BILINGUAL PUBLISHING GROUP

Forum for Linguistic Studies

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

ARTICLE

A Corpus Approach in Language Discovery: A Word Frequency Analysis Based on the Corpus Outcomes in Kazakh

Sofiya Omarova [1], Dana Ospanova [1], Nurlykhan Aitova [2], Gulnaz Tokenkyzy [1], Assel Ormanova [3*], Madina Alshynbekova [4]

[1] National Scientific and Practical Center «Til-Qazyna», Astana 010000, Kazakhstan

[2] Department of Kazakh Linguistics, Eurasian National University, Astana 010000, Kazakhstan

[3] Department of General Education Disciplines, Astana IT University, Astana 010000, Kazakhstan

[4] Branch campus of Beijing Language and Culture University (BLCU), Astana International University, Astana 010000, Kazakhstan

ABSTRACT

This study examines the most frequently used parts of speech and grammatical forms in the texts of the Sub-corpora of the National Corpus of the Kazakh Language (qazcorpora.kz). The frequency of word forms based on the 13-million-word usages in the 2023 corpus database was collected and analyzed both manually and using the functional setting of the corpus software. The study provided key insights into Kazakh journalistic texts' frequency distribution, grammatical variability, and comparative patterns. The results indicated that: (1) conjunction 'žäne' [and], demonstrative pronoun 'bul' [this], auxiliary verb 'dep' [no translation], noun 'Kazakh' [Kazakh], modal verb 'žoq' [not], adjective 'aq' [white], adverb 'köp' [many/much], numeral 'eki' [two] showed the highest frequency indicators emphasizing their functional and stylistic roles in text construction in their word class. (2) functional words were the most frequently used part of speech. (3) conjunction 'žäne' [and], postposition 'üšın' [for] and particle 'ɣana' [only] possessed the highest frequency indicators among functional words. This corpus-based research highlights the alignment of Kazakh frequency patterns with global linguistic trends, such as Zipf's law, while also showcasing unique features attributed to the language's

*CORRESPONDING AUTHOR:

Assel Ormanova, Department of General Education Disciplines, Astana IT University, Astana 010000, Kazakhstan;
Email: assel.ormanova@yandex.kz

ARTICLE INFO

Received: 5 November 2024 | Revised: 20 January 2025 | Accepted: 23 January 2025 | Published Online: 20 February 2025
DOI: https://doi.org/10.30564/fls.v7i2.8317

CITATION

Omarova, S., Ospanova, D., Aitova, N., et al., 2025. A Corpus Approach in Language Discovery: A Word Frequency Analysis Based on the Corpus Outcomes in Kazakh. Forum for Linguistic Studies. 7(2): 869–881. DOI: https://doi.org/10.30564/fls.v7i2.8317

agglutinative nature.

Keywords: National Corpus; Frequency Indicator; Part of Speech; Grammatical Form; Corpus Linguistics

1. Introduction

The potential and capabilities of language corpora provide significant benefits not only for linguists but also for society at large, playing a vital role in shaping the future of language studies[1]. This is crucial for studying the richness of a language, its diachronic and synchronic development, various changes in its stylistic usage, its lexical-grammatical aspects, as well as the dynamics of the language over a certain period, and for creating linguistic models[2]. Additionally, language corpora serve as valuable resources for developing linguistic models, designing curricula and standards, and creating textbooks, educational tools, reference materials, and guidelines[3].

This research focuses on analyzing frequency patterns in Kazakh journalistic texts using the Sub-corpora of the National Corpus of the Kazakh Language (https://qazcorpora.kz/). It investigates the usage characteristics of linguistic elements such as parts of speech, functional words, pronouns, and their grammatical forms in Kazakh. Furthermore, it compares these findings with word frequency patterns in other corpora.

The study addresses two primary research questions:

(1) What are the most frequently used parts of speech and grammatical forms in the Kazakh language according to the Sub-corpora of the National Corpus of the Kazakh Language?
(2) How do these frequencies reflect the functional and stylistic features of the Kazakh language?

2. Literature Review

2.1. Word Frequency

The research in this field is related to issues of corpus linguistics, the comparison of large corpora using various statistical, linguistic, and mathematical methods, and the study of word frequency in texts of different genres.

Kilgarriff[4] studies word frequency by identifying frequently used elements in the German language based on various scientific data. He explored the similarity and homogeneity of corpora through measurement methods, examined the connections between grammatical and lexical similarities, and conducted statistical experiments on word frequency.

Nilsson[5] examines word frequency in popular science and research articles. He conducted a corpus linguistic analysis to carry out both quantitative and qualitative research on these two genres, each containing over 200,000 words. The analysis included word frequency, comparison, cluster analysis, and phraseology.

A study by Shin and Paul[6] finds that demonstrative pronouns are used significantly more frequently in popular science articles compared to research articles, while research articles tend to have a higher occurrence of nouns and specific terminology.

Baayen[7] compares the impact of word frequency across different languages using parallel statistical methods. They discovered that languages with varying morphological systems exhibit different word frequency characteristics. For instance, even the most frequently used complex word forms in polysynthetic languages tend to have lower frequency compared to the word frequency in more isolating languages. The researchers created a frequency rank graph based on absolute frequency to assess word frequency. They concluded that word frequencies in the ten languages studied align with the typical Zipf's law, which states that the most frequent word occurs twice as often as the second most frequent word, three times as often as the third, and so on.

Zasorina[8] demonstrates that the size of the corpus allows for the study of variability and frequently occurring linguistic phenomena, opening effective results in several areas: (1) The study of morphological variations and the development of parts of speech, such as nouns and verbs. (2) The examination of word formation variants and related issues, such as paronyms, word formation models, and the productivity of word formation tools. (3) The analysis of variations in agreement, concord, and other constructions. (4) The study of accentological variations and changes in the Russian language's stress system. (5) The investigation of lexical variability, including the semantic relations and composition changes in synonymic series and thematic groups.

Zotina and Solovyov[9] emphasize that word frequency is an important marker of semantic and formal changes in words, reflecting shifts in the lexical system of a language. They examined diachronic changes in the frequency of nouns representing concrete and abstract concepts and those considered significant and relevant in different historical periods of the Russian language.

Kim[10] states that if a word is high-frequent in one language, there is a high likelihood that it shows high frequency in another. Exceptions are polysemous words. Since polysemy can be inherited from a proto-language, it is impossible to compile a representative universal list of meanings whose frequency would not depend on polysemy.

There are retrospective studies on word frequency in several Turkic languages over specific periods. Ölker systemized word frequency from 1945 to 1950 ("Yazılı Türkçenin Kelime Sıklığı Sözlüğü"), Çal compiled word frequency data from 1975–1980, and Göz studied word frequency from 1995–2000, resulting in the dictionary "Yazılı Türkçenin Kelime Sıklığı Sözlüğü". Çal provides an overview of the history of word frequency research. According to his data, word frequency research was first conducted manually in English in 1944 by E.L. Thorndike and I. Lorge, resulting in "The Teacher's Word Book of 30,000 Words". Subsequently, in 1967, H. Kucera and W.N. Francis analyzed word frequency using a computer later presented in "Computational Analysis of Present-Day American English"[11].

2.2. The Frequency of Word Classes

Generally, the frequency of word classes in the text is identified using a simple calculation method and presented as follows: (1) Nouns. (2) Verbs. (3) Pronouns. (4) Prepositions. (5) Adjectives. (6) Adverbs. (7) Conjunctions. (8) Numerals and Particles[8]. In some works, the frequency of word classes is examined in the texts of the belles-lettres style. For example, the frequency of word classes in the 100-word Bunin's story "Kostsy" was identified. The text contains 22 nouns, 15 adjectives, 15 pronouns, and 14 verbs, while the conjunction 'и' (and) appears 6 times[12]. Since the text is mostly descriptive, the frequency of nominal words (nouns and verbs) constitutes higher indicators.

Kolár and Plecháč[13] researched the frequency of word classes in the Czech language. They stated that poetry shows a higher percentage of the use of nominal parts of speech (nouns, adjectives), whereas in artistic prose, correspondingly, verbal parts of speech (verbs, adverbs) are more prevalent; this is likely related to the division into epic and lyrical genres of literature.

In the current study, the frequency tends to be identified by using corpus linguistics tools.

2.3. Current Linguistic Corpora

The linguistic corpus is a unified, structured, philologically competent collection of linguistic data presented in an electronic format. With the development of computer technologies and artificial intelligence, corpus-based research practices have been significantly well-developed in linguistics[14].

World languages have been represented in linguistic corpora. The British National Corpus (http://corpus.byu.edu/bnc), which includes 100 million words in English, covers a wide range of sources for both written and spoken language samples. A distinctive feature of this corpus is the separate presentation of word frequency for written and spoken language. The largest structured corpus of historical English, the Corpus of Historical American English (COHA), represents over 475 million sources. The German language corpus of the Berlin-Brandenburg Academy of Sciences (DWDS) (https://www.dwds.de/) includes 106 million words. This information system consists of three components: dictionaries, text corpora, and word statistics. It includes both modern and historical dictionaries, along with an etymological dictionary of the German language. The Russian National Corpus (http://www.ruscorpora.ru) conducts linguistic analysis on 2 billion words. It consists of a collection of individual corpora, each aimed at solving specific linguistic problems. Its features include frequency dictionaries, accentological, educational, panchronic, multimedia, and MultiPARKi sub-corpora.

Additionally, there are other corpora such as the Eastern Armenian National Corpus (http://www.eanc.net/EANC/), the Czech National Corpus (http://ucnk.ff.cuni.cz/), the Turkish National Corpus (turkic_corpora.html), the Spanish Language Corpus (http://www.corpusdelespanol.org/), the Italian Corpus (corpora.dslo.unibo.it), the Lancaster Corpus of Chinese (www.lancaster.ac.uk/fass/projects/corpus/LCMC), the Polish National Corpus (nkjp.pl), the Bulgarian National Corpus (www.ibl.bas.bg/BGNC_bg.htm), the Corpus of Modern Ukrainian (www.mova.info/corpus.aspx), and others.

Recently, the corpora of modern Turkic languages have been systematized. They include the Bashkir Poetic Corpus, the Almaty Kazakh Language Corpus, the Crimean Tatar Corpus, the Electronic Corpus of the Khakas Language, the Turkish National Corpus, and the Electronic Corpus of Texts in the Shor Language[15].

2.4. The Development of Corpus Linguistics in Kazakhstan

Work on creating the National Corpus of the Kazakh Language began in 2010. The project was first initiated by the A. Baitursynov Institute of Linguistics within a scientific research project. The created corpus consists of ten sub-corpora: the main corpus, the oral corpus, the dialectal corpus, the cultural-representative corpus, the corpus of proverbs and sayings, the historical corpus, the parallel corpus, the onomastic corpus, and the advertising corpus. The total volume of texts in the corpus amounts to 40 million words. The Kazakh Speech Corpus 2 (KSC2) developed by the Institute of Smart Systems and Artificial Intelligence, Nazarbayev University, is the first open-source Kazakh speech corpus. It includes both the Kazakh speech corpus and the Kazakh text-to-speech conversion corpus[16]. The corpus consists of approximately 1,200 hours of high-quality transcribed data containing 600,000 sentences. The Almaty Corpus of the Kazakh Language where more than 86% of the over 40-million-word forms have undergone grammatical analysis. One more corpus is by the Sh. Shayakhmetov National Scientific and Practical Center "Til-Qazyna" since 2021. This corpus is called 'Sub-Corpora of the Kazakh Language National Corpus' (https://qazcorpora.kz/) that provides the data for the present research. Within this corpus, linguistic analyses are being conducted on texts in journalistic style. According to the 2024 statistics, the corpus encompasses 13-million-word usages. Media texts were sourced from Kazakh-language newspapers such as Ana Tili, Egemen Kazakhstan, Zan, Turkistan, Kazakh Adebieti, Aikyn, Astana Akshamy, Zhas Alash, and the Akikat journal. The texts in the database vary in genre: they include biographies, interviews, retrospective interviews, diaries, domestic narratives, stories, opinions, news, events, epics, memoirs, recollections, appeals, essays, reflections, informational reports, announcements, presentations, and others.

The corpus data has become a starting point for the research in frequency. The first frequency dictionary of the Kazakh language was compiled in 2016 by researchers A. Zhubanov, A. Zhanabekova, B. Karbozova, and A. Kozhakhmetova. The dictionary is organized into three sections based on frequency principles: alphabet-frequency, frequency-alphabet, and reverse-alphabet, covering 2,102,496 words and featuring a word list of 30,515 words. Later, E. Kazybek and A. Fazylzhanova published "A Frequency Dictionary of the Kazakh Language for General Education." In 1973, A. Akhabaev published "An Alphabet-Frequency Dictionary of the Language of Modern Kazakh Newspapers" and in 2020, A. Zhubanov, A. Zhanabekova, D. Toqmurzayev, and B. Otegenova prepared the "Frequency Dictionary of Oral Kazakh Texts".

In recent years, Kazakh scholars have also explored the theoretical and methodological effectiveness of corpus materials, as seen in books and monographs as "Using Corpora as Authentic Materials in the Language Learning Classroom", "Linguistic Research Based on National Corpora (Based on Kazakh, Russian, and English Language Materials)", and "The Application of Corpus Methods in Teaching Russian Language and Literature to Foreign Students"[17].

2.5. The Typological Uniqueness of Kazakh

Kazakh, as an agglutinative language, exhibits distinct structural features that set it apart from isolating languages like English and inflectional languages like Russian and Chinese. Agglutinative morphology is based on the sequential attachment of affixes, where each affix contributes a specific grammatical or semantic function. This results in a regular one-to-one correspondence between form and meaning, enabling efficient expression of complex ideas within single words.

For instance, the word 'üýlerimizden' [from our houses] consists of: üý [root: house], -ler [plural: houses], -imiz [1st person plural possessive: our], and -den [ablative case: from]. This compact encoding contrasts with how the same meaning is expressed in other languages. In English, it would be 'from our houses'; in Russian, 'из наших домов' [iz nashikh domov], combining case endings and possessive forms; and in Mandarin Chinese, '从我们的房子' [cóng wǒmen de fángzi], using the possessive particle 'de'[18].

Kazakh's agglutinative system not only enables efficient word formation but also showcases a high degree of

morphological productivity and flexibility. For example, the word 'jazylmadym' [I was not written about] is formed by combining several affixes: jaz- [to write], -yl- [passive], -ma- [negation], and -dym [1st person singular, past tense]. Each affix here contributes a distinct grammatical function, reinforcing the efficiency and regularity of the Kazakh agglutinative structure[19].

Moreover, Kazakh is characterized by vowel harmony, a feature absent in both isolating and inflectional languages. Vowel harmony ensures that suffixes harmonize with the root vowel, as demonstrated in words like 'üyde' [in the house] and 'orda' [in the horde], where the suffix '-de/-da' harmonizes with the root vowel, contributing to fluid expression. Additionally, Kazakh displays *homonymy*, as seen in the word 'jazylmadym', which can mean both 'I was not written about' and 'I have not recovered', depending on the context[20].

These features underscore the typological uniqueness of Kazakh within the Turkic language family. They highlight its ability to convey complex meanings efficiently and its distinctive morphological dynamics, which are central to understanding Kazakh's linguistic structure as explored in the present corpus-based study.

3. Materials and Methods

3.1. Data Collection

The data for this study were collected from the texts of the Sub-corpora of the National Corpus of the Kazakh Language (qazcorpora.kz) using the functional setting of the corpus software. The texts comprise diverse genres such as literature, news articles, and spoken language. Word frequency analysis was conducted on the corpus outcomes, providing a representative sample of language usage in contemporary Kazakh across various contexts and registers.

3.2. Data Preprocessing

The linguistic preprocessing involved a series of essential steps to ensure the accuracy and reliability of the frequency analysis. These included tokenization, lemmatization, and morphological annotation, tailored to the agglutinative nature of the Kazakh language. Tokenization segmented the corpus into analyzable units such as words and sentences, accommodating language-specific features like vowel harmony and extensive affixation. Lemmatization normalized word forms to their base (lemma), addressing the complexities of Kazakh's inflectional system and compound structures. Morphological annotation assigned grammatical categories to tokens, enabling precise part-of-speech tagging and grammatical form analysis. The preprocessing utilized tools integrated within the qazcorpora.kz platform, specifically optimized for the Kazakh language's unique structural and phonological characteristics. This approach ensured the accurate representation of linguistic phenomena, such as the functional variability of affixes and morphophonemic alternations, thereby enhancing the validity of the corpus-based analysis and its alignment with advanced methodologies in corpus linguistics.

3.3. Data Analysis

The analysis of the corpus, particularly in the context of Kazakh's unique linguistic features, involves both manual and automated processes. Initially, linguistic experts analyzed the data using the morphological analyzer, which correctly handled the vowel harmony and extensive affixation present in Kazakh. The software was designed to identify and process phonological patterns, such as vowel harmony, where suffixes harmonize with root vowels. This ensures accurate morphological parsing.

Since this manual process, the analysis has become increasingly automated, with the system now relying on a data-driven approach based on the same foundational corpus. Additionally, the system is capable of distinguishing lexical and grammatical homonyms within context, ensuring that meanings are properly disambiguated. This enhanced accuracy in analyzing the complex morphological structures of Kazakh—such as affixation and vowel harmony—helps maintain the integrity and precision of the corpus, which is essential for advanced linguistic studies.

4. Results

4.1. Frequency of Word Classes in Kazakh

Table 1 presents the frequency indicator (hereafter FI) of 78 words, including 11 functional words, 11 pronouns, 22 verbs, 12 nouns, 7 modal words, 9 adjectives, 7 adverbs, and 2 numerals. These are selected for analysis because their FIs are at the level of 4,000, that is high.

Table 1. FI of word classes.

1. Functional Words		2. Pronouns		3. Verbs		4. Nouns		5. Modals		6. Adjectives		7. Adverbs		8. Numerals	
Word	**FI**	**Word**	**FI**	**Word**	**FI**	**Word**	**FI**	**Word**	**FI**	**Word**	**FI**	**Word**	**FI**	**Word**	**FI**
žäne	*29846*	bul	*28524*	dep	*23769*	Kazakh	*14803*	žoq	*11692*	aq	*10040*	köp	*8728*	eki	*8024*
üšın	*17643*	osï	*19446*	degen	*17272*	adam	*8849*	kerek	*11023*	žaŋa	*7909*	eŋ	*6951*	mïŋ	*5262*
twralï	*10080*	sol	*14817*	emes	*13607*	žumïs	*7951*	bar	*10609*	memlekettik	*7486*	qazir	*5197*		
ɣana	*9530*	onïŋ	*13004*	al	*11725*	Kazakhstan	*7674*	sekildi	5274	ulttïq	*6852*	endi	*4858*		
boyïnša	*9500*	öz	*11316*	edi	*11110*	žïl	*6875*	qažet	*4282*	ülken	*5247*	äri	*4503*		
deyin	*7662*	biz	*7389*	bolïp	*11051*	Abay	*6232*	siyaqtï	*4014*	älewmettik	*5109*	žoɣarï	*4354*		
pen	*7278*	bizdiŋ	*6184*	eken	*8697*	Bilim	*5494*	tärizdi	*838*	žalpï	*4857*	kezinde	*4208*		
biraq	*7031*	ne	*5407*	boldï	*7932*	söz	*5277*			žaqsï	*4804*				
keyin	*6989*	onï	*5348*	bolɣan	*7861*	memleket	*4984*			ulï	*4745*				
arqïlï	*5947*	olardïŋ	*4664*	bolïp	*7196*	žol	*4314*								
soŋ	*4387*	barlïq	*4662*	boladï	*6992*	halïq	*4171*								
				kelgen	*6200*	qatar	*4016*								
				bolsa	*6047*										
				žatqan	*6047*										
				otïr	*5498*										
				ötken	*5431*										
				deydı	*5427*										
				keledi	*5273*										
				dedi	*5188*										
				bolatïn	*4477*										
				žatïr	*4474*										
				alɣan	*4124*										

(1) Functional words. The word in the first position has FI 29,846, indicating that the conjunction ‘žäne’[and] is used more frequently compared to others. In the Kazakh language, the coordinating conjunction ‘žäne’ is used to connect nouns, verbs, phrases, and clauses that are coordinated within compound sentences. The frequency of the conjunction ‘žäne’ is observed to be 5% higher than the pronoun ‘bul’ [this] in the second position, 21% higher than the verb ‘dep’ in the third position, 51% higher than the noun ‘Kazakh’ in the fourth position, 61% higher than the modal word ‘žoq’ [not], 67% higher than the adjective ‘aq’ [white], and 71% higher than the adverb ‘köp’ [many/much].

Medetbekova[21] describes the functional-semantic and functional-stylistic features of the conjunction ‘žane’. She identifies ontological consistencies and inconsistencies in its use. The author points out that through analyzing the numeric occurrence of the conjunction ‘žane’ across journalistic, scientific, and literary writings, particularly in relation to nouns, it is clear that, similar to the conjunction ‘men’ [and], the conjunction ‘žane’ is mainly prevalent in journalistic style. This statement supports the active use of this conjunction in journalistic texts.

Although the conjunction ‘žäne’ and the conjunctions ‘men /ben/ pen’ [and] are functionally synonymous, there are instances where they cannot substitute for one another. Moreover, another conjunction ‘üšin’ [for], which conveys purpose and causality in the Kazakh language when combined with a root word, is also frequently used in journalistic style. Its frequency is 10% lower than that of the demonstrative pronoun ‘osï’ [this] and 3% higher than the verbal noun ‘degen’ [that].

(2) Pronouns. The frequency of the demonstrative pronoun ‘bul’ [this] in expressing intermediate relations in Kazakh, which is used as a substitute for other words, shows 28,524 occurrences, while ‘osï’ [this] itself occurs 19,446 times. The frequent use of these demonstrative pronouns in journalistic style is likely related to their anaphoric and deictic functions.

(3) Verbs. Among verbs, the use of the auxiliary verb ‘dep’ [by saying] in comparison to other forms is high with FI 23,769. ‘Degen’, the participle form of ‘de’ [say] occurs 17,272 times, the present form ‘deydi’ [says] appears 5,427 times, and the past form ‘dedi’ [said] is used 5,188 times. Therefore, the auxiliary form ‘dep’ is more frequently used than the others.

(4) Nouns. In journalistic texts, the noun ‘Kazakh’ occurs 14,803 times. Its frequent use in journalistic style is likely tied to the genre’s focus on Kazakh public-political and cultural issues. It is often employed to showcase national identity or to construct a particular image of Kazakhstan and its people. In the Kazakh language, the noun ‘Kazakh’ is frequently found in possessive structures, such as ‘Kazakh qogamï’ [Kazakh society], ‘Kazakh halqï’ [Kazakh people], ‘Kazakh ultï’ [Kazakh nation], ‘Kazakh dalasï’ [Kazakh steppe], ‘Kazakh žeri’ [Kazakh land], ‘Kazakh tili’ [Kazakh language], ‘Kazakh ädebieti’ [Kazakh literature] and so on.

It was determined that the word ‘Kazakh’ being an ethnonym referring to the people, culture, and history of Kazakhstan, is also frequently used in informal speech. This word is one of the key concepts related to the country and describes

the nationality, identity, and distinctiveness of its inhabitants. It has a distinct combinatorial characteristic in informal texts, for example: every Kazakh, all Kazakhs, today's Kazakh, which Kazakh; Kazakh knows, as a Kazakh, what Kazakhs like, what is called Kazakh, that which is Kazakh; in Kazakh districts, in Kazakh homes; it often combines with prepositions such as for and about – for Kazakhs, about Kazakhs, as a Kazakh; it also combines with nouns in the third-person possessive form – Kazakh humor, Kazakh schools, Kazakh authority, Kazakh men, Kazakh women. Additionally, it is frequently used with emotionally descriptive words like Kazakh – boastful, Kazakh – ambitious, Kazakh – timid.

(5) Modals. The modal verbs 'žoq' [not] and 'kerek' [need] are actively used in journalistic style, while the usage of 'sekildi' [similar], 'qažet' [necessary], and 'siyaqtï' [like] is around 45–34% compared to the first two. In the Kazakh language, the modal words 'kerek' [need] and 'qažet' [necessary] differ in frequency of use by 61% while both convey a semantic nuance of necessity or obligation to perform an action; however, they have functional differences as the modal 'kerek' possesses its higher collocational capacity.

The modal verb 'kerek' expresses the speaker's desire or intent to perform an action (e.g., 'oqwïm kerek' [I need to study], 'žazgïm kerek' [I need to write] and signifies a subjective relationship arising from the necessity of an object, e.g., 'aqša kerek' [money is needed], 'scenariy kerek' [scenario is needed]. Thus, the modal word 'kerek' conveys the semantic category of necessity and performs functional roles of personal evaluation, obligation, or command. Meanwhile, 'qažet' predominantly conveys the importance or inevitability of a particular event or process. Its functional characteristic lies in expressing necessity from the perspective of logic, ethics, laws, or accepted standards. It is often used to formalize requirements and recommendations. For example, phrases such as 'qoldanisqa alw öte qažet' [it is very necessary to put into practice], 'zertteudi qažet etetin' [requires research], 'estetikalïq damwdï qažet etedi' [needs aesthetic development], 'memlekettik qoldaw qažet' [state support is needed], 'maquldanyan recenziyasï qažet bolsa' [if an approved review is needed], 'ulttïq mekteptiŋ žaŋa ülgisi qažet' [a new model of the national school is needed], 'tüsindirme sözdikterin žasawdï qažet etedi' [needs to create explanatory dictionaries], 'tolassïz izdenis qažet' [continuous research is necessary], and so on.

The following are cases where these modal words can be interchangeable. For example, 'alwïm kerek' [I need to get] can be replaced with 'alwïm qažet' [I need to get], 'istew kerek' [need to do] can be replaced with 'istew qažet' [need to do] and so on. However, there are certain cases where kerek cannot replace qažet, such as 'awaday qažet' [essential as air] cannot be replaced with 'awaday kerek' [essential as air], 'šïnï kerek' [to be honest] cannot be replaced with 'šïnï qažet' [to be honest] and 'žatsa kerek' [it must be the case] cannot be replaced with 'žatsa qažet' [it must be the case].

(6) Adjectives. The most frequently used adjective is 'aq' [white] with FI 10,040. This adjective is often used in its literal sense in journalistic texts. For example, 'aq tüs' [white color], 'aq lalagül' [white lily], 'aq maqta' [white cotton], 'aq kiyiz' [white felt], 'aq žolaq' [white stripe], 'aq köylek' [white dress], etc. It is also found in a figurative sense. The metaphorical use of adjectives with meanings related to color, size, temperature, tactile sensation, and sound occupies a certain place in the texts of modern print media [22]. Its active use in journalistic texts is linked to its symbolic meaning as well. This word is associated with concepts such as purity, innocence, kindness, generosity, openness, and truthfulness, and is often used to describe moral or ideological qualities. For instance, 'aq niyet' [pure intention], 'aq köŋil' [kind-hearted], 'aq söyleytin' [truthful speaker], 'aq batam' [blessing], 'aq žüziŋ' [pure face], 'aq tarih' [true history], and others.

In journalistic style, it often appears in contexts of national, cultural, and political associations. Examples include 'Aq Orda' [White Palace], 'Aq Arwana' [White Camel], 'aq patša' [white king], 'aq saqaldï' [white-bearded], 'aq nayza' [white spear], 'aq bilek' [white arm], 'aq öleŋ' [white verse], 'aq žol' (blessing for a successful way), and more.

The adjective 'žaŋa' [new] is also frequently used in journalistic texts to vividly describe certain phenomena or events serving as a method or tool for the journalist's expressive language. For example, 'suw žaŋa kiyindirip' [dressed brand new], 'žaŋa qïrïnan körsetedi' [shows from a new perspective], 'žaŋa keyipkerlerdi ömirge äkelw' [bringing new characters to life], '100 žaŋa esim' [100 new names], 'qïrïq ötiriktiŋ žaŋa türi' [a new type of forty lies], 'žaŋa ay men köne ay (new moon and old moon), žaŋa estetıkalïq keŋistikte (in a new aesthetic space), and so on.

In addition to the mentioned above, adjectives such 'memlekettik' [state], 'ulttïq' [national, and 'älewmettik' (so-

cial) also show high frequency in journalistic texts. These adjectives are employed to provide a precise and comprehensive description of various phenomena occurring within the state and society. These adjectives are important in the context of public information because they lend a sense of formality, authority, and seriousness to the discussion. Additionally, they play a key role in describing socio-political and socio-cultural realities, making them an integral part of journalistic style.

(7) Adverbs. The adverb 'köp' [many/much] often combines with verbs and nouns. This quantity adverb tends to collocate with nouns, describing the general number of something as large. It does not necessarily provide an exact number, as a general indication of abundance or scarcity is sufficient[23]. In the corpus, it is used more frequently than the others, with a frequency of 8,728. The intensifying adverb 'eŋ' [most] is used 20.4% less than köp, 40.5% less than 'qazir' [now, 44.4% less than 'endi' [now], 48.5% less than 'äri' [and, further], 50.2% less than 'žoɣarï' [higher], and 51.8% less than 'kezinde' [at the time]. In the corpus, the adverb 'köp' [many/much] frequently collocates with verbs, nouns, derived adjectives, intensifying adverbs, quantity adverbs, and restrictive particles like in 'köp kördim' [I saw many], 'köp otïratïnmïn' [I used to sit a lot], 'köp bolsa' [if there are many], 'mümkindikter köp' [many opportunities]), köp qoy [it's too much], etc.

In a text from the newspaper "Kazakh ädebieti" in 2020, which belongs to the literary genre and discusses political and public life, the adverb 'köp' [many/much] appears 8 times in a historical account in phrases like 'köp adam' [many people], 'žaralanɣandar öte köp' [many were injurded], 'köp balalï' [many children], 'köp bolganïn' [there were many], etc. The author employs these expressions to describe the number of subjects, the repetition of certain actions, the severity of the situation, the duration of an event, and temporal concepts. The use of 'köp' in journalistic style strengthens the storyline and enhances the emotional impact on the reader.

(8). Numerals. The most frequently used numeral is 'eki' [two] with FI 8,024. It serves an informational function and is used in its direct and specific meaning. In journalistic style, its role in evoking emotions and making an impact on the reader is also prominent. It is highly functional in conveying the quantitative and qualitative data of social life.

Regarding its usage in artistic text style, for example, in the "Kazakh Ädebieti" newspaper article Žïlan Monšaq [The Snake Bead] by M. Etemadzada, translated from Persian and published in 2020, the numeral 'eki' [two] is used 8 times. Examples include: 'eki-üš tösek žaymanï' [two or three bed sheets], 'eki bölmesi' [two rooms], 'eki mïŋdïq tiyïn' [two-thousand coin], 'eki ayaɣïnïŋ astïna' [under his two feet], 'eki bala' [two children], etc. Additionally, in an interview from an internet resource in the public domain, the numeral 'eki' demonstrates active collocational capacity with other parts of speech. For example, 'eki žïl' [two years], 'eki mïŋ teŋge' [two thousand tenge], 'eкi topka' [into two groups], eki žigit [two guys], and others.

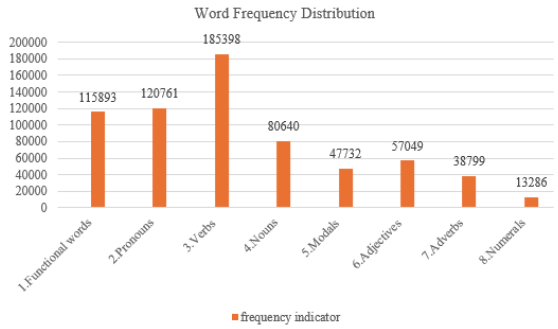

Figure 1. Word classes distribution in the sub-corpora of the national corpus of the Kazakh language.

Figure 1 represents a distribution of different parts of speech or word classes, with corresponding frequency counts. As it shown, functional words (FI 115,893) are common in the dataset; these are words that have little lexical meaning but serve to express grammatical relationships between other words in a sentence. Pronouns (FI 120761) have a slightly higher frequency than functional words, indicating a significant use of pronouns in the dataset. The largest category is verbs (FI 185,398), which are typical, as verbs are central to sentence construction and tend to appear frequently. Though the number of nouns (FI 80,640) is substantial, it is less frequent than verbs, indicating verbs might dominate the action or structure in this dataset. The frequency of modals (FI 47,732) is relatively high but not as much as the main verb category. Adjectives (FI 57,049) have a moderate frequency in the data and adverbs (FI 38,799) appear less frequently compared to adjectives. Numerals (FI 13,286) are the least frequent category, indicating that specific numerical references are relatively rare in the corpus. Overall, the data indicates a typical language distribution with a higher frequency of functional words, pronouns, and verbs, while

numerals and other specific word categories like adverbs appear less frequently.

Thus, Kazakh journalistic texts represent a great use of conjunctions, pronouns, verbs, nouns related to concepts of nationality and country, some modal words, adjectives describing socio-political and cultural realities, adverbs of quantity, intensification, and time, and cardinal numerals. Their frequent usage is closely tied to their multifunctional grammatical roles in the language, as well as the author's intention to structure and connect ideas, apply personal stylistic choices, narrate events emotionally, exaggerate certain details, and other related factors.

4.2. Frequency of Functional Words in Kazakh

In the Kazakh language, as well as in other agglutinative languages, the role of functional words is consistent. Banguoglu[11] admits that functional words serve to link nouns to other elements within a sentence. They lack independent meaning when used alone, but they play a critical role in expressing the connection between pairs of words or sentences.

Thus, according to their distinctive features functional words are divided into conjunctions, postpositions, and particles. **Table 2** provides the FI of mostly used functional words in the corpus.

Table 2. FI of functional words.

1. Conjunctions		2. Postpositions		3. Particles	
Word	**Frequency**	**Word**	**Frequency**	**Word**	**Frequency**
žäne	*29846*	üšın	*17643*	ɣana	*9530*
biraq	*7031*	twralï	*10080*	tek	*4094*
nemese	*3406*	boyïnša	*9500*	qana	1403
öytkeni	*3178*	deyin	*7662*		
		keyin	*6989*		
		arqïlï	*5947*		
		soɳ	*4387*		

Functional words, with their function of connecting words and sentences, are among the most actively used parts of speech compared to other word classes. Based on statistical data from the corpus, the various types of functional words (as shown in **Table 2**) demonstrate high activity. Among conjunctions, 'žäne'[and] and 'biraq' [but] are frequently used; among postpositions, 'üšın' [for], 'twralï' [about], 'boyïnša' [according to]; among particles, 'ɣana' [only] and 'tek' [just] show high usage. In contrast, conjunctions like 'nemese' [or], 'öytkeni' [because], and the particle 'qana' [only] show lower frequency. Thus, among functional words, there are 2 conjunctions, 7 postpositions, and 2 particles.

If we compare the frequency of the Russian conjunction 'i' [and] with the Kazakh 'žäne' [and], according to the Russian National Corpus, which includes 131,488 texts and 374,449,975 words, the frequency of 'i' [and] is 34,272.1 in literary texts and 33,196.28 in non-literary texts. According to corpus statistics, its frequency is lower. Meanwhile, in the Sub-corpora of the Kazakh National Corpus, with 13 million words, the frequency of the conjunction 'žäne' [and] in journalistic texts is 29,846, representing 0.22%.

In the Russian frequency dictionary on ruscorpora.ru, with a total of 374,449,975 words, the frequency of the preposition 'dlya' [for] is 1,908.75 in literary texts and 3,854.88 in non-literary texts. Meanwhile, in Kazakh, within the corpus of 13,000,000 words, the frequency of the postposition 'üšın' [for] is 17,643, or 0.14%.

Functional words are eligible to subordinate one word or sentence to another, requiring specific forms for the words they follow and expressing various grammatical, temporal, spatial, and purposive relations[19]. Among postpositions, their role in official styles is more significant compared to coordinating and limitative postpositions. In official styles, postpositions such as 'qatïstï' [referring to], 'qarsï' [against], 'baylanïstï' [related] often appear as they play an important role in expressing purpose, reason, and comparisons between subject and object movements[24].

When analyzing the frequency of functional words, an interesting linguistic fact about the usage of the particles 'ɣana/qana' [only] was discovered. The particle 'ɣana', which expresses restriction, is used 85% more frequently than its 'qana' variant, which occurs after hard consonants. The particle 'ɣana' commonly appears with words ending

in vowels, voiced consonants and sonorants. For example: ‘qïzmetine ɣana’ [only for the service], ‘säl ɣana’ [just a little], ‘aspaptarïn ɣana’ [only their instruments], etc. Additionally, the particle ‘ɣana’ is used after interrogative pronouns like ‘qaydan’ [where from], ‘nege’ [why], and ‘qalay’ [how], placing emphasis on the word and often adding a nuance of regret, as in ‘qaydan ɣana’ [where only], ‘nege ɣana’ [why only], and ‘qalay ɣana’ [how only].

Meanwhile, the particle ‘qana’ is used with words that end in hard consonants, as well as voiced consonants such as /b/, /v/, /g/, and /d/. Examples include: ‘13 ret qana’ [only 13 times], ‘mektep qana’ [only the school], ‘keŋistik qana’ [only the space], and so on. In the corpus, it is observed that ‘qana’ frequently combines with verbs ending in -ïp, -ip, and -p. Examples include: ‘daäleldep qana’ [only proving], ‘oqïp qana’ [only reading, ‘barïp qana’ [only going], and others.

Thus, the more frequent usage of one variant (‘ɣana’) over the other (‘qana’) in Kazakh word combinations, especially in publicistic and colloquial styles, is due to: 1. Kazakh words tend to end more often in vowels and sonorant sounds than in hard consonants. This natural tendency of Kazakh phonology—where voiced sounds dominate at the end of words, and harmony between consonants and vowels plays a role—explains the prevalence of ‘ɣana’ over ‘qana’. 2. The functional and stylistic role of these particles, which add meanings of limitation or restriction, supporting logical flow in speech or writing, also contributes to this difference in frequency.

Therefore, when analyzing functional words in the corpus, it becomes evident that the frequency of the conjunction ‘žäne’ [and] is particularly high in publicistic texts. This can be attributed to its grammatical function of connecting words (both nouns and verbs), coordinating clauses in compound sentences, and emphasizing or clarifying ideas within a sentence. The frequent use of the postposition ‘üšın’ [for] can be attributed to its function of attaching to nouns in the nominative case or verbal nouns, adding causal and purposive grammatical meanings. Thus, the broad functional-grammatical roles of the conjunction ‘žäne’, as well as the postpositions ‘üšın’ in Kazakh, explain their frequent usage. These linguistic tools play a significant role in helping the speaker or writer express their thoughts clearly and effectively within a text.

4.3. Grammatical Form Frequency of Pronouns in Kazakh

Corpus materials allow us to determine not only word frequency and part-of-speech frequency but also the frequency of grammatical forms. **Table 3** shows the frequency of the grammatical forms of some pronouns. The reason for focusing on the frequency of pronouns is that, according to the frequency statistics, pronouns rank second place in the overall frequency.

Table 3. FI of grammatical forms of pronouns.

Bul	FI	**Osï**	FI	**Sol**	FI	**Öz**	FI	**Biz**	FI
Bul	28524	Osï	19446	Sol	14817	Öz	11316	Biz	7389
buɣan	2228	osïɣan	1307	sonïmen	3789	özi	6355	bizdiŋ	6184
budan	1763	osïnïŋ	709	sonïŋ	2876	öziniŋ	3236	bizge	1797
bular	371	osïnï	636	sodan	2240	özin	1756	bizde	896
bulardïŋ	294	osïlardïŋ	149	soɣan	1281	özine	1615	bizdi	799
bunïŋ	150	osïmen	128	sonï	901	özim	1057	bizden	254
bunï	133	osïlar	61	solardïŋ	842	özderi	759	bizben	165
bunïmen	28			solar	146	özderiniŋ	582		
						özinen	459		
						özimen	249		

According to the frequency rates, the pronouns ‘bul’ [this], ‘osï’ [this], ‘sol’ [that], ‘öz’ [own], ‘biz’ [we], are ones of the most frequently used in the nominative case. ‘Bul’ [this] in the nominative case has FI 28,524, while its dative case form ‘buɣan’ has FI 2,228, the ablative case form ‘budan’ has FI 1,763, the plural form ‘bular’ has 371, the possessive plural form in the genitive case ‘bulardïŋ’ has FI 294, the singular form in the genitive case ‘bunïŋ’ has FI 150, and the accusative case form ‘bunï’ has FI 133. The form ‘bunïmen’ is rarely used in journalistic texts, appearing only 28 times.

Among reflexive pronouns, which express the idea of separating the speaker or subject from other substances or

phenomena, the most frequently used form is ‘öz’ [own]. It is used much more often than its other grammatical forms such as ‘özinің’, ‘özin’, ‘özine’, ‘özim’, ‘özderi’, ‘özderinің’, and ‘özinen’, with FI 11,316. In the corpus texts, ‘öz’ is often used in conjunction with nouns, for example: ‘öz uaqyty’ [own time], ‘öz müddesi’ [own interest], ‘öz žumïsï’ [own work] and many others. In these constructions, the noun following ‘öz’ typically appears with the third-person possessive suffix and then is declined (e.g., ‘öz waqït+ï+ң’), or the noun is in the first-person plural possessive form and then declined (‘öz qatar+ïmïz+ɣa’). In some cases, the noun is first pluralized, then possessively marked, and declined (‘öz eңbek+ter+i+niң’ [of their own works]).

The pronouns ‘barlïq’ [all] and ‘bäri’ [everyone], which convey a generalizing grammatical meaning, have different usage frequencies. The pronoun ‘barlïq’ is used 67.7%, while ‘bäri’ is used 32.3%. Examples of ‘barlïq’ in use include: ‘barlïq ulttarï’ [all nations], ‘barlïq žerde’ [everywhere], etc. From these examples, it is clear that ‘barlïq’ frequently combines with nouns and verbal forms that have been substantivized. As a result, ‘barlïq’ is used more often than ‘bäri’. The model ‘barlïq + noun’ is commonly found in corpus-based journalistic texts. However, there are instances where ‘barlïq’ is incorrectly used with plural suffixes (e.g., ‘barlïq-tarï’, ‘barlïqtarïñïz’). According to the rules of the Kazakh language, since ‘barlïq’ semantically implies a plural meaning, it should not take a plural suffix, but some language users violate this rule.

In contrast, pronoun ‘bäri’ is used in examples like: ‘bäri bos’ [everything is empty], ‘bäri qïzïq’ [everything is interesting], ‘istïң bäri’[(all the work], etc. The combinatorial properties of ‘bäri’ are broader than those of ‘barlïq’, as it frequently combines with adjectives, nouns, pronouns, and verbs. Especially when it is used with nouns and pronouns, it often forms possessive constructions (e.g., ‘osïnïң bäri’ [all of this]). In the corpus, the frequency of ‘bäri’ is 54%, while ‘barlïɣï’ is used 46% more often. However, the frequency of their accusative forms shows the opposite trend where ‘barlïɣïn’ is 28% less frequent than ‘bärin’.

In conclusion, the pronouns ‘bul’, ‘osï’, ‘sol’, ‘öz’, ‘biz’, ‘barlïq’, ‘bäri’, and ‘kim’ are frequently used in journalistic texts. Their grammatical forms, such ‘buɣan’, ‘budan’, ‘osïɣan’, ‘sonïmen’, ‘sodan’, ‘soɣan’, ‘özi’, ‘özinің’, ‘özim’, ‘bizdің’, ‘bizge’, ‘barlïɣï’, and ‘bärïn’, while also common, rank second in frequency, with more than 1,000 occurrences. The remaining grammatical forms of these pronouns are rarely used.

5. Discussion

This study marks a pioneering effort in employing a corpus-based word frequency analysis for the Kazakh language. The absence of prior similar research in Kazakh highlights the novelty and significance of this work. By focusing on word frequency, this research aligns with methodologies widely used in linguistic studies of other languages, such as English, Spanish, and Chinese. For instance, studies like Leech et al.[25] for the British National Corpus (BNC) and He et al.[26] for Mandarin Chinese emphasize the value of corpus-based frequency data in understanding linguistic trends. Additionally, this study builds on the methodology established in the previous research[15] focused on the generating the Kazakh Russian parallel corpus and concordances. So, the current study logically extends the efforts by applying the corpus-based frequency analysis to further enrich the understanding of Kazakh linguistic features, while continuing the development of digital resources for the Kazakh language.

A comparison reveals that while foreign corpora benefit from decades of refinement and larger datasets, this study’s smaller scale provides a foundation for future expansions. Moreover, cultural and structural differences, such as agglutinative morphology in Kazakh, necessitate unique adaptations to corpus techniques. The findings contribute to global corpus linguistics by demonstrating the potential for expanding corpus methods into underrepresented languages.

Future research should prioritize enhancing the search functionality of the Kazakh National Corpus by enabling genre-based text categorization. Advanced filters could allow researchers to differentiate texts authored by professionals and non-professionals, as well as unedited texts, facilitating a more nuanced analysis of linguistic changes in Kazakh. Furthermore, systematic development of the corpus would support extensive research into accentological and morphological variants, the productivity of word formation processes, paronyms, synonym groups, and structural variations in agreement and government. These advancements would provide a robust foundation for exploring the complexities of Kazakh linguistic phenomena.

The limitations of this study include the potential bias in the corpus selection, as it primarily consists of written texts, which may not fully represent spoken language usage. Additionally, the corpus may not cover all regional dialects or specialized domains, potentially limiting the generalizability of the word frequency findings. Furthermore, changes in language usage over time were not considered, which could affect the representativeness of the data.

The implications of this study highlight the potential for corpus-based approaches to inform language teaching, lexicography, and language policy in Kazakh. By identifying high-frequency words and linguistic patterns, the study provides valuable insights into contemporary language usage, supporting the development of educational materials and language resources.

6. Conclusions

The study of frequently used parts of speech and grammatical forms in mass media texts plays an important role in identifying dynamic changes in the Kazakh language and in explaining sociolinguistic variation and language change. Our quantitative research presents models of the frequency of functional words and pronouns based on the corpus method. As the content of the database expands, the FI of word forms will continue to change.

To address the research gaps identified in this study, a more focused exploration of the Kazakh National Corpus is necessary. While the corpus includes journalistic and literary texts, its range should be expanded to incorporate specialized sub-corpora, such as spoken and historical texts, to deepen the understanding of language variation and change. Additionally, the study highlights the need for detailed stylistic and functional analyses of specific text types. For example, examining conjunctions, modal verbs, and pronouns in journalistic texts could uncover their role in emphasis and persuasion, providing valuable linguistic insights.

A comparative cross-linguistic framework is also essential to contextualize findings about the Kazakh language within global research. Comparing frequency patterns of parts of speech and grammatical forms with other languages can reveal typological distinctions and shared trends. Furthermore, applying these findings to practical contexts - such as creating more precise frequency dictionaries, educational tools, and NLP applications - would ensure the corpus serves not only academic interests but also broader societal and technological needs.

Author Contributions

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

Funding

This research is funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR24993001 “Creation of a large language model (LLM) to maintain the implementation of Kazakh language and increase the technological progress”).

Acknowledgments

We express our gratitude to the Sh. Shayakhmetov National Scientific and Practical Center «Til-Qazyna», which supports the implementation of the project.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Data will be available upon request.

Conflicts of Interest

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript;

or in the decision to publish the results.

References

[1] Mastrantuono, A., Regan, B., 2024. Present perfect and preterit variation in the Spanish of Lima and Mexico City: Findings from a corpus analysis. Corpus Linguistics and Linguistic Theory. 20(2), 375–405. DOI: https://doi.org/10.1515/cllt-2022-0060

[2] Stefanowitsch, A., 2020. Corpus linguistics: A guide to the methodology. (Textbooks in Language Sciences 7). Language Science Press: Berlin, Germany. pp. 1–490.

[3] Jung, Y., Gablasova, D., Brezina, V., et al., 2024. Developing a coding scheme for annotating opinion statements in L2 interactive spoken English with application for language teaching and assessment. Research in Corpus Linguistics. 12(2), 146–173. DOI: https://doi.org/10.32714/ricl.12.02.07

[4] Kilgarriff, A., 2001. Comparing corpora. International Journal of Corpus Linguistics. 6(1), 1–37.

[5] Nilsson, F., 2019. A comparative analysis of word use in popular science and research articles in the natural sciences: A corpus linguistic investigation [Ph.D. Thesis]. Västerås, Sweden: Maraldalen University. pp. 1–89.

[6] Shin, D., Paul, N., 2007. Beyond single words: The most frequent collocations in spoken English. ELT Journal. 62(4), 339–348. DOI: https://doi.org/10.1093/elt/ccm091

[7] Baayen, H., 1992. Statistical models for word frequency distributions: A linguistic evaluation. Computers and the Humanities. 26, 347–363. DOI: https://doi.org/10.1007/BF00136980

[8] Zasorina, L.N., 1997. Chastotnyy slovar russkogo yazyka [The frequency dictionary of the Russian language]. Russkij yazyk: Moscow, Russia. pp. 1–923.

[9] Zotina, E.V., Solovyov, V.D., 2012. Diachronic changes in the frequency of nouns based on the material of the national corpus of the Russian language. Scientific notes of Kazan University. Humanities Series. 154(5), 34–44.

[10] Kim, N.M., 2010. Funktsionirovanie imen chislitel'nykh v publitsisticheskikh tekstakh [The functioning of numerals in journalistic texts]. Vestnik TGPI Gumanitarnye Nauki. 2, 153–168.

[11] Banguoğlu, T., 2004. Dil edatlari [Prepositions]. Isparta: Istanbul, Turkey. pp. 1–56.

[12] Makarenko, F.D., 2016. Rol, mesto i chastota upotrebleniya samostoyatel'nykh i sluzhebnykh chastey rechi v tekste [The role, place and frequency of the use of independent and official parts of speech in the text]. Molodoy Uchonyy. 2(106), 908–912.

[13] Plecháč, P., Kolár, R., 2015. The corpus of Czech verse. Studia metrica et poetica, 2(1), 107–118.

[14] Ormanova, A.B., Anafinova, M.L., 2022. Linguistic interference in information space terms: A corpus - based study in Kazakh. Theory and Practice in Language Studies. 12(12), 2497–2507. DOI: https://doi.org/10.17507/tpls.1212.04

[15] Baishukurova, G., Irgebayeva, A., Aitova, N., et al., 2024. The creation of concordance as an effective tool for studying the text: On the example of A. Baitursynov's concordance. Forum for Linguistic Studies. 6(5), 51–64. DOI: https://doi.org/10.30564/fls.v6i5.6856

[16] Mussakhojayeva, S., Khassanov, Y., Varol, H.A., 2022. KSC2: An industrial - scale open - source Kazakh speech corpus. Proceedings of the 23rd InterSpeech Conference; Incheon, South Korea, 18–22 September 2022. pp. 1367–1371.

[17] Aitova, N., Ospanova, D., 2024. Verb - based emotive structures in the linguistic corpus base. Toraygyrov University Bulletin. Philological Series. 1, 55–69.

[18] Hung - Yeh Tiee, H., 1979. The productive affixes in Mandarin Chinese morphology. Word. 30(3), 245–255. DOI: https://doi.org/10.1080/00437956.1979.11435670

[19] Zhanpeisov, E., 2002. Qazaq gramatikasy: fonetika, sözjasam, morfologia, sintaksis [Kazakh grammar: Phonetics, word formation, morphology, syntax]. Astana: Astana, Kazakhstan. pp. 1–132.

[20] Muhamedowa, R., 2016. Kazakh: A Comprehensive Grammar. Routledge: London, UK. pp. 1–299.

[21] Medetbekova, P.T., 2015. Linguostatistical analysis of conjunctions "Men" vs. "Zhäne". Qazaq Universiteti: Almaty, Kazakhstan. pp. 1–124.

[22] Yuneev, V.V., 2007. Metaphorization of words in the texts of modern journalism [Candidate of Philological Sciences Thesis]. Moscow, Russia: State Pedagogical Institute. pp. 1–230.

[23] Ermukhamet, M., 2020. Mölsher kategoriya'synyn tarikhi paradigmacy (lingvistikalyk aspektide) [The historical paradigm of the measure category (in the linguistic aspect)] [Ph.D. Thesis]. Almaty, Kazakhstan: Al - Farabi Kazakh National University. pp. 1–160.

[24] Alkebaeva, D.A., 2020. Qazaq tilining pragmastilistikasy: Oqulyq [Pragmastilistics of the Kazakh language: Textbook]. Almaty, Kazakhstan: Qazaq Universiteti. pp. 1–62.

[25] Leech, G., Rayson, P., Wilson, A., 2001. Word Frequencies in Written and Spoken English: Based on the British National Corpus. Longman: London, UK. pp. 1–304.

[26] He, Y., Chow, J.Y.J., Nourinejad, M., 2017. A privacy design problem for sharing transport service tour data. Proceedings of IEEE ITS Conference; Yokohama, Japan, 16–19 October 2017. pp. 1–1359.