, іргелі зерттеулеріне қарап, аналитикалық құрылымдардың да аннотациялануы бірте-бірте автоматтандырылатындығы уақыт еншісінде ғана деген пайымдау жасалды.

Кілт сөздер: қазақ тілінің корпусы, есімді-етістікті аналитикалық құрылымдар, аналитикалық форма, морфологиялық белгіленім, сөзжасамдық белгіленім.

ӘДЕБИТТЕР ТІЗІМІ

1 Жұбанов А.Қ., Жаңабекова А.Ә. Корпустық лингвистика. – Алматы.: Қазақ тілі, 2016. – 336 б.
2 Beauzée N. Encyclopédie, ou Dictionnaire raisonné des sciences, des arts et des métiers par une Société des gens de lettres. T.9. Paris: chez Samuel Faulche et Compagnie, Libraires et Imprimeurs. – 1765. – 307 p.
3 Rivarol A. de. Discours de l'Universalité de la langue Française. – Pierre Belfond. – 1784.
4 Smith A. The Theory of Moral Sentiments or an Essay Towards an Analysis of the Principles by Which Men Naturally Judge Concerning the Conduct and Character, First of Their Neighbours, and Afterwards of Themselves, to Which is Added, a Dissertation on the Origin of Languages. – London, 1774. – 478 p.
5 Венцов А.В., Грудева Е.В. Орыс әдеби тілінің ұлттық корпусындағы аналитикалық формалар [Электрон. ресурс].https://cyberleninka.ru/article/n/aktsentno-razmechennyy-korpus-russkogo-literaturnogo-yazyka-kak-istochnik-novyh-slovarey-slovar-omografov-russkogo-yazyka-i-chastotnyy.pdf (дата обращения: 01.07.2024).
6 Қазақ тілінің ұлттық корпусы [Электрон. ресурс]. URL: https://qazcorpus.kz/about/1/ (дата обращения: 01.07.2024).
7 Қазақ тілі ұлттық корпусының кіші корпустары [Электрон. ресурс]. URL: https://qazcorpora.kz/search (дата обращения: 01.07.2024).
8 «Туган тел» Татар тілінің ұлттық корпусы [Электрон. ресурс]. URL: https://tugantel.tatar (дата обращения: 01.07.2024).
9 Алматы қазақ тілінің корпусы [Электрон. ресурс]. URL: http://web-corpora.net /KazakhCorpus/search/?interface_language=kz (дата обращения: 01.07.2024).
10 Башқұрт тілінің поэтикалық корпусы [Электрон. ресурс]. URL:http://web-corpora.net/bashcorpus (дата обращения: 01.07.2024).
11 Орыс әдеби тілінің корпусы [Электрон. ресурс]. URL: https://narusco.ru/search(дата обращения: 01.07.2024).
12 Алтаева А. Көмекші етістіктердің семантикасы. – Алматы.: Арыс, 2006. – 90 б.
13 Жұбанов Қ. Қазақ тілі жөніндегі зерттеулер. – Алматы.: Мемлекеттік тілді дамыту инты, 2010. – 607 б.
14 Құлманов С., Жаңабекова А., Әшімбаева Н., Бисенғали А., Шүленбаев Л.Н., Қордабай Б. Л. Корпусқа енгізілетін мәтіндердегі сөздерге белгіленім қою және оларды компьютерлік бағдарламаға енгізу мәселелері // Гумилев ат. ЕҰУ Хабаршысы, фил.сер. – 2022. – Т. 140. №3. – Б. 103-117.
15 Оралбай Н. Қаз. – Алматы.: – Ғылым, 1979. – 196 б.
16 Жұбанов А.Қ. Қазақ тілінің ұлттық корпусы және метабелгіленім мәселесі // А.Байтұрсынұлы ат. ТБИ «Тілтаным» журналы, - 2015. – Т. 57. №1. Б.– 23-33.

ISSN 2788-7979 (online)
17 Момынова Б., Иманғазина М., Анесова Ү. Қазақ тілінің ұлттық корпусындағы етістіктердің лексика-семантикалық әзірлемесі: әлемдік тәжірибе, классификациялау, корпуста белгілеу//Абылайхан ат. ҚХҚТУжӘТУ Хабаршысы, фил.сер. – 2022. – Т. 66. №3. – Б, 128-146.

Материал 25.07.2024 баспаға түсті

Аннотация аналитических структур в языковых корпусах

Г.Б. Мадиева[1], Л.Т. Алимтаева[1], Ж.Б. Саткенова[1], К.К. Пирманова[1]
[1]Казахский национальный университет имени аль-Фараби, Алматы, 050000, Республика Казахстан

Наряду с целью дифференциации морфологических и словообразовательных разметок аналитических структур в корпусах казахского языка, был выполнен обзор разметок аналитических структур в языковых корпусах других тюркских языков. Было проведено сравнение с корпусом русского языка, который находится в ряду сбалансированных, объемных корпусов. В качестве объекта исследования проанализированы материалы существующих корпусов современного казахского языка.

Установлено, что с точки зрения словообразования не существует особой проблемы с аннотированием сложных и парных слов, которые определяются машиной как одна сложная единица, т.е. могут быть объектом подобного поиска и представлять в соответствии с морфологической, словообразовательной, лексической разметками единую сложную единицу. Было обнаружено, что нет возможности искать составные единицы словосочетания, записываемые с пустыми интервалами, как одну лемму, наряду с этим их морфологические, словообразовательные, лексические разметки аннотируются индивидуально, а не как единое целое. Известно, что эта проблема снижает функциональность корпусов. Однако аннотирование фразеологизмов как одной целостной единицы признается преимуществом этого корпуса. Несомненно, что аннотирование имен (степеней), глагольных аналитических форм и формантов, вспомогательных слов также максимально улучшает функции корпуса. Современный словарный фонд национальных корпусов казахского языка и масштабные, фундаментальные, ориентированные на будущее исследования ученых, позволяют сделать вывод о том, что аннотирование аналитических структур с течением времени постепенно автоматизируется.

Ключевые слова: корпус казахского языка, номинативно-глагольные конструкции, аналитическая форма, морфологическая разметка, словообразовательная разметка.

СПИСОК ЛИТЕРАТУРЫ

1 Жубанов А.К., Жанабекова А.А. Корпусная лингвистика. – Алматы.: Қазақ тілі, 2016. – 336 с.
2 Beauzée N. Encyclopédie, ou Dictionnaire raisonné des sciences, des arts et des métiers par une Société des gens de lettres. T.9. Paris: chez Samuel Faulche et Compagnie, Libraires et Imprimeurs. – 1765. – 307 p.
3 Rivarol A. de. Discours de l'Universalité de la langue Française. – Pierre Belfond. – 1784.
4 Smith A. The Theory of Moral Sentiments or an Essay Towards an Analysis of the Principles by Which Men Naturally Judge Concerning the Conduct and Character, First of Their Neighbours, and Afterwards of Themselves, to Which is Added, a Dissertation on the Origin of Languages. – London, 1774. – 478 p.
5 Венцов А.В., Грудева Е.В. Аналитические формы в национальном корпусе русского литературного языка [Электрон. ресурс]. URL: https://cyberleninka.ru/article/n/aktsentno-razmechennyy-korpus-russkogo-literaturnogo-yazyka-kak-istochnik-novyh-slovarey-slovar-omografov-russkogo-yazyka-i-chastotnyy.pdf.

ISSN 2788-7979 (online)
6 Национальный корпус казахского языка [Электрон. ресурс]. URL: https://qazcorpus.kz/about/1/ (дата обращения: 01.07.2024).
7 Малые корпуса национального корпуса казахского языка [Электрон. ресурс]. URL: https://qazcorpora.kz/search (дата обращения: 01.07.2024).
8 Национальный корпус татарского языка «Туган тел» [Электрон. ресурс]. URL: https://tugantel.tatar (дата обращения: 01.07.2024).
9 Корпус казахского языка Алматы [Электрон. ресурс]. URL: http://web-corpora.net /KazakhCorpus/search/?interface_language=kz (дата обращения: 01.07.2024).
10 Поэтический корпус Башкирского языка [Электрон. ресурс]. URL: http://web-corpora.net/bashcorpus (дата обращения: 01.07.2024).
11 Корпус русского литературного языка [Электрон. ресурс]. URL: https://narusco.ru/search(дата обращения: 01.07.2024).
12 Алтаева А. Семантика вспомогательных глаголов. – Алматы.: Арыс, 2006. – 90 с.
13 Жубанов Қ. Исследования по казахскому языку. – Алматы.: Институт по развитию государственного языка, 2010. – 607 с.
14 Кулманов С., Жанабекова А., Ашимбаева Н., Бисенгали А., Шуленбаев Л.Н., Кордабай Б. Л. Проблемы морфологической разметки слов в текстах корпуса и их включения в компьютерную программу // Гумилев ат. ЕҰУ Хабаршысы, фил.сер. – 2022. – Т. 140. №3. – С. 103-117 с.
15 Оралбай Н. Структура и семантика аналитических формантов глаголов в казахском языке. – Алматы.: – Наука, 1979. – 196 с.
16 Жубанов А.К. Национальный корпус казахского языка и проблемы метаразметки // ИЯ им. А.Байтұрсынулы, научный ж у р н а л «Языкознание», - 2015. – Т. 57. №1. С. – 23-33.
17 Момынова Б., Имангазина М., Анесова У. Разработка лексико-семантической разметки глаголов в национальном корпусе казахского языка: мировой опыт, классификация, разметка в корпусе // КУМОиМЯ им. Абылайхана, фил.сер. – 2022. – Т. 66. №3. – С. 128-146.

Материал поступил в редакцию журнала 25.07.2024