BILINGUAL PUBLISHING GROUP

Forum for Linguistic Studies

https://journals.bilpubgroup.com/index.php/fls

ARTICLE

Usage of the Statistical Method to the Lexico-Morphological Structure of the Language

Pirmanova Kunsulu [1*] **, *Tokmyrzayev Darkhan*** [2] **, *Zaurbekova Gulkhadisha*** [3] **, *Madiyeva Gulmira*** [1] **, *Serikbayeva Aizhan*** [1] **, *Sabirova Dina*** [4]

[1] *Faculty of Philology, Al-Farabi Kazakh National University, Almaty 050010, Kazakhstan*

[2] *Space Technology Center, Al-Farabi Kazakh National University, Almaty 050010, Kazakhstan*

[3] *Faculty of Bussines, Media and Administration, International IT University, Almaty 050010, Kazakhstan*

[4] *Faculty of Philology, Abai Kazakh National Pedagogical University, Almaty 050010, Kazakhstan*

ABSTRACT

All words in the modern Kazakh language are divided into three large groups: naming words, auxiliaries, simple words. These groups are further differentiated according to lexico-grammatical aspects, where naming words are first divided into naming parts of speech and verbs and then the naming parts of speech themselves are further divided into naming nouns and naming adverbs. It is known that naming words are divided into nouns, adjectives, numerals and pronouns and adverbs into adverbs and similes. To summarize, words in the modern Kazakh language are divided into 9 groups called “parts of speech”: nouns, adjectives, numerals, pronouns, verbs, adverbs, imitation words, auxiliary words, interjections. The content of the parts of speech consists of a unit of two components called “lexical-grammatical”. These components encompass both lexical and grammatical semantics. The content of the word “grammatical” in the second part of the double word includes the meanings of the grammatical categories specific to a particular word, as well as their forms and transformations. This means that the term “grammatical” here encompasses the meanings and individuals of the systems of word formation, word transformation, word combination and thus all morphological and syntactic signs (secret features) of the whole word.

Keywords: Modern Kazakh Language; Statistical Linguistics; Lexico-Grammatical Part of Speech; Language Units; Parts of Speech

*CORRESPONDING AUTHOR:

Pirmanova Kunsulu, Faculty of Philology, Al-Farabi Kazakh National University, Almaty 050010, Kazakhstan; Email: dtokmyrzaev@mail.ru

ARTICLE INFO

Received: 14 January 2025 | Revised: 25 February 2025 | Accepted: 28 February 2025 | Published Online: 18 March 2025
DOI: https://doi.org/10.30564/fls.v7i3.8419

CITATION

Kunsulu, P., Darkhan, T., Gulkhadisha, A., et al., 2025. Usage of the Statistical Method to the Lexico-Morphological Structure of the Language. Forum for Linguistic Studies. 7(3): 890–905. DOI: https://doi.org/10.30564/fls.v7i3.8419

1. Introduction

One of the aims of the applied field of Kazakh language, namely statistical linguistics, is to statistically explore the meaning, morphological form (identity) and syntax of the word on the basis of the lexical (semantic), grammatical (morphological and syntactic) signs of the parts of speech, i.e., on the basis of these three signs — to be able to recognize the function of the system on the basis of digital data. As a result, the formal aspects of the indexing of the text content are defined[1].

The types of subclassified parts of speech classes can be used more or less in a language. According to the experience of statistical research, it is necessary to launch programs that specify some measures to determine the frequency of their use[2].

For example, if we need to know the statistical data about the relational character of word classes in different language styles, we should think of convenient ways for the researcher to set a conventional sign (code) for each word class. This is because the possibility of formally distinguishing word classes or other linguistic units within a text according to their type is still outside the scope of research. Therefore, the assignment (coding) of the aforementioned linguistic units and their internal components to conventional symbols is very important for the cognitive role in carrying out statistical counts by computer. A universal approach should be adopted in coding linguistic units[3].

Therefore, it is possible to use the model of "encoding" in research to distinguish the above parts of speech of Kazakh and to determine their statistics, which we consider below:

(1) by number: Noun — 01, Adjective — 02, Number — 03, Pronoun — 04, Verb — 05, Pronoun — 06, Preposition — 07, Adverb — 08, Preposition — 09, Imitation word — 10, Modal words — 11;
(2) by letter: Noun [zat esim] — zt, adjective [syn esim] — sn, numerals [san esim]— sa, pronoun [esimdik] — es, verb [etistik]— et, participle [esimşe]— eş, preposition [kösemşe] — kş, adverb [üsteu]— üs, imitation words [elikteuiş sözder] — el, modal words [modal sözder]— md;
(3) mixed by marking numbers and letters (mixed method): Noun — z1, adjective — s1, number — n2, verb — e1, pronoun — e2, participle — e3, imitation words — e4, preposition — k1, adverb — y1, modals — м1.

There are many ways to set conventional signs (codes), although it is considered better to strive for consistency (standard) than to set them according to everyone's wishes. Morphological, syntactic and semantic branching types and modes of creation within each word class can be of different nature depending on the objective of the particular researcher. Therefore, each word, sentence, phrase, clause or larger unit of text may contain different grammatical information. The adoption of conventional character codes can be organized as a series of "slots" corresponding to multiple pieces of information in a single linguistic unit[4].

Of course, in most cases such a complex character code is entered into the text, depending on the direction of statistical research. In the practice of researching the Kazakh language using the statistical method, various ways of introducing conventional signs into the text in order to distinguish the character of use (frequency) of word classes in our language have been shown above.

With the aim of complicating such a method of marking, distinguishing the morphological types and structure of each word class, taking into account the types of formation, i.e., with the aim of maximally capturing other grammatical information associated with this linguistic unit and determining the statistics of each of them, we will focus on some models of "coding" in more detail in our further report[5].

The peculiarity of our proposed model for creating a program for matching morphological information signs is that it takes into account the structural nature of the internal parts of the word classes of the Kazakh language and the way they are formed, and it aims at universality[3].The main parts of speech of the Kazakh language, such as nouns, verbs, adjectives, numerals and pronouns, were made the subject of the program creation. The program for assigning morphological information to these parts of speech using the character-code model was reproduced in the corresponding tables[6].

2. Materials and Methods

There are no unique morphological features for nouns. It is known that they are transformed according to their lexical-semantic properties by special conjunctions and un-

dergo a personal change by suffixes and interact freely with other words in the sentence. Their mysterious character can only be fully revealed by looking at the morphological, syntactic and semantic features of nouns. These features include the nature of their frequent or infrequent use in the text [7].

The grammar of the modern Kazakh language distinguishes the following groups with semantic and grammatical features unique to nouns: nominative (who) and common (what) nouns, plural nouns, emotional-expressive nouns, auxiliary nouns. It is also known that nouns can be grouped according to their morphological characteristics, i.e., their structure, the way they are formed and the system of transformation. Thus, it can be said that it is very important to study the Kazakh language using the methods of computational and statistical linguistics [8].

The complexity of the composition of verbs in comparison to other parts of speech is closely related to their extensive lexical-semantic character, their richness of lexical-grammatical forms and their comprehensive syntactic function. The reason for the richness of the lexico-grammatical characters of the verb is that they include the time of the action, the period of its execution, the type of transition, i.e., the direction, the tempo, the rest of the action. If we simultaneously include the systems of synthetic and analytical methods of forming verbs from words related to other parts of speech, we can say that there is no parts of speech that can match the richness of personalization of a verb. This can be seen in its unique lexical-semantic, lexical-grammatical and grammatical categories. The meanings and syntactic functions of these different categories of verbs become clear when they come into contact with words of other parts of speech.

Although each word in the lexical system of the verb has its own unique lexical meaning, it is categorized into a word group according to its semantic features. In order to correctly determine the grammatical persons and functions of the verb, it is necessary to classify and group them according to the semantic structure of the root forms, derived synthetic and analytic forms to reveal their nature and determine their personal and content relations. This is because grammatical semantics is very closely linked to the composition of the word [9].

In the field of morphology of the modern Kazakh language, the part of speech adjective is defined as follows: *"The lexico-grammatical part of speech expressing the quality, nature, property, volume, weight, color (type-color) and other characteristics of an object is called adjective"* ([1], pp. 166–172, 384). As far as morphological features are concerned, it is known that adjectives have unique features compared to other nouns. For example, in addition to suffixes that form both common and personal words with nouns, adverbs, and other parts pf speech, there are special suffixes that form only derived adjectives, and adverbial suffixes that form adjectives from adjectives.These are also considered morphological forms of adjectives.

Another special feature of adjectives is that they can express various qualities and features of things directly, but also through their relationship to other things. In this context, it is also known that adjectives are divided into two branches according to their semantic meaning and grammatical features, which are called qualitative (basic) and relative (derived) adjectives [10].

What we mean by basic or qualitative adjectives is that they usually consist of root words without suffixes that express different properties of things. However, if we look at some words that are considered root words from the perspective of their historical development, they may also turn out to be derived words. Therefore, both the number and the quality of words that are considered base adjectives in terms of their structure are constantly changing. And it should be remembered that the classification of adjectives into base and derived words according to their morpheme structure is conditional.

Thus, the adjective words that occur in any sentence (in the text) can belong, firstly, to the qualitative adjectives (white, black, yellow, blue, gray, high, large, low, warm, cold, large, small, etc.), secondly, according to the morphological (synthetic) method, i.e., the derivative formed by corresponding suffixes can belong to the adjective.

Thirdly, the next adjective to be considered is a syntactic (analytic) approach, i.e., by combining simple adjectives (white-yellow, red-yellow, white-chubar, white dress, big-small, etc.) and derived adjectives can also be formed according to the morphological-syntactic (semantic) approach.

It can even be said that adjectives formed according to the latter method are unproductive from the point of view of usage. It can therefore be said that the most basic method of forming adjectives is first the morphological or synthetic

method and then the analytical, i.e., the syntactic method.

For this reason, we will base the program for identifying the information symbol code on the lexico-morphological structure of the adjective only on the way of forming the adjective in a synthetic way, more precisely on the way of forming the adjective from the nouns with the help of productive suffixes [11].

To study the lexical and morphological structure of the Kazakh language, a research methodology was developed that allows achieving accuracy, objectivity and repeatability of results. The study consists of several stages: primary classification of words, their coding, data verification and statistical analysis. Each stage was carefully planned to minimize errors and increase the reliability of the results [12].

2.1. Word Classification

Words were assessed individually by two independent experts, which made it possible to avoid subjectivity and increase the accuracy of the results. Working with text materials, specialists divided words into nine main word classes (nouns, adjectives, numerals, verbs, etc.) and their subtypes. For example, nouns are divided into "human nouns" (who?) and "universal nouns" (what?), as well as into morphological categories (root morphemic noun, derivative noun, compound noun, etc.).

To reduce the number of errors, cross-checking of the data was carried out. The classification results of two experts were compared. In case of disagreement, a third expert was involved in making the final decision. This ensured a high level of data consistency.

2.2. Coding Rules

Three different markers were used to aggregate the data: numeric, alphanumeric, and mixed. The choice of marker type depended on the purposes of the analysis and ease of use. For example:

a. A noun is coded as "01" (numeric), "zt" (alphabetic), or "z1" (mixed).
b. An adjective is coded as "02", "сн", or "c1".
c. A verb is coded as "05", "et", or "e1".

Morphological features of words were taken into account during coding. For example, for nouns, additional codes were considered indicating their type (human names — "020", universal names — "021") and grammatical characteristics (conjugated/unconjugated form, proper name, plural name, etc.). A similar approach was applied to other word classes.

2.3. Analysis Tools

The statistical analysis was performed using R and Python software tools. These tools allowed us to automate the data processing process, identify frequency patterns of word forms, and create graphs to visualize the results. For example, R was used to calculate the frequency distribution of words by word classes and determine correlations between different morphological categories.

Python scripts were used to check the data, automatically comparing the encoded values with the original texts. This allowed us to identify potential errors at the coding stage and correct them in a timely manner. In addition, machine learning methods allowed us to speed up the analysis of large text corpora.

2.4. An Example of Using the Methodology

Let's consider this sentence as an example:

> "When Asan went to school, he carefully packed his books, pens, notebooks, and other supplies into his bag."

1. Word classification:
   - Asan is a noun (human names).
   - To school is a noun (common noun).
   - Barарda is a verb.
   - Book, pen, notebook are nouns (proper names).
   - Accessories are a noun (compound noun).
2. Coding:
   - Asan: ZT/020/030/040/050/060/070/080.
   - To school: ZT/021/031/040/050/060/070/083.
   - Book: ZT/021/031/040/050/060/070/080.

Using R software, a word frequency distribution by word class was created. For example, the noun turned out to be the most frequently occurring word class in this text, which corresponds to general patterns in the Kazakh language.

The proposed methodology allows for precise classification and coding of words, and helps to identify hidden patterns in their use. Automation of data analysis and verification processes allows for scaling up the research and applying it to large volumes of text.

3. Results

The most frequently used part of speech in Kazakh language texts are nouns and verbs. Therefore, one of the goals of the "Applied Linguistics" discipline is to understand the statistics of their morphological and syntactic changes in the text. The central focus of article is the stabilization of the place ('slot') of grammatical information corresponding to the word, phrase, or phraseological units found in the text—referred to as the text unit or word usage—and the conventional sign-code that represents that information. We think that it is essential to thoroughly consider the conditional parameters (such as 'nest' location and sign-code) matched to each of the nouns of any linguistic unit, for example, which we are going to consider first and foremost.

For example, the number of 'slots' or the number of groups matching nouns to words is eight, i.e., n = 1, 2, 3, 4, 5, 6, 7, 8. In each 'slot', there is a conventional sign-code number (set number), and the 'code value' may change depending on the characteristics of the nouns considered in that 'slot'.

If we refer to the table of data containing these statements as the "program", a concise collection related to the noun is presented in the form of a "program" in **Table 1**.

Table 1. Nouns in the form of a "program".

Assan:	ЗТ/	020	030	040	050	060	070	080
мектепке/mektepke [to school]:	ЗТ/	021	031	040	050	060	070	083
кітап/kitap [book]:	ЗТ/	021	031	040	050	060	070	080
қалам/Qalam [pen]:	ЗТ/	021	031	040	050	060	070	080
дәптер/däpter [notebook]:	ЗТ/	021	031	040	050	060	070	080
керек-жарақтарын/ kerek-karaqtaryn [accessories]:	ЗТ/	021	031	042	050	060	073	084
сөмкесіне/sömkesine[in his bag]:	ЗТ/	021	031	040	050	060	070	083

Let's comment on some of the details in the table.

This table consists of three "columns":

(1) location of conventional signs (number of the batch);
(2) lexical-morphological nature of nouns;
(3) conditional sign-code.

It should be assumed that the first set of 'slots' related to each parts of speech is marked (encoded) with the shortened name of that parts of speech and occupies the 1st position. Therefore, the very first 'slots' in the chain (the first column) is equal to the number "1" (the first set). In this row, the second column contains information about the nature of the word class, i.e., the name of the noun, and the third column contains the value of the "sign-code" of the information – "ZT". Similarly, in the number "2" position of conventional signs, in the second group, two different descriptions of nouns are given. The first of them is human names (who?), and the second is universal names (what?). Depending on which question (who? what?) the next noun to be considered in the position of the conditional sign, i.e., in the 2nd place, only one of the sign-codes corresponding to it in the third column is selected. **Table 1** shows 8 such groups related to nouns.

In the seventh group, six different characteristics of the noun word are mentioned: 1.Noun with base morpheme; 2. Derived noun (base + suffix); 3. Complex noun: (a) closed compound words; (b) open compound words; (c) hyphenated compound words; (d) shortened words.

In the third column, their sign-code values are given by three-digit numbers: 070, 071, 072, 073, 074, 075 (see **Table 2**).

As mentioned above in connection with the 2nd place, similarly, when conditionally "coding" the next word under study, only one of the six sign-code values corresponding to the 7th place (between 070 and 075) is taken, which is characteristic of the character of the word. The other cluster locations listed in the table can also be explained in the same way.

Now, as an example, let's try to write some noun words in the form of conventional sign-code based on the data in **Table 2**. For instance, the nouns in the following sentence should be separated and written using a sign-code: "Асан

мектепке барарда кітап, қалам, дәптер және басқа да керек-жарақтарын аса ұқыпты-лықпен сөмкесіне салды”/ “When Assan went to school, he carefully put his books, pens, notebooks, and other accessories in his bag”. The nouns in this sentence are: Assan; to school; book; pen; notebook; accessories; in his bag.

Table 2. Program for matching lexical-morphological information to nouns.

Symbol-Code Location	Lexical-Morphological Character	Symbol-Code
1	Noun word	ЗТ/ZT
2	Human names (Who?)	020
	Universe names (what?)	021
3	Nominative name: a) nominative names (onomastics); b) geographical names (toponymy)	030
	Common names	031
	Mixed meaning (moon, sun, earth)	032
4	Words with no plural meaning (considered without the end of words)	040
	Plural-Only nouns (aiaq [feet], qears, boots)	041
	Words with plural meaning (without the end of words)	042
5	Colorless noun	050
	An emotional-expressive noun (Äkei, Sauleş)	051
6	Nouns that not enter to auxiliary noun	060
	Auxiliary noun (aldy [front], arty [back], qasy [side])	061
7	Noun with base morpheme	070
	Derived noun (base + suffix)	071
	Complex noun: a) closed compound words	072
	b) open compound words	073
	c) hyphenated compound words	074
	d) shortened words	075
8	Changing forms of noun:	
	The infinitive form of the noun without suffixes (who? what?)	080
	With plural endings: -лар/-lar, -лер/-ler,-дар/-dar,-дер/-der,-тар/-tar,-тер/-ter	081
	Through subjunctive conjunctions: -нікі/-nıkı,-дікі/-dıkı,-тікі/-tıkı, -м/-m,-ым/ym, -ім/-ım; -ң/-ň, -ың/-уň, -ің/-ıň, -ңыз/-ňyz, -ңіз/-ňız, -ыңыз/-уňyz, - іңіз/-ıňız; -сы/-sy, -сі/-sı, -ы/-y,-і/-ı	082
	Through case declensions of noun: ілік/ılık [genetive] (кімнің?/kımnıň? [whose], ненің?/nenıň?[whats?]); Барыс/barys (кімге?/kımge?[for whom?], неге?/nege? [why?], қайда?/qaida? [where]); табыс/tabys (кімді?/kımdı? [whom], нені?/nenı? [what?]); Жатыс/jatys (кімде?/kimde? [who got?], неде?/nede? [on where]); шығыс/şygys [ablative] (кімнен?/kimnen? [from whom?], неден?/neden? [from what?]); Көмектес/kömektes (кіммен?/kimmen? [with whom?], немен?/nemen? [with what?]).	083
	Through personal endings: singular: -мын/-myn, -мін/-min, -сың/-syň, -сің/-siň, -сыз/-syz, -сіз/-siz; plural: -мыз/-myz, -міз/-miz.	084

Let’s match the conditional sign-codes to these words. That is, it is necessary to place the appropriate sign-code value in the eight-place ‘slots’ for each of the mentioned nouns. It is evident that all of them have the symbol-code “ZT/” in the first place.

Now, if we consider the word “Assan” separately, it is in the 2nd place — “020” (because it is a human name), in the third — “030”, in the fourth — “040”, in the fifth — “050”, in the sixth — “060”, in the seventh — “070”, in the eighth — “080”. So, the ‘code’ spelling of the word “Assan” and other nouns in the sentence will be as follows:

In the indicated way, let’s match the idioms related to these nouns with the sign-code: *scientist, Assandar, my ear, to Almaty, writer, mushrooms, my soul, to my walking,* which presented in the **Table 3**.

It is possible to create frequency dictionaries by converting specific text characters to symbol-code type or by arranging characters (units) in alphabetic-frequency, frequency, or reverse-alphabetic dictionaries created from the text in letter form by changing it to symbol-code type. If we write such dictionaries in a ‘compressed’ form using an appropriate computer program, it is possible to obtain a ‘characteristic frequency dictionary’ of nouns (or other word classes). Such a frequency dictionary will undoubtedly be important material necessary for the study of the morphological structure of the Kazakh language and further text linguistics [13].

Revealing the secrets of the morphological structure of verb forms approaches the scope of research in the field of contextual or textual linguistics. From this perspective, organizing research involves extensive use of statistical linguistics methods. It should be understood that one of the reasons for the complex and rich nature of the semantic structure of derivative verbs lies in their rare or frequent use [14].

Table 3. Matching of idioms.

Ғалым/galym [science]	ЗТ/ZT	020	032	040	050	060	070	080
Асандар/Assandar	ЗТ/ZT	020	030	040	050	060	070	081
Құлағым/qylagym [ears]	ЗТ/ZT	021	032	041	050	060	070	082
Алматыға/Almatyga [to Almaty]	ЗТ/ZT	021	030	040	050	060	070	083
Жазушы/jazyşy [writer]	ЗТ/ZT	020	032	040	050	060	071	080
саңырау құлақтан [from mushrooms]	ЗТ/ZT	021	032	041	050	060	074	083
жаным-ай [my soul]	ЗТ/ZT	021	032	040	051	060	070	082
Жүрісіме [to my walking]	ЗТ/ZT	021	032	042	050	060	071	083

Therefore, we believe it is necessary to consider the morphological structure of derivative verbs as a separate system, group them according to their characteristic features, and introduce a special sign-code system to reveal their probabilistic-statistical patterns. As mentioned above, the problem of recognizing linguistic units, such as word classes of the Kazakh language, through formal signs and automatically separating them from written or spoken text is not resolved in relation to nouns. Conducting statistical research on the morphological structure of the verb based on the conventional sign-code uploaded "mechanicly" reveals many aspects, including the problem of automatic separation and categorization of text units. At the same time, aspects of the lexical-grammatical structure of the verb, which cannot be observed in the traditional method of research, can be revealed (identified). Considering modern Kazakh verbs (base and derivative) in this context, we divide them into groups based on their closeness in terms of meaning, compatibility in terms of mutual functions, and lexical-morphological structure. We offer a code-matching program [15].

For example, from the point of view of meaning, verbs can be divided into the following groups [16]: амал-әрекет/action, қимыл-қозғалыс/movement, қалып-сапа/quality, ойлау-сөйлеу/speech verbs, өсу-өну/growth-development verbs, and so on, as shown in **Table 4**.

As a rule, it is known that derived words have a certain morphological structure formed according to one system. The first component of that structure is a word with an independent meaning, and the second component is a suffix that turns the first component into a certain word class (e.g., тіс[teeth]+те[te] = bite, кел[come]+тір[tir] = make it done, сабын[soap]+да[da] =wash with soap). Derivative verbs are divided into two branches, verbs formed from nouns and verbs. They are called noun-based verbs and verb-based verbs. We have considered only the first of these, i.e., the ways of forming verbs based on nouns, separated into 16 groups, and we have presented a sample of marking them using a conventional sign-code using **Table 4**.

Similar to the program for matching lexical-morphological information to nouns, the program in **Table 4** also consists of three columns:

(1) symbol-code position; (2) the nature of noun-based derived verbs; (3) sign-code symbol.

The sign-code position of the first column is labeled as "1" (or Group №1), and its description in the second column is termed 'verb word'. If the next word to be studied throughout the text belongs to a verb, then the symbol-code of that verb is 'ET' in the third column. In the second place, or in the second column of the second group, there is a description of each of the 11 groups of verbs that are closer to each other in terms of meaning and functionality. In the third column, you can distinguish the sign-code symbol corresponding to these characteristics. For example, the third part of the "2nd" place (2nd set) of the sign-code is called quality verbs, and the corresponding third column contains the three-place sign-code of the same description – '102'. Similarly, let's provide an example of how derivative verbs are formed from nouns using suffixes. The formation of derived verbs with the suffix -ла/-la (-ле/-le, -да/-da, -де/-de, -та/-ta, -те/-te) is given by the characteristic of names grouped by internal 5 parts:

(1) body parts (аяқта, жұдырықта, өкшеле);
(2) labor tools (арала, балтала, арқанда);
(3) names of things that are the object of action (жүнде, майла, тұзда, алтында);
(4) address or place, measurement, sound, etc. (мекенде, өрле; аршында, метрле; мыңқылда, шыңқылда);
(5) adjective, numeral, adverb, adverb, etc. cases attached to words (ақта, қарала, екеуле, аһла, уһле).

These 5 different groups are assigned 5 three-digit numbers (sign-code): 113, 114, 115, 116, 117. Similar characteristics and conventional sign-codes in other places of this

group are shown in **Table 4**.

The question may arise about how to practically use such a set of symbol-codes. This was briefly mentioned above in relation to nouns, and it can be said that there is no significant difference in the use of **Table 4** regarding verbs. At this point, it is worth noting that the number of group places (sign-code place) can be assigned (marked), firstly, in accordance with the purpose of the study, and secondly, depending on the lexical-morphological and semantic features of each word class. For example, in **Table 2** for nouns, eight places are enough, in **Table 4** for verbs, the number of such places is only four. The number of grouping places may be different in relation to other word classes or according to the purpose of research. Similarly, from the data in **Tables 2** and **4**, we can see that the number of subsets of each group is also different. Now let's go back to the use of these mentioned tables. First, after the conventional signs that separate the word usages in the part of the studied text into partd of speech, the necessary frequency dictionaries should be obtained from them by a computer. It is necessary to separate the verb forms (or words related to other word classes) in the mentioned frequency dictionary and match them with the sign-codes according to **Table 4**. As a caveat, it is possible to conduct the study directly with text without a frequency dictionary, but repeating text units in this way takes a lot of time. In any case, the research work needs to separate and consider verbs (or other word classes) from the vocabulary system that forms the chain.

Table 4. Program for matching lexical and morphological information to verbs.

Symbol-Code Location	Lexical-Morphological Character of Verbs	Symbol-Code
	Verbs	ET/ET
1	**Semantic and functional grouping:**	
	(1) Action verbs	100
	For example: босат, көтер, күрес, қи, сыз, өлше etc.	
	(2) Movement verbs	101
	For example: ауна, аудар, домала, секір, бүкірей etc.	
	(3) Quality verbs. For example: жат, жантай, тұр, отыр, ұлғай etc.	102
	(4) Speech act verbs. For example: айт, сөйле, де, ескер, жатта, ұмытпа etc.	103
	(5) Growth-development verbs. For example: балала, жапырақта, гүлде, өс, қозыла etc.	104
	(6) Verbs of direction. For example: бар, кел, қайт, әпер, әкет, қаш, түс, көтер etc.	105
	(7) Mood verbs. For example: жыла, қайғыр, өкін, күл, қуан, алақайла etc.	106
	(8) Figurative-imitation verbs. For example: жарқыра, күркіре, дүркіре, тарсылда etc.	107
	(9) Sound verbs	108
	(10) Verbs of seeing and hearing.	109
2	(11) Verbs of behavior	110
	(12) Forms of the verb not included in the list	111
3	**Base verbs**	112
	Definition: The part of the verb base that remains after removing all the suffixes that give rise to various forms (noun, preposition, mood, verb, tense, etc.).	
	For example: аз, ал, айт, алда, ат, ас etc.	
	Derivative verbs formed from nouns.	113
	Ways to create with suffixes:	
	1. Derived verbs formed using the suffix: **–ла (-ле, -да, -де, -та, -те)**:	
	a) attached to the names of body parts. For example: аяқта, жұдырықта, өкшеле.	
	b) Attached to names of labor tools. For example: арала, балтала, арқанда.	114
	c) It is attached to objects that are the subject of action. For example: жүнде, майла, тұзда, алтында.	115
	d) Attached to names of location or place, measurement, sound, movement, appearance, native and social, political, cultural, lifestyle, tradition, nature, etc. For example: мекенде, өрле, төменде; аршында, метрле, тоннала; мыңқылда, шыңқылда; жарқылда, бұрқылда, ирелеңде; ботала, қозда; тәрбиеле, еркеле, жазала, қаумала, үймеле etc.	116
	e) It is added to adjectives, nouns, adverbs, simple words and etc. For example: ақта, қарала, жаманда; онда, жүзде, екеуле; төменде, кейінде, ілгеріле; жымыңда, күлімде; аһла, уһле, etc.	117
	2. Derived verbs formed using the suffix: **–лан (-лен, -дан, -ден, -тан, -тен)** For example: ашулан, арлан, борышлан, иелен …	118
	3. Derived verbs formed using the suffix: **–лас (лес, -дас, -дес, -тас, -тес)**. For example: бірлес, көмектес, бәстес, достас, кезектес, сабақтас etc.	119
	4. Derived verbs formed using the suffix: **–лат (-лет, -дат, -дет)**. For example: дауылдат, тездет, түнделет etc.	120
	5. Derived verbs formed using the suffix: **–а (-е)**. For example: аша, деме, жаса, өрте, тіле, сына, міне etc.	121
	6. Derived verbs formed using the suffix:**–ай (-ей, -й)**. For example: күшей, мұңай, қартай, кеңей, көбей, молай etc.	122

Table 4. *Cont.*

Symbol-Code Location	Lexical-Morphological Character of Verbs	Symbol-Code
4	7. Derived verbs formed using the suffix: **-қар (-ғар, -кер, -гер)**. For example: басқар, ескер, аңғар, теңгер, ақыр, қақыр, жазғыр, ысқыр, кекір etc.	123
	8. Derived verbs formed using the suffix: **-ар (-ер, -р)**. For example: жаңар, тазар, ескір, өзгер etc.	124
	9. formed using the suffix: **-ал (-ел, -ал, -іл, -л)**. For example: жоғал, оңал, тарыл, тіріл, теңел etc.	125
	10. formed using the suffix: **-ық (-ік)**. For example: ашық, бірік, зарық, дәнік, өшік etc.	126
	11. formed using the suffixes: **-сы (-сі)** and **-ымсы (-імсі)**. For example: батырсы, босаңсы, көлгірсі, үлкенсі, апамсы etc.	127
	12. formed using the suffix: **-сын (-сін)**. For example: адамсын, білгішсін, көпсін, жамансын etc.	128
	13. formed using the suffix: **-сыра (-сіре)**. For example: айсыра, әлсіре, қансыра, жетімсіре etc.	129
	14. formed using the suffix: **-ыра (-іре)**. For example: барқыра, бұрқыра, дүркіре, күркіре etc.	130
5	15. formed using the suffix: **-ырай, (-ірей)**. For example: бақырай, шақырай, кішірей, шікірей, etc.	131
	16. formed using the suffixes: **-ы, -і; -шы, -ші; -ан, -ен, -ын, -ін, -н; -ырқа, -ірке, -ырқан, -іркен; -ына, -іне; -қа, -ке, -ға, -ге** etc. For example: байы, жасы, желпі, кеңі; аунақшы, дөңбекші; тасырқа, мүсірке, ашырқан, шіміркен; есіне, қатына, пысына; бүрке, иіске, қозға; басын, жирен, оян, үйрен etc.	132
	Verbs are derivative verbs formed from base words	133
6	Complex verbs	140
	Singular forms of verbs	141

It should always be remembered that only one of the symbol-codes in that collection is selected for one designated place.

As an example, let's take a look at the verb forms in the following sentence: "Кешке әрі тоңып, әрі шаршап аһлап, уһлеп отырған кемпір өзін өлтіріп кете жаздаған кім екенін есіне алды" / "In the evening, the old woman, who was freezing and tired and crying, remembered who almost killed her" (S. Kobeev).

Those belonging to the verb forms: froze, tired, cried, cried, was sat, killed, almost leaved, took.

Now let's match the sign-code values to the verbs according to the data shown in **Table 5**.

Thus, different verbs can take several sign-code values depending on their construction and semantic features. Examine it in more detail in **Table 6**.

Table 5. Verbal forms and their code formations.

Тоңып/ toŋyp [froze]	ET/	106	133	141
Шаршап/ şarşap [tired]	ET/	106	133	141
Аһлап/ ahlap [cried]	ET/	106	133	141
Уһлеп/ uhlep [cried]	ET/	106	133	141
Отырған/ otyrgan [was sat]	ET/	102	133	141
Өлтіріп/öltirip [killed]	ET/	100	133	141
кете жаздаған/ kete jazdagan [almost leaved]	ET/	105	140	140
Алды/aldy [took]	ET/	111	133	141

Table 6. Program for identification of information sign-code to the lexical-morphological structure of the adjective.

Symbol-Code Location	Lexical -Morphological Character of Adjectives	Symbol-Code
1	**If the word (adverb) is a qualitative adjective (main adjective) according to its morpheme structure:**	CH1
2	1) determines the type and color of an object or phenomenon. For example: ақ, қара, қызыл, жасыл etc.	201
	2) if it determines the secret and quality of a thing or phenomenon. For example: жақсы, жаман, тәуір, нашар etc.	202
	3) determine the volume and area, length and weight of the object. For example: үлкен, кіші, ұзын, ауыр, қысқа, жеңіл etc.	203
	4) if it means the taste and smell of a thing. For example: ащы, тәтті, күлімсі etc.	204
	5) if it means other characteristics of a thing or phenomenon.	205
3	**If the following word (vocabulary) belongs to a relative adjective according to its morpheme composition, that is, it expresses the sign of an object through the relation of another object, and such derivative adjectives are formed using productive (unproductive) suffixes.**	CH2

Table 6. *Cont.*

Symbol-Code Location	Lexical -Morphological Character of Adjectives	Symbol-Code
4	**Derivatives from nouns are created using productive suffixes that form adjectives and pertain to the object's appearance and color, shape and form, secrecy and criticism, place and time, and others if they express a critical concept about the attributes.**	
	1. Derived adjectives formed from some nouns, pronouns, adverbs, as well as from words in the genitive and genitive case with the suffixes **-қы, -кі; -ғы, -гі** are used in the following sense:	
	а) if it is attached to some nouns, adverbs, as well as to nouns in genitive (sometimes eastern) adverbial forms. For example: ауызғы, төргі, түпкі, ішкі, төменгі, соңғы, санаулы, бүктеулі etc.	210
5	ә) if the temporal concept is attached to some nouns and pronouns, temporal adverbs. For example: кешкі, түскі, күзгі, жазғы, көктемгі, түнгі, күндізгі etc.	211
	2. The following derivative words are formed by adding **-лы, -лі, -ды, -ді, -ты, -ті**:	
	а) if it is added to nouns to express the existence of a certain thing (or phenomenon). For example: арлы, сулы, әсерлі, гүлді, икемді, байыпты, инабатты, балалы etc.	212
	b) compound adjectives from compound nouns, adjectives, numerals, and adverbs are formed using these additions. For example: ағалы-інілі, ойлы-қырлы, таулы-тасты, өзенді-сулы, үлкенді-кішілі, бұрынды-соңды etc.	213
	c) complex adjectives consisting of two or three words and used as descriptive names. For example: ақ басты, қаз мойынды, ай қабақты, теке сақалды etc.	214
	3. If the suffix **-сыз (-сіз)** is added to the nouns, and the derived adjectives with negative meaning are formed. For example: баласыз, көліксіз, білімсіз, сенсіз, бізсіз, мұңсыз etc.	215
	4. **-шыл (-шіл)** suffix is added to nouns, pronouns, modal words (each side) to form relative adjectives expressing inclination, flexibility, passion. For example: ұйқышыл, ұйымшыл, өзімшіл, турашыл, ойшыл etc.	216
	5. **-дай, (-дей, -тай, -тей)** жұрнағы жалғанып, сөздерді салыстыру, ұқсату мәнді туынды сын есімдер жасалса. If the suffix **-дай, (-дей, -тай, -тей)** is added, it form the adjectives with the meaning of comparison and similarity. For example: аттай, әкедей, мендей, сендей, жүздей, өлердей etc.	217
	6. Derived adjectives with the following meaning are formed using the suffixes **-лық, -лік, -дық, -дік, -тық, -тік**:	
	а) if they are attached to nouns and form derived adjectives that express the concreteness and meaning of the words. For example: орталық, қоғамдық, қалалық, азаматтық, жолдастық etc.	218
6	ә) if they are connected to the names of seasons and various product names, and derivative adjectives related to the season and measurement amount are made. For example: айлық, жылдық, апталық, тәуліктік, көйлектік, пальтолық, қайнатым-дық, екі-үш асымдық etc.	219
	б) if they are connected to pronouns and form derived nouns from them that indicate belonging to a certain party. For example: өздік, мендік, сендік, қандайлық, қаншалық etc.	220
	7. Derived adjectives with the following meanings are formed using the suffixes **-лас, -лес, -дас, -дес, -тас, -тес**:	221
	а) if it is attached to a noun related to a person's character, it means that such a character also applies to other people. For example: ниеттес, сырлас, тілеулес, пікірлес etc.	
	ә) If human beings and animals are connected with nouns related to the genealogy, and an adjective is formed, indicating that they are relatives. For example: туыстас, аталас, бауырлас, қарындас, тумалас etc.	222
	б) if they are connected to the concepts of place and made derivative adjectives with the meaning of settlement with them. For example: ауылдас, көршілес, жерлес, елдес, тумалас, ұялас, жапсарлас, іргелес etc.	223
	в) attached to some nouns, depending on their specific lexical meanings, they express critical concepts that are related to different relations, terms and quantities. For example: жолдас, замандас, сабақтас, кәсіптес, қызметтес, дәмдес, өкшелес etc.	224
	8. Derived adjectives are formed using the suffixes **-шаң, -шең**:	
	а) if adjectives are added to the names of clothes and represent the appearance of a person related to his clothes. For example: көйлекшең, етікшең, пальтошаң, шалбаршаң etc.	225
	ә) if attached to some nouns, derived adjectives are formed that express a certain special characteristic of a person or thing. For example: ашушаң, сөзшең, бойшаң, тершең, кіршең etc.	226
7	**9. By unproductive suffixes that form adjectives derived from nouns:**	
	а) If the suffixes **-дар, -дер, -тар, -тер** are attached to some nouns, a critical concept is created, which means that one has been subjected to a certain action. For example: қарыздар, хабардар, борыштар etc.	227
	ә) If the suffix **«-и»** transferred from Iranian languages (Persian, Tajik, Afghan) is added to some nouns and a derived adjective is formed. For example: әскери, мәдени, тарихи etc.	228
	б) If the suffix **-паз** is added to create a new word that tends to an adjective. For example: өнерпаз, аспаз, білімпаз, ойынпаз etc.	229
	в) If the suffix **-мпаз (-м+паз), -ымпаз (-ым+паз), -імпаз (-ім+паз)** (and components) are added to some nouns and form a derived adjective. For example: сезімпаз, алымпаз, білімпаз, жағымпаз, жасампаз etc.	230
	г) If the suffix **-қой, (-ғой, -гөй)** is added to some nouns and participants, it creates a derived adjective that is the name of actions and behaviors. For example: кәсіпқой, әзілқой, сәнқой, жәдігөй, әзілқой, сәнқой, жанжалқой etc.	231
	д) If the suffix **-қор** is added to nouns, it form a derivative adjective meaning a specific practice. For example: жемқор, жалақор, айлақор, бейнетқор, қамқор, мансапқор, ызақор, намысқор etc.	232
	Adjectives derived from verbs using unproductive suffixes:	317
	1. **-қ, -к, -ық, -ік, -ақ, -ек**: а) It is connected to a ritual and lesson verb and creates nouns with different critical concepts. For example: ашық, тұнық, шірік, ілік, дөңгелек etc.;	
	б) if it is connected to imitation words and creates a relative adjective. For example: бұлтақ, жалтақ, еңкек, жалпақ etc.;	
	в) if it is attached to some nouns and forms a derived adjective. For example: жолақ, қасырақ, ортақ, ирек etc.;	

Table 6. *Cont.*

Symbol-Code Location	Lexical -Morphological Character of Adjectives	Symbol-Code
	г) if from the verb **-уық (-уік)** suffixes form a derived adjective expressing a tendency to a certain action. For example: жылауық, сөйлеуік, сұрауық, сыбырлауық etc.	
8	2. **-ыңқы, -ңқы, -іңкі.** For example: жатыңқы, салбыраңқы, көтеріңкі, кебіңкі, батырыңқы etc.	318
	3. **-ынды, -інді, -нды, -нді**. For example: асыранды, серпінді, жаттанды, шұбырынды, түйінді etc.	319
	4. **–малы (-мелі, -балы, -белі, -палы, -пелі)**. For example: ауыспалы, көшпелі, таңдамалы, аумалы-төкпелі etc.	320
	5. **-қыш, -кіш, -ғыш, -гіш.** For example: білгіш, оңғыш, тапқыш, сенгіш, айтқыш etc.	321
	6. **-шақ (-шек)**. For example: мақтаншақ, еріншек, қызғаншақ, ашуланшақ, жасқаншақ, тартыншақ etc.	322
	7. **-ымды (-імді, -мды, -мді)**. For example: жағымды, ұғымды, үйлесімді, жарасымды, сенімді etc.	323
	8. **-улы, -улі**. For example: жинаулы, ерттеулі, үюлі etc.	324
	9. **–қақ, -кек, -ғақ, -гек**. For example: асқақ, тоңғақ, жабысқақ, майысқақ, тайғақ, оңғақ, қатқақ, ұрысқақ etc.	325
	10. **ма, -ме, -ба, -бе, -па, -пе**. For example: жалдама (ақы), қызба (адам), көшпе (құм), сырма (бешпет), аспа (шам), сырма (бешпет) etc.	326
	Adjectives derived from verbs using unproductive suffixes:	327
	1. **–ыс, -іс, -с**. For example: келіс, кетіс, ұқсас, таныс, тіркес, жалғас, талас etc.	
	2. **–қыр, -ғыр, -кір, -гір**. For example: білгір, ұшқыр, үшкір, алғыр, өткір, білгір etc.	328
	3. **–мыс (-мыш)**. For example: жасамыс, айтылмыш, тарамыс, жазылмыш, жаралмыш etc.	329
	4. **–ымтал (-імтал)**. For example: ұғымтал, өсімтал, сезімтал etc.	330
	5. **–қы (-кі, -ғы, -гі)**. For example: бұралқы, оралғы, жинақы, күлдіргі etc.	340
	6. **–ыр (-ір, -ар, -ер, -р)**. For example: жұмыр, иір, обыр, қыңыр, тықыр, шымыр, құзар, былжыр, жалтыр, жылтыр, сылбыр, былбыр etc.	341
	7. **–у.** For example: жарау, қызу, тақау, жадау, жабырқау, түзу, таяу, бітеу, баяу, көлбеу, қату etc.	342
	8. **–аған (-еген)**. For example: тебеген, қабаған, сүзеген, қашаған, береген, алаған, безеген, көреген etc.	343
	9. **–ын (-ін, -н)**. For example: ортан, бүтін, еркін, ұзын, жайын etc.	344
	10. **–қалақ (-келек, -ғалақ, -гелек)**. For example: ұшқалақ, сасқалақ, қозғалақ etc.	345
9	11. **–алақ (-елек)**. For example: еңкелек, қаңғалақ, бұлталақ, шыжалақ, жалталақ etc.	346
	12. **–анақ**. For example: сұғанақ, шұқанақ, шығанақ etc.	347
	13. **–ғылықты (-гілікті, -қылықты, -кілікті)**. For example: жеткілікті, тұрғылықты, жергілікті etc.	348
	14. **–мсақ (-мсек, -ымсақ, -імсек)**. For example: сұрамсақ, тілемсек, жарамсақ, берімсек, өлімсек etc.	349

Now, based on the information sign-code matching program for the lexical-morphological structure of the adjective, as considered in **Table 3**, examples of conditional sign-code matching for some basic and derived adjectives are provided in **Tables 7** and **8**.

It is clear that adjectives are divided into simple and complex, depending on the composition of words. Simple adjectives consist of only one component, but they do not necessarily have an additional morpheme [17].

Table 7. Codes of the main adjectives.

Qualitative (Base) Adjectives	Symbol-Code	Location
	1	**2**
Red, green, etc.	CH1	201
Good, bad, etc.	CH1	202
Short, long, heavy, etc.	CH1	203
Sweet, salty, etc.	CH1	204

Table 8. Codes of derived adjectives.

Relative (Derivative) Adjective	Symbol-Code	Location
	1	**2**
Сыртқы/external, төргі/on the base place, etc.	CH2	210
Кешкі/at night, күндізгі/at the daetime, etc.	CH2	211
Әсерлі/impressive, инабатты/kind, etc.	CH2	212
үлкенді-кішілі/with big and small, өзенді-сулы/ with river-watered, etc.	CH2	213
қаз мойынды/with long neck, теке сақалды/ with long beard, etc.	CH2	214
Мұңсыз/not sad, білімсіз/uneducated, etc.	CH2	215
Ойшыл/ thoughtful, өзімшіл/ selfish, etc.	CH2	216

Table 8. *Cont.*

Relative (Derivative) Adjective	Symbol-Code	Location
	1	**2**
Балғадай/like a hammer, әкедей/like a father, шөлмектей/ like a thirst, etc.	CH2	217
Қалалық/urban, қоғамдық/public, etc.	CH2	218
Айлық/monthly, жылдық/yearly, etc.	CH2	219
Көйлектік/from shirt, пальтолық/from coat, etc.	CH2	219
Өздік/ours, қаншалық/how much, etc.	CH2	220
Сырлас/sharing person, ниеттес/intended person, etc.	CH2	221
Туыстас/relative, қарындас/young sister, etc.	CH2	222
Ауылдас/from same village, көршілес/neighbour, etc.	CH2	223
Жолдас/friend, қызметтес/colleague etc.	CH2	224
Көйлекшең/small shirt, пальтошаң/small coat etc.	CH2	225
Ашық/open, шірік/rotten, дөңгелек/round etc.	CH2	317
Көтеріңкі/ upper, жатыңқы/laying etc.	CH2	318
Жаттанды/memorized, серпінді/ dynamic etc.	CH2	319
Таңдамалы/chosen, көшпелі/moved etc.	CH2	320
Білгіш/ connoisseur, оңғыш/ right, сенгіш/bealiver etc.	CH2	321
Мақтаншақ/proud, еріншек/arrogant etc.	CH2	322
Жағымды/pleasant, үйлесімді/harmonious etc.	CH2	323
Жинаулы/collected, ерттеулі/early etc.	CH2	324
Ұрысқақ/quarrelsome, тоңғақ/stubborn etc.	CH2	325
қызба (адам)/ hotspur, сырма/slide etc.	CH2	326
Таныс/familiar, келіс/agree etc.	CH2	327
Сұрамсақ/onion, өлімсек/listless etc.	CH2	349

On the other hand, complex adjectives include words composed of two, three, or more components and are created syntactically (analytically). It is known that complex adjectives are formed from basic adjective words through their combination, doubling, and union. They serve as a complex persona in our language.

We decided that words created according to the syntactic method of the adjective found in the text should be marked with conventional signs or matched with the appropriate sign-code using the above table (**Table 9**).

Table 9. A program for matching the information sign-code to the syntactic structure of the adjective.

Symbol-Code Location	Lexical -Morphological Character of Complex Adjectives	Symbol-Code
	Complex Adjectives	CH3
1	1. Qualitative (main) adjectives combine with each other. For example: black spotted, yellow spotted, red spotted, etc.; dark blue, dark red, dark maroon, dark red, etc.	400
	2. Single or basic (qualitative) or derived (relative) adjective words are either repeated or doubled. For example: кішкене-кішкене/small-small, аппақ-аппақ/ white-white, үлкен-үлкен/ big-big, etc.; малды-малсыз/with and without livestock, үлкенді-кішілі/large and small, қоралы-қопсылы/ stable and free etc.	401
	3. The main adjective is combined with the suffixes of derived adjective **-лы/-ly, -лі/-li**. For example: кең маңдайлы/ intelligence, қызыл шырайлы/ with a ruddy and lovely face etc.; both components of which are formed by the suffixes **-лы/-ly, -лі/-li** or with the suffixes to creat the comple adjectives **-сыз/-syz, -сіз/-siz**. For example: таулы-тасты/ mountainous-rocky, әкелі-балалы/ father-child, ессіз-түссіз/ crazy-colorless etc.	402
	4. Compound adjectives are formed by combining nouns or numerals with the suffixes **-лық (-лік, -дық, -тік, -тық, -тік)**. For example: халық аралық/ international, бес жылдық/ five-year, екі кісілік/ two-person, екі-үш күндік/ two-three-day etc.	403
	5. Compound adjectives formed by adding suffixes **-лық, -лік, -дық, -дік, -тық, -тік, -и** into both components or adding suffix **-н** into one component. For example: әскери-саяси/military-political, ғылыми-әдеби/scientific-literary, қоғамдық-тарихи/social-historical, әлеуметтік-экономикалық/socio-economic, статистикалық-лингвистикалық/statistical-linguistic, etc.	404
	6. Addition of the 3rd side of the dependency is used in compound adjective calculation by adding an adjective to the added noun.For example: көзі ашық/open mind, тілі майда/quite, not talkative, басы бос/single, қолы қысқа/doesn't have power, жүзі жылы/lovely etc.	405

In order to clarify the data in this table, we present **Table 10** below as an example of correspondences. Based on this conventional designation of the types of adjectives in the text or in the frequency dictionary created by the syntactic method, it is possible to use the probabilistic-statistical method during their research.

Table 10. Conditional designation of adjective types in the text.

Complex Agjectives	Symbol-Code	Location
	1	**2**
black spot, yellow spot; dark blue, dark red, dark maroon, etc.	CH3	400
кішкене-кішкене/small-small, аппақ-аппақ/ white-white, үлкен-үлкен/ big-big, etc.;	CH3	401
малды-малсыз/with and without livestock, үлкенді-кішілі/large and small, қоралы-қопсылы/ stable and free etc.		
кең маңдайлы/ intelligence, қызыл шырайлы/ with a ruddy and lovely face etc.;	CH3	402
таулы-тасты/ mountainous-rocky, әкелі-балалы/ father-child, есссіз-түссіз/ crazy-colorless etc.		
халық аралық/ international, бес жылдық/ five-year, екі кісілік/ two-person, екі-үш күндік/ two-three-day etc.;	CH3	403
көңіл көтерерлік/ entertainment, қабырға қайысарлық/ wall resistance, etc.		
әскери-саяси/military-political, ғылыми-әдеби/scientific-literary, қоғамдық-тарихи/social-historical,	CH3	404
әлеуметтік-экономикалық/socio-economic, статистикалық-лингвистикалық/statistical-linguistic, etc.		
көзі ашық/open mind, тілі майда/quite, not talkative, басы бос/single, қолы қысқа/doesn't have power, жүзі жылы/lovely etc.	CH3	405

4. Discussion

The data obtained by linguistic-statistical research from frequency dictionaries (alphabetic, frequency, reverse-alphabetic, dictionary) created from the texts of Kazakh language fiction and other styles with the help of a computer are very useful for automatic processing of Kazakh texts, long-term storage and searching and finding the necessary information [18].

According to the findings of scientists studying other languages, some statistical features are common to many languages, while others differ depending on the originality of the language, the differences between genres and the language-specific features of the author. Texts in Kazakh also have their own characteristics. For example, the frequency of words in the frequency dictionaries compiled from the texts of different styles is not uniform in terms of text or dictionary coverage [19].

Statistical research on texts in Turkic languages holds great theoretical importance. The results of studies on the agglutinative structure in Turkic languages enable us to compare them with the rich statistical data of Indo-European languages with an inflectional-analytical structure. For example, let's consider the following: In Russian, English, German, French, Romanian, Spanish, etc., 100–150 high-frequency words are sufficient to cover 50 percent of any text. In contrast, 700–800 high-frequency words are required in Turkic languages [20].

The amount of information description in Turkic languages may differ from that in Indo-European languages. Therefore, it should be understood that linguistic statistical research also falls under the typological experiment [21].

A probabilistic-statistical and informative model based on Kazakh texts was created in the work, and their capabilities were determined. According to the results of such research, when frequency is considered from the beginning to the end of the dictionary, the characteristic features of the statistical and informational coverage of the words throughout the text were determined by appropriate theoretical criteria. It has been established that, in most cases, the words in the frequency dictionaries based on the texts of the Kazakh language are independent of the text [22].

In short, some linguistic phenomena are very weakly manifested, remain hidden, and in most cases, are beyond the direct control of the researcher. Here, a mathematical tool for determining such linguistic phenomena can be found through statistical research. Such phenomena include the use of words, phrases, etc., related to some parts of speech, and the features of the use of units in the language.

The mathematical tools mentioned in our research are:

(a) statistical laws of distribution of the variable (frequency of linguistic units) within the text;
(b) theoretical criteria for evaluating the ratios of statistical and probabilistic properties of language units.

We are confidently believe that the mentioned educational tool will be helpful in studying the structure of texts in Turkic languages, including Kazakh, objectively using

quantitative methods.

The results of this study have a wide range of applications in the fields of linguistics, computational linguistics, and education. Below are some of the possibilities for improving existing tools and developing new approaches using the data[23].

1. NLP (Natural Language Processing) in Kazakh
   The study provides an important basis for developing technologies for automatic processing of Kazakh texts. The system of encoding words and their morphological characteristics can be integrated into the following tasks:
   - Identification of word classes: creating more accurate models for automatic detection of word classes in text. This is especially important for the Kazakh language, which has a rich morphological structure.
   - Morphological analysis: Developing algorithms for analyzing and generating word forms taking into account the morphemic composition, categories, and meanings of words.
   - Syntactic analysis: improving systems that build dependency trees or perform other types of syntactic analysis.
   - Machine translation: Improving the quality of machine translation from Kazakh to other languages and vice versa by integrating data on word frequency and morphological structure.
2. Language learning
   The results of the research will be used to create modern teaching materials to improve the teaching of Kazakh as an analytical and foreign language. Examples:
   - Textbooks and dictionaries: including statistical data on word frequency and morphological forms in textbooks to better understand the structure of the language.
   - Online courses: developing interactive courses for learning the language using practical examples based on real texts.
   - Adaptive learning: creating adaptive learning programs that focus on the most frequently used words and constructions using frequency data.
3. Lexicographic database
   Creating a new generation of dictionaries and reference books that take into account not only the lexical meaning of words, but also their morphological structure, frequency, and contextual features:
   - Electronic dictionaries: adding information on morphological types and frequency of use of words to improve search functions.
   - Specialized dictionaries: developing dictionaries for scientific, technical or literary texts focused on specific categories of words.
   - Corpus linguistics: creating an annotated corpus of texts using the proposed coding system, allowing for deeper study of the language.
4. Comparative study of Turkic languages
   The results of the study can serve as a basis for a comparative analysis of the Kazakh language with other Turkic languages. This is especially important:
   - Identifying similarities and differences: comparing the morphological and lexical features of the Kazakh language with other Turkic languages.
   - Developing universal models: Creating linguistic models applicable to all Turkic languages based on the identified patterns.
5. Automate text analysis
   Using the proposed coding system opens up possibilities for the automatic analysis of large texts:
   - Text analytics: identifying key themes, trends and patterns based on word frequency and morphological characteristics.
   - Stylistic analysis: identification of stylistic features of texts (e.g. fiction, scientific, professional) taking into account the distribution of word classes and their forms.

5. Conclusion

Thus, the program of conventional marking of nouns, verbs and adjectives from the main vocabulary of the Kazakh language, i.e., the identification of signs and codes for their types, creates an opportunity to apply the statistical method in the morphological structure of the Kazakh language. We are confident that such a program can be implemented in the described way for other word classes of the Kazakh language, taking into account their specific features. The model of the program we propose is a result of theoretical research, so it has not yet been fully implemented in practice. But such statistical studies will be carried out for the Kazakh language

in the near future[24].

The article presents the results of statistical analysis of the lexical and morphological composition of the Kazakh language. The proposed coding system allows for effective classification of words and their forms, which opens up new possibilities for automatic text processing and linguistic research[25].

The main achievements of the study include the following:

1. Develop a universal coding system that can adapt to different classes of words.
2. Obtain statistical data on the frequency and morphological forms of words.
3. Create a basis for natural science programming in the Kazakh language, linguistic teaching and comparative research of Turkic languages.

In the future, the analysis will be expanded to other classes of words (e.g. pronouns, adverbs) and comparative studies with other Turkic languages. This will allow for a deeper understanding of the structure of the Kazakh language and the creation of universal models applicable to a wide range of languages.

Thus, this study is an important step in the development of modern linguistics and computational linguistics and opens up new prospects for the study and use of the Kazakh language in the digital age.

Author Contributions

Conceptualization, T.D. and P.K.; methodology, T.D.; software, Z.G.; validation, T.D., P.K., and Z.G.; formal analysis, M.G.; investigation, T.D. and S.A.; resources, S.A. and S.D.; data curation, T.D.; writing—original draft preparation, P.K. and T.D.; writing—review and editing, S.D.; visualization, M.G.; supervision, T.D.; project administration, M.G.; funding acquisition, T.D. All authors have read and agreed to the published version of the manuscript.

Funding

This research has been 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»).

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Due to confidentiality considerations, the research data cannot be disclosed.

Acknowledgments

This research has been 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»).

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Johnson, L., 2017. Abai's Impact on Kazakh Literature: A Critical Review. Journal of Comparative Literature. 42(4), 567–580.
[2] Williams, R., 2019. Symbolism in Abai's Poetry: A Cross-Cultural Perspective. Journal of World Literature. 36(1), 89–104.
[3] Jones, E., 2018. Abai's Cultural Legacy: Influence and Reception. Cultural Studies Review. 25(2), 211–228.
[4] Zhumagulova, V., Kazhigaliyeva, G., Onalbayeva, A., et al., 2019. Representation of modernization of public consciousness in the mass media discourse. Opcion. 35(23), 1171–1187.
[5] Abisheva, S.D., Khavaidarova, M.M., Sabirova, D.A., et al., 2024. Experience of Implementing Trilanguage in Kazakhstan on the Example of Pedagogical Universities. Journal of Ecohumanism. 3(7), 283–290.
[6] Pirmanova, K.K., Serikbayeva, A.D., Nurlybayev, N.M., et al., 2024. Potential of Using Abai's Poetic Corpora for Research, Educational and Methodical Purposes. Journal of Ecohumanism. 3(4), 618–626.

[7] Sabirova, D.A., Abisheva, S.D., Polyak, Z.N., et al., 2020. Transformation of Family Norms in Modern Literature of Kazakhstan. Journal of Research in Applied Linguistics. 11, 20–29.

[8] Madiyeva, G.B., Pirmanova, K.K., Kunsulu, K., et al., 2025. Problems of Using the Poetic Corpus of Abay for Pedagogical Purposes and Ways to Solve it. Journal of Ecohumanism. 4(1), 1477–1482.

[9] Pirmanova, K.K., Zaurbekova, G., Tokmyrzaev, D.O., 2023. Derivatives and word formation methods (structural aspect of word formation designations. Polish Journal of Science. (67), 28–34.

[10] Taylor, K., 2016. Abai's Poetry and the Romantic Tradition. Romanticism. 33(4), 477–492.

[11] White, A., 2014. Abai's Humanism and Social Consciousness. Humanities Today. 12(3), 301–318.

[12] Garcia, M., 2011. Abai's Poetic Vision: A Comparative Study. Comparative Literature Review. 20(1), 55–70.

[13] Brown, L., Abai's Poetry: A Linguistic and Cultural Perspective. Journal of Linguistic Studies. 22(2), 211–228.

[14] Pirmanova, K.K., Torekhanova, A., Ashimova, N., 2023. Theory and methodology of the world's national linguistic corpora. «Bulletin. Series: Philological sciences» of the Joint Stock Company «KazUIR&WL named after Ablai Khan». 2 (69), 158–181.

[15] Akar, A., Pirmanova, K.K., Shaimerdenova, N.Z., et al., 2023. Methods of transferring full word-formation affixes in the national corpus of the Kazakh language. Tiltanym, (3), 3–14.

[16] Klimenko, S.V., Rykov, V.V., Logical induction and deduction as principles of reflection of the subject area in the corpus of texts [Electronic resource]. Available from: www.narod.yandex.ru (cited 24 December 2024).

[17] Pankova, I.M., 2006. Comparison as a meaning-forming dominant of an advertising text: abstract of cand. diss. Stavropol, 23p. [Electronic resource]. Available from: https://surl.li/mddtvp (cited 10 December 2024).

[18] Baimyrza, A., Serikbayeva, A., Pirmanova, K., et al., 2024. The Impact of Linguistic Identity on the Perception of Abai's Poetry. Forum for Linguistic Studies. 6(5), 239–246. DOI: https://doi.org/10.30564/fls.v6i5.6879

[19] Suleimenova, E.D., Shaimerdenova, N.Z., Akanova, D.X., 2007. Dictionary of Sociolinguistic Terms. Kazakh University: Almaty, Kazakhstan.

[20] Shabalova, I.O., 2016. Features of the Perception of Poetry by Russian and French Students: Comparative Analysis. Quarterly peer-reviewed, Scientific Journal Bulletin of ASU. 3(184), 95–101.

[21] Francis, W.N., 1983. Problems of the Formation and Machine Representation of a Large Text Corpus // New in Foreign Linguistics. Problems and Methods of Lexicography. 14, 334–352.

[22] Sharov, S.A., 2003. Representative Corpus of the Russian Language in the Context of Global Experience // Scientific and Technical Information. Series 2: Information Processes and Systems. 6, 9–18.

[23] Fillmore, C.J., 1992. 'Corpus linguistics' or 'Computer-aided armchair linguistics' // Directions in Corpus Linguistics. Berlin: de Gruyter, Available from: https://is.muni.cz/el/phil/jaro2008/FJ0B738/um/Corpus_linguistics_verze1.pdf (cited 10 November 2024).

[24] Madiyeva, G.B., 2016. Computational Linguistics. Textbook. Kazakh University: Almaty, Kazakhstan. 240p.

[25] Suleimenova, E.D., 2015. Is a National Corpus of the Kazakh Language Needed? // "TILTANYM" Journal of A. Baitursynov Institute of Linguistics. 2, 42–45. Available from: https://www.tiltanym.kz/jour/article/view/441?locale=kk_KZ (cited 1 November 2024).