ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 15, No. 3, pp. 677-687, March 2025
DOI: https://doi.org/10.17507/tpls.1503.02

A Corpus-Based Approach in Vocabulary Research: Defining the Word of the Year 2023 in Kazakh

Assel B. Ormanova
Department of General Education Disciplines, Astana IT University, Astana, Kazakhstan

Madina L. Anafinova
Pedagogical Institute, Astana International University, Astana, Kazakhstan

Dana Zh. Ospanova
Terminology Department, National Scientific and Practical Center "Til-Qazyna", Astana, Kazakhstan

Zhibek K. Tleshova
Department of General Education Disciplines, Astana IT University, Astana, Kazakhstan

***Abstract*—The Word of the Year (WOTY) is an event held in various countries and regions to determine the most relevant, significant, and popular words and expressions that reflect not only the linguistic but also the socio-cultural aspects of the country. This paper aims to identify the most frequently used words/phrases in Kazakh for 2023 to be nominated for the WOTY title. The research methods include media discourse analysis and quantitative analysis using a corpus-based approach. A computer program, #LancsBox 6.0, generated a dataset—a research corpus consisting of 500 texts published on Kazakh news platforms throughout 2023. The results indicated that: 1) the conjunction "jáne" [and] had the highest frequency and occurrence in the research corpus; 2) the extracted words with high frequency indicators might serve as candidates for WOTY 2023, such as "Kazakhstan", "jana" [new], "jyly" [year], "kerek" [need], "jumys [work]"; 3) WOTY "artificial intelligence" named by other global sources showed a high frequency indicator in Kazakh media texts. The study contributed with the generated corpus of media texts in Kazakh for 2023. The significance of our study is highlighted by the pioneering linguistic assessment in Kazakh language, which involves the analysis of media discourse publications based on corpus outcomes.**

***Index Terms*—The Word of the Year (WOTY), Kazakh, corpus, vocabulary, frequency indicator**

I. Introduction

Linguistics has always been a dynamic field, constantly evolving to analyze and understand the intricate nature of human language. One of the current trends in linguistics is the application of computational methods and techniques (Peterson, 2013). Computational linguistics applies to scientific methods of language analysis such as corpus linguistics, which allows for processing large texts or working with the already existing world corpora. The analyst needs to provide empirical evidence in the form of data extracted from language corpora to support any statement made regarding language (Brezina, 2018).

The potential and capabilities of a corpus have offered significant benefits not only to language researchers but also to the broader social context, with substantial importance for the future (Mastrantuono & Regan, 2024). A corpus-based approach is applicable in vocabulary analysis, where researchers use corpora to explore and determine the usage, frequency, and distribution of words in different contexts. This provides a possibility to define words as the most popular and/or frequently used words in a one-time period.

The essence of this linguistic and cultural procedure lies in the fact that at the end of each year, the most widespread and frequently used word, phrase, or phraseological unit in the media and online communication receives the title "Word of the Year" (WOTY). Ryabova and Sergeichik admit that the WOTY is a word (or phrase) characterized by the greatest frequency and occurrence in media discourse during a calendar year, which, according to the results of expert rating evaluation, is defined as the most linguistically resonant. Experts like scholars, publishers, leaders of organizations and societies, are usually initiators of these ratings, together with linguists and researchers from other fields of science (2019).

In recent decades, in many countries, selecting the Word of the Year has become an important linguistic and cultural tradition. Since the 1990s, this practice has expanded into the English-speaking linguistic and cultural sphere: in 2003, Words of the Year emerged in Holland, in 2005 - in France, in 2006 - in Denmark, in 2007 - in Russia, in 2011 - in Poland, and so on (Nikolaeva, 2017). To our knowledge, this tradition has not yet been initiated in Kazakhstan.

Thus, this study seeks to define the WOTY 2023 in Kazakh using the corpus-based approach. Consequently, the research questions are as follows:

1. What is the WOTY 2023 in Kazakh?
2. What are the WOTYs 2023 in other linguistic and cultural contexts? What are their frequency indicators in Kazakh?

The significance of our research lies in conducting the first linguistic rating in Kazakh, which is based on analyzing a variety of media discourse publications and performing a statistical analysis of the lexicon used. The contribution is a comprehensive dataset – a research corpus comprising 500 publications, totaling 267,205 words and 1,846,818 characters, sourced from Kazakhstani online news platforms for the year 2023.

II. Literature Review

A. Definition

The conceptual and cognitive understanding of the word meaning is one of the challenging aspects of Word of the Year practice as it refers to how words convey ideas and represent objects, actions, and concepts. This includes understanding the relationship between words and their referents, as well as the mental processes involved in interpreting and assigning meaning to linguistic signs.

Selecting Word of the Year has gained popularity in recent years, with various organizations and publications marking a word that best represents the year's cultural, social, and political climate. According to Ryabova and Sergeichik (2019), the occurrence of the Word of the Year is viewed as both a reflection and a product of how specific cultural values are expressed linguistically within a society at a specific time. This is classified as a linguo-cultural idea, recognizing its linguistic essence.

Nikolaeva (2017) characterizes the Word of the Year as a concept pertaining to language and culture that highlights the most significant political, economic, and socio-cultural occurrences of a specific year. Additionally, it captures new significant components within the evolving and dynamic semantic framework of the national language's cultural landscape.

As stated by Melnik (2016), identifying the Word of the Year enables language experts to evaluate language contributions over the course of a full year, to condense its verbal and conceptual meanings, and to understand the focus on societal values and the potential for its ongoing evolution.

B. Historical Background

The first attempt to identify the Word of the Year in the verbal space of mass communication took place in 1971 in Germany at the initiative of the German Language Society, when “aufmüpfig” (obstinate, recalcitrant) was chosen as the WOTY. In 1984, the Japanese publishing house "Dajisen Kokuminsha" first identified the "Most Popular Words". Since 1991, the WOTY has been announced by the American Dialect Society. In 2000, the WOTY version from the Global Language Monitor agency appeared, along with additional categories such as "Phrase of the Year" and "Name of the Year." In 2004, another version of the English "Words of the Year" was added by the Oxford English Dictionary publishing house, with separate editions for the British and American languages. Since 2006, Merriam-Webster has been offering its own version of WOTY. At the same time, The New York Times began publishing annual WOTY lists compiled by American lexicographer G. Barrett (Martseva et al., 2018).

In Russia WOTY has been selected since 2007 to continue an existing global practice, initiated by M. Epstein, a Russian-American linguist, cultural researcher, and critic. Stepanov (2007) claims that the initiative is sociolinguistic as it incorporates a large group of informants and specialists, alongside a thorough examination of media ratings. Between 2009 and 2016, popular terms were selected in collaboration with experts via public voting on the imhonet.ru website. The initial Russian term of the year – “гламур” [glamour] – was chosen in 2007. This term demonstrated Russian society's inclination towards consumer culture, entertainment, and a leisurely lifestyle.

The selection process for the WOTY typically involves reviewing trends in language usage, analyzing cultural and political events and considering the overall mood and atmosphere of the year. Words that have been selected in the past have ranged from political terms like "fake news" and "Brexit" to cultural phenomena like "selfie" and "vape". These words are often reflective of the major events and trends that have shaped the year, providing a snapshot of the collective consciousness at that time (Zimmer, n. d.).

In determining the Word of the Year, it is essential to clarify the various linguistic approaches to its selection. The British approach is centered on the criterion of its frequency in the media. The French Word of the Year is chosen in two directions: a popular word reflecting the pressing issues of society is selected through voting, while "le nouveau mot de l'année" refers to a neologism that is added to dictionaries. The Russian approach is based on the vote of native speakers (Buryakovskaya & Dmitrieva, 2017).

C. WOTYs 2023 in Different Linguistic and Cultural Contexts

In 2023 WOTYs were also selected across different linguistic and cultural contexts, revealing the diverse linguistic issues, trends, and values that shape global and regional discourses.

After carefully analyzing over 32,000 votes, a team of language experts announced that Oxford's Word of the Year for 2023 was a shortened form of "charisma" — "rizz." This term refers to someone's ability to captivate another person through their style, charm, or irresistible appeal. According to the corpus, the use of this word significantly increased in 2023, reaching its peak in June of this year.

The Cambridge Dictionary (Online https://dictionary.cambridge.org/) has chosen "hallucinate" as the Word of the Year for 2023. This decision comes after a notable rise in fascination with generative artificial intelligence (AI) technologies such as ChatGPT, Bard, and Grok. People's focus has turned to the shortcomings of AI and whether these challenges can be addressed. Tools powered by AI, particularly those that utilize large language models (LLMs), can produce seemingly accurate text but frequently rely on incorrect, deceptive, or entirely fabricated "facts." They "hallucinate" with a level of confidence that can sometimes appear credible. In response to this shift, the Cambridge Dictionary has revised its meaning of "hallucinate" to capture this updated context.

Collins Dictionary (Online https://www.collinsdictionary.com/) declared artificial intelligence (AI) as the Word of the Year for 2023, defining it as "the simulation of human cognitive processes by computational programs." As we consider the forthcoming significant technological advancement, AI has progressed swiftly and has become a hot topic throughout 2023. Analysts examined the Collins Corpus, which is a collection of over 20 billion words sourced from global websites, newspapers, magazines, and books, as well as spoken language from radio, television, and daily conversations. They found that the frequency of the term had increased markedly, establishing it as the primary subject of discussion in 2023.

Merriam-Webster (Online https://www.merriam-webster.com/) has chosen "authentic" as the Word of the Year for 2023, a term frequently associated with both personal and national identity. The word "authentic" encompasses multiple interpretations, such as "genuine or not a copy", similar to the concepts of real and actual; and "faithful to one's own character, spirit, or nature." Typically, a word that many look up each year, "authentic" experienced a notable surge in interest in 2023, fueled by discussions and narratives surrounding artificial intelligence, celebrity culture, personal identity, and social networking.

The Pushkin Institute of the Russian Language (https://slovogoda.pushkin.institute/) determined "нейросеть" (neural network, neuronet) as Word of the Year 2023. A two-step methodology was used. In the first stage, specialists turned to a survey of colleagues, and as a result, a list of candidate words was formed. It also included "искусственный интеллект" (artificial intelligence), ChatGPT, "импортозамещение" (import substitution), "СНГ" (CIS), "заимствование" (borrowing), "педагог" (teacher), "наставник" (mentor), and "волонтёр" (volunteer). In the second stage, it was decided to evaluate the frequency of candidates using tools such as Wordstat, Google Trends, and Medialogy. As a result, the undisputed winner was revealed – the word "нейросеть" (neural network, neuronet). According to Wordstat, starting in February 2023, the frequency of use of this word has been consistently high. Medialogy revealed that in the year 2023, the media and social platforms featured 104,680 postings containing this specific term. In contrast, the figures for 2022 stood at 33,438 postings, while in 2021, there were 31,097 postings. Therefore, the word's frequency tripled in a year. In 2023, "нейросеть" (neural network, neuronet) was included in the Spelling academic resource "Academos", suggesting that the word's frequency in mass media texts has exceeded the threshold for moving to its codification. Moreover, within the Russian-speaking community, the term "нейросеть" (neural network) successfully surpasses its nearest rival – the expression "искусственный интеллект" (artificial intelligence). However, The Moscow Times (Online https://www.themoscowtimes.com/) – Russia's oldest independent media outlet – announced "искусственный интеллект" (artificial intelligence) as its choice for a word that best characterized 2023.

TABLE 1
WOTYS 2023 IN SOME LINGUISTIC AND CULTURAL CONTEXTS

Source	Word of the Year 2023
Oxford Languages Dictionary	Rizz (charisma)
Merriam-Webster Dictionary	Authentic
Cambridge Dictionary	Hallucinate
Collins Dictionary	Artificial intelligence
The Moscow Times	Искусственный интеллект [artificial intelligence]
The Pushkin Institute of the Russian Language	Нейросеть [neural network, neuronet]

Analyzing all these candidates for Word of the Year in different lingua-cultures (Table 1), there is a trend that links them all, and that is the development and integration of ICT. Technology has been driving our lives in this century. Four words directly or indirectly related to AI. Collins went for it directly, while Merriam-Webster and Cambridge opted for words that describe what AI has done. As AI can create a copy of anything and anyone, the identity and originality of the real world are being challenged. This way, AI has posed challenges for authenticity and originality. This is why Merriam-Webster and Cambridge chose "authentic" and "hallucinate" as the Words of the Year. So, 2023 is largely driven by Artificial Intelligence, starting with the revolutions brought about by ChatGPT and other AI tools (Panda, 2024).

D. A Corpus-Based Research in Kazakhstani Settings

(a). Frequency Detection

In recent decades, numerous empirical studies have been conducted on the frequency of Kazakh words. One such study is the "Frequency Dictionary of Kazakh oral language texts", which determines the statistical structure of the Kazakh vocabulary. The volume of text from all styles recorded from oral language exceeds 50 thousand words, according to the application. The dictionary reveals the most frequent and rare variants of words in Kazakh and variants of competitive words (doublets) in the literary norm (Jubanov et al., 2020).

The dictionary "The frequency dictionary of the Kazakh language in general education" was prepared and submitted to the press on behalf of the Minister of Education and Science as part of the 2016 research project "Development and creation of the National Corpus of the Kazakh language." The dictionary is based on a text database with more than 7 million phrases. The dictionary's language base (36,265 words) provides daily public vocabulary and consists of texts of various discourses that enter communication in everyday life. The dictionary represents five styles of the Kazakh language (scientific, publicistic, official, spoken, and belles-lettres style) and a mixed text base containing texts of different styles (Qajybek & Fazyljanova, 2016).

(b). Corpora Representation

Currently, the Kazakh language is represented by four corpora: 1) The National Corpus of the Kazakh Language (40 million words), 2) The Almaty Corpus of the Kazakh Language (40 million words), 3) The Sub-Corpus of the National Corpus of the Kazakh language (13 million words), and 4) The Kazakh Speech Corpus 2 (KSC2).

The National Corpus (Online https://v2.qazcorpus.kz/) was initiated in 2010 and 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, with a total volume of 40 million words. It includes information on agreement, base forms of words, and language indicators such as structure, word formation, vocabulary, sounds, and meaning-related aspects. The Almaty Corpus of the Kazakh language (Online http://web-corpora.net/) is a variant of the National Corpus of the Kazakh language serving as a reference framework, established through significant financial support for annotated texts in literary Kazakh. The Sub-Corpus of the National Corpus of the Kazakh Language (Online https://qazcorpora.kz/) developed by the National Scientific and Practical Center “Til-Qazyna” since 2021, compiles biographies, interviews, retrospective interviews, diaries, domestic narratives, stories, opinions, news, events, epics, memoirs, recollections, appeals, essays, reflections, informational reports, announcements, presentations, and others. The Kazakh Speech Corpus 2 (KSC2) (Online https://issai.nu.edu.kz/) is an attempt by the local research community of Nazarbayev University. It is designed to be a large-scale corpus containing over 135 million words and includes approximately 1,200 hours of high-quality transcribed data containing 600,000 sentences (Makhambetov et al., 2013).

(c). Natural Language Processing

There are some studies on machine learning, deep learning techniques, or natural language processing for the Kazakh language. The field of word extraction for Kazakh papers is quite advanced, with a few relevant publications. Most research in this area is conducted by specialists in information technologies and communications (ITC) technologies who focus on natural language processing (NLP) rather than by linguists.

Myrzakhmetov and Kozhirbayev (2018) used both LSTM- and classic n-gram-based neural networks to undertake language modelling experiments on a newspaper dataset. They found that neural-based models performed better than n-gram-based models.

Nugumanova and Mansurova (2019) looked at statistical and graph-based techniques and demonstrated how the term derivation works and the theme modelling task relates to one another. Additionally, a broad overview of automatic word recognition is provided, along with discussion and implementation of examples in Python libraries and the R ecosystem, as well as working methodologies for difficult topics like terminology.

Rakhimova et al. (2021) provide the study and development of a question-answer system for the Kazakh language that is based on the BERT paradigm. In the study, the questions and replies were translated from English into individual files, resulting in a corpus of 60,000 sentences.

Abibullayeva and Çetin (2022) created the Kazakh News Data Set (KND), which consists of 7,060 online news articles. They utilized the Bidirectional Encoder Representations from Transformers (BERT) model, a widely used method in natural language processing (NLP), to explore keyword extraction in the Kazakh language.

Aitova and Ospanova (2024) based on the verb-based structures taken from the National Corpus of the Kazakh language, examined how words change emotionally to articulate ideas and feelings, as well as their emotional condition in both a standalone context and in verbal exchanges. This investigation unveiled the semantic, grammatical, lexical, and functional characteristics of emotive elements in the Kazakh language.

III. Methodology

A. Data Collection Procedure

Our research involved two stages.

In the initial stage, we determined the WOTYs 2023 of English and Russian by analyzing data from well-established dictionaries like Oxford, Merriam-Webster, Cambridge, Collins, and institutions and organizations responsible for such linguistic ratings. In our search, we focused on three languages (Kazakh, Russian, and English) that operate in Kazakhstan. Currently, in the country, there is a great influence of Russian and English on the Kazakh language due to globalization, integration, historical reasons, wide spread of English and other extra-linguistic factors (Ormanova & Anafinova, 2022).

The next stage aimed to generate a dataset. The reason for building the research corpus was that the available Kazakh corpora could not provide statistical data on the frequency of the determined WOTYs 2023 in Kazakh, nor did any other organizations. Conversely, the generated research corpus met our goals by providing frequency indicators for the already determined WOTYs.

B. Instrumentation

Among a wide range of specialized software tools for working with texts of various formats, a new generation corpus tool, #LancsBox 6.0, was used (see Figure 1). This instrument was developed by the Corpus Approaches to Social Science Center at Lancaster University (UK) (Brezina et al., n. d.) and is designed for the analysis, processing, and visualization of corpus language and text data. The computer program allows users to work with and upload existing world corpora to a personal computer. Additionally, the program's functionality enables the creation of research text corpora. #LancsBox 6.0 offers search functions at different levels of corpus annotation using: a) simple search; b) wildcard search; c) intelligent search; d) regular expression search. The corpus tool provides the following functions: KWIK, Graphicall, Whelk, Words, Ngrams, Text, Wizard. The program supports working languages such as Chinese, English, French, and Russian.

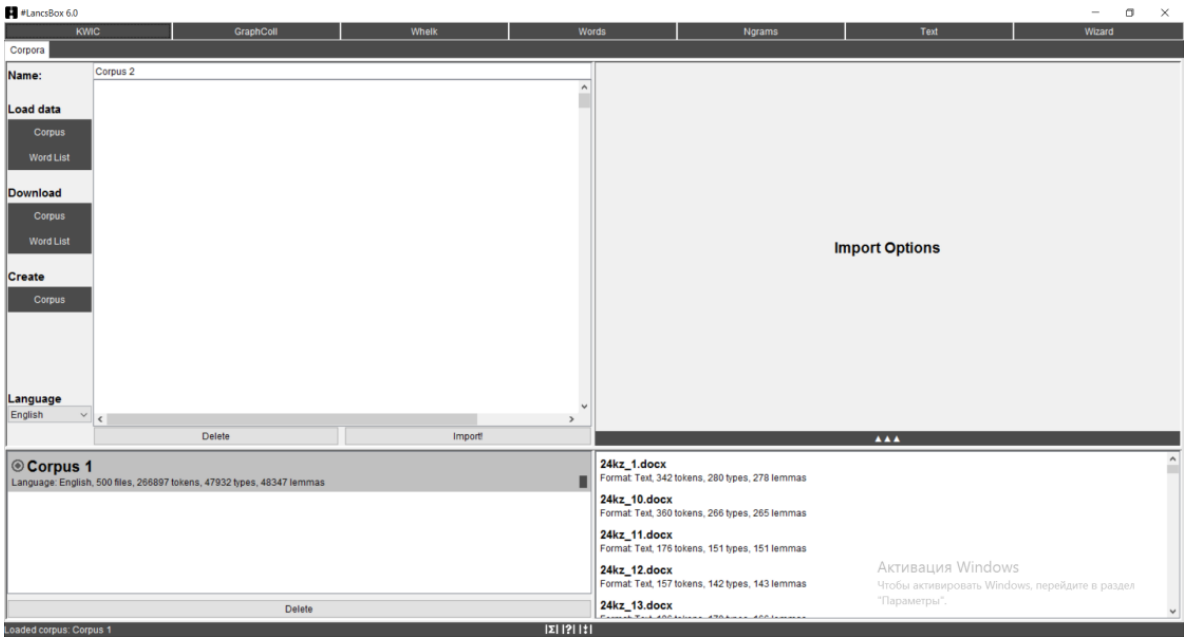

Figure 1. The Interface of #LancsBox 6.0 Program

In gathering data, we chose popular online news websites that are well-regarded by the public in Kazakhstan. To create the corpus, we picked five hundred articles from twenty different news platform sites published in Kazakh. The selection of newspapers was based on the ratings available on https://aqparat.info/feed. The texts were then loaded into the #LancsBox 6.0 computer program, where the Text function combined and presented all the publications in one text; the Words function provided a list with frequency; and the KWIC function generated collocations with possible word forms. The statistics of the dataset are presented in Table 2.

TABLE 2
STATISTICS OF THE RESEARCH CORPUS

News platform	Access	Publications	Tokens	Types	Lemmas
7kun.kz	https://7kun.kz/	15	3417	2033	1989
24kz	https://24.kz/	15	4085	2154	2115
365info.kz	https://365info.kz/	15	7323	3502	3522
Aiqyn.kz	https://aikyn.kz/	15	6120	3313	3262
Akorda.kz	https://www.akorda.kz/	15	8916	3611	3597
Stan.kz	https://stan.kz/	15	5618	2878	2852
Bugin	https://www.bugin.kz/	34	21693	8381	8362
Forbes.kz	https://forbes.kz/	50	56584	14907	14980
Inbusiness.kz	https://www.inbusiness.kz/	15	9298	4091	4036
Informburo	https://informburo.kz/	15	3818	2015	1953
Inform.kz	https://www.inform.kz/	30	13190	5780	5748
Massaget	https://www. massaget.kz/	33	13005	5669	5673
NUR.KZ	https://www.nur.kz/	50	40436	11651	11616
Redtram	https://rus.redtram.com/	15	3633	2102	2062
Spuntik Қазақстан	https://sputnik.kz/	8	5572	2667	2617
Tengrinews.kz	https://tengrinews.kz/	10	5285	2651	2559
Zakon.kz	https://www.zakon.kz/	15	10195	4615	4555
Егемен Қазақстан	https://egemen.kz/	70	35464	13741	13717
КНБ	https://www.gov.kz/memleket/	15	4559	1845	1713
Kazpress	https://kazpress.kz/	50	8686	3926	3916
Total: 20		**500**	**266897**	**47932**	**48347**

As shown, the corpus includes 500 articles, totaling 266,897 tokens, with 47,932 different word forms and 48,347 lemmas. While the overall size of this collection is considered limited, it guarantees a wide variety of subjects within the news articles and offers diverse content. The tagged media publications were released throughout a year-long timeframe in 2023.

IV. RESULTS

Firstly, the corpus tool provided the data for the already established WOTYs 2023, which covers other lingua-cultures and their representation in the Kazakh texts. Secondly, the software generated a frequency list of all word entries to be used in identifying buzzwords in Kazakh.

A. Research Corpus Data of the WOTYs 2023

Based on the research data from the corpus, the frequency indicators for established WOTYs are as follows, as presented in Table 3.

TABLE 3
FREQUENCY INDICATORS FOR ESTABLISHED WOTYS IN THE KAZAKH MEDIA TEXTS

WOTY	Kazakh Equivalent	Frequency Indicator
artificial intelligence	‘жасанды интеллект’ [jasandy intellekt]	72
neuronet	‘нейрожелі’ [neirojelı]	15
authentic	‘сайма-сай’ [saima-sai]	10
rizz (charisma)	‘харизма’ [harizma]	2
hallucinate	‘галлюцинация’ [galüsinasia]	2

According to the table, the phrase “jasandy intellekt” [artificial intelligence] was mentioned 72 times in 11 publications. For example, the most frequent occurrence (23 times) was found in the publication "The Head of State took part in the international forum Digital Bridge 2023" (Online https://www.akorda.kz/kz/memleket-basshysy-digital-bridge-2023-halykaralyk-forumyna-katysty-1294420), where the role of artificial intelligence was emphasized in the current reality. Another publication, "Artificial Intelligence saves time: 5 skills that no longer need to be learned" (Online https://www.bugin.kz/31101-zhasandy-intellekt-uaqytty-unemdeydi-endi-uyrenudinh-qadgeti-dgoq-5-daghdy), with 14 occurrences, discussed current competencies that could easily be performed by AI, such as writing, art design, data entry, data analysis, and video editing. "A startup that does not receive income, but the cost is growing every year" (Online https://forbes.kz/articles/tabyis_tappaytyin_bra_nyi_jyil_sayyin_setn_startap_1702006137) used “artificial intelligence” 10 times to describe the Kazakhstan-based Cerebra AI project, which is among the top five startups in the world for diagnosing brain diseases, particularly detecting strokes using AI. Thus, 11 out of 500 news publications used “jasandy intellekt” [artificial intelligence] as a reflection of the current trends in the political, economic, and social spheres of the country. The software provided a list of collocations, frequency indicators, and word forms of “jasandy intellekt” [artificial intelligence] in the research corpus (see Figure 2).

Figure 2. Corpus Data of 'Jasandy Intellekt' [Artificial Intelligence]

Looking at the word formation, the Kazakh compound term "jasandy intellekt" is a direct translation of the English "artificial intelligence". "Jasandy" is a Kazakh adjective meaning fake, untrue, or false, while "intellekt" is a borrowing from Russian. After examining the current terminological dictionaries of the Kazakh language, we could not find its definition. Since this term is relatively new, we can suggest that the definition in Kazakh is almost identical to that in English or Russian. Although the worldwide web offers the following: 'Жасанды интеллект (ЖИ, ағылш. artificial intelligence, AI) — интеллектуалды компьютерлік бағдарламалар мен машиналар жасау технологиясы әрі ғылым (*Eng.* technology and science of creating intelligent computer programs and machines)' (Online https://kk.wikipedia.org/wiki/Жасанды_интеллект). This term is registered in the Kazakhstani terminological database Termincom.kz (Online https://termincom.kz/search/?termin=жасанды%20интеллект%20&cid) in the sections Culture and art, Philosophy and political science, starting from 2021. The online translation source Sozdik.kz (Online https://sozdik.kz/ru/dictionary/translate/kk/ru/жасанды%20интеллект/) introduces "jasandy intellekt" as artificial intelligence, referring to the information technologies group.

B. The Most Frequent Word in the Research Corpus

The second part of our study focused on working with the built research corpus and identifying the most frequently used word or phrase within it. While processing the data, we utilized the "Words Searches" function to compile a list of the most commonly occurring words. The annotation of this list included 266,897 tokens, 47,932 types, and 48,347 lemmas. The search revealed that the conjunction "jáne" had the highest frequency of 103.26, occurring 2,756 times and appearing in 440 out of 500 texts. It is used to connect two or more related elements in a sentence, adding information or combining ideas.

Furthermore, the next most frequently used word with 1,953 occurrences was "men" which in Kazakh functions in two ways: the first singular personal pronoun "I" and the conjunction "and". It is worth noting that the National Scientific and Practical Center "Til-Qazyna" (Online https://tilqazyna.kz/), together with several domestic newspapers, proposed the "TOP 100 words". The first place in this list was taken by the word "men". According to analytical data, "men" was repeated 39,337 times in a 13-million-word corpus. Thus, in this way our results partially coincided with similar research.

C. The Candidates for WOTY 2023 in Kazakh

By expanding the scope of the frequency list, we identified the first twenty words (including nouns, adjectives, and verbs) that possess a conceptual worldview characteristic to describe the linguistic and cultural context in the country during the year 2023 (see Figure 3).

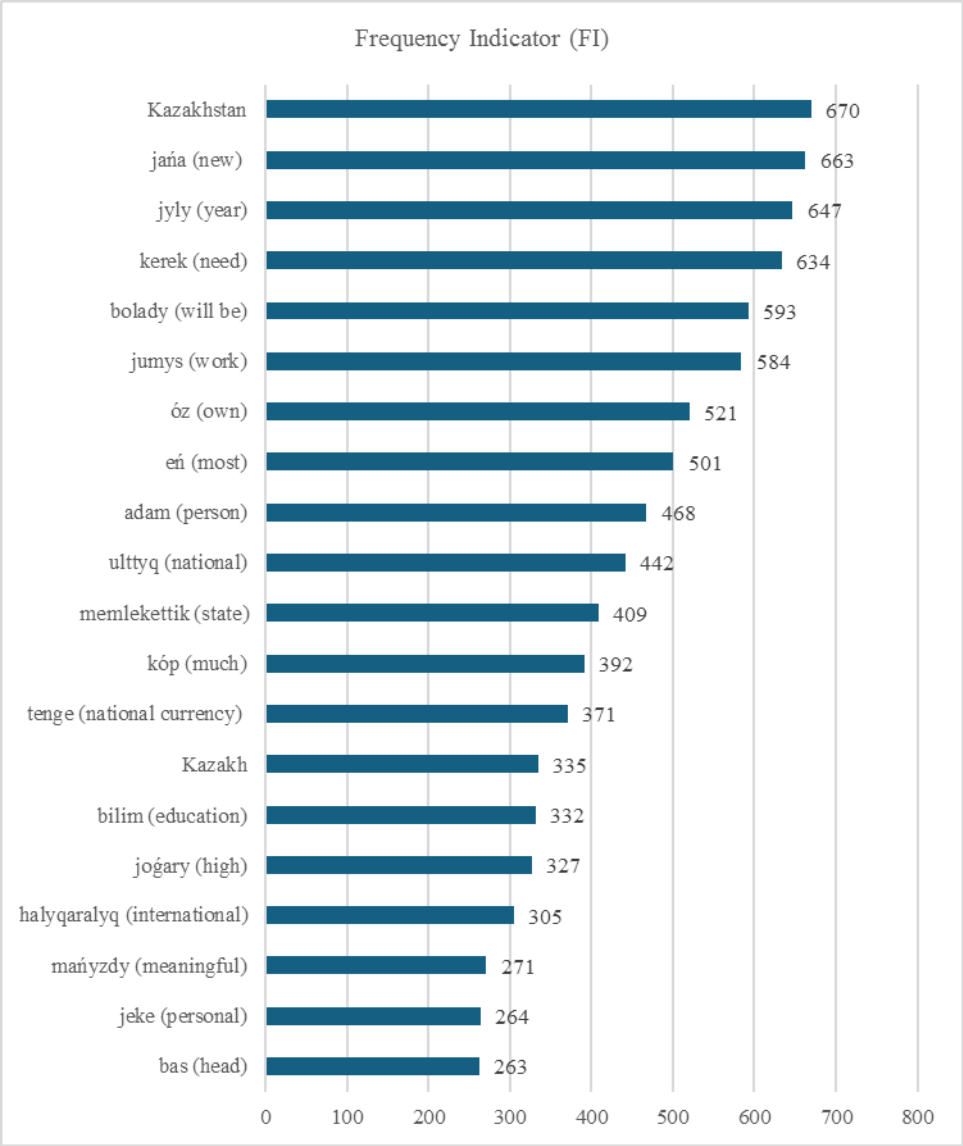

Figure 3. Frequently Used Concepts in Kazakh in the Research Corpus

The name of the country “Kazakhstan” (FI 670) and the adjective “jańa” [new] (FI 663) had the highest frequency indicators. We suggest that the combination of these words means “New Kazakhstan”, which stands for the updated policy of the country, regulating new strategies and directions for the country's development, along with people with new orientations and trends. Some of the words have a political coloring, such as “ulttyq” [national] (FI 442), “memlekettik” [state] (FI 409), and “halyqaralyq” [international] (FI 305), which could imply the country's connection with other states. The verbs “kerek” [need] (FI 634) and “bolady” [will be] (FI 593) suggest the necessity of the future or plans described in the texts. These verbs are likely to refer to “jumys” [work] (FI 584) that needs to be done. The adjectives “eń” [most] (FI 501), “joǵary” [high] (FI 327), and “bas” [head] (FI 263) can be united in meaning by emphasizing a sense of superiority or importance and interpreted as describing something or someone that is the best, greatest, or most important in a particular context.

V. Discussion

The study employed the corpus-based approach to analyze the vocabulary of the Kazakh language and determine the Word of the Year for 2023. By examining the collection of text samples, researchers were able to identify the most frequently used words and determine the term that best encapsulated the spirit of the year. This method proved effective in capturing the linguistic trends and shifts in the Kazakh language over time.

By generating the research corpus with the #LancsBox 6.0 program, researchers can gain insights into the current linguistic trends, issues, concerns, and interests of the Kazakh-speaking population. On the one hand, the most frequently used word was the conjunction “jáne” with 2,756 occurrences. On the other hand, the software provided us with the words with high frequency indicators, such as “Kazakhstan”, “jańa” [new], “jyly” [year], “kerek” [need], “bolady” [will be], “jumys” [work], etc. that have ‘meaning’ and denote a particular concept of the linguistic worldview

to characterize the linguo-cultural situation in the country during 2023. We suggest that the publications tend to be politically motivated and focus on the country's life itself, as there were words like the name of the country (Kazakhstan) with 670 occurrences, national currency (tenge) with 371 occurrences, nation (Kazakh) with 335 occurrences, personal (jeke) with 264 occurrences, and others. Looking from the morphological perspective, the dominance of words consisted of adjectives (11 elements) and nouns (8 elements), while there were only two verbs and an adverb. Such distribution can be interpreted as though items, events, or people were much spoken about and described or characterized, but lacking actions. In comparison with the candidates for WOTYs in other languages, the words that represent technology, ICT, and AI are not reflected in Kazakh.

As there were no similar works in domestic linguistic research that focused on defining the Word of the Year in Kazakh using a corpus-based approach, this highlights the novelty and importance of this study in the field of Kazakh linguistics. Some non-scientific attempts were made on social media sites where readers could propose their ideas for the candidates for WOTY in Kazakh, such as 'OBSE', 'Nur-Otan', but without any statistical data (Online https://adebiet.wordpress.com/2008/01/20/жыл-сөзі/).

One relevant work that supports the importance of a corpus-based analysis in vocabulary research is the study by McEnery and Hardie (2012), who emphasize the significance of corpora in understanding language use and variation. Another study by Stubbs (2001) discusses the advantages of using corpus data in linguistic research, such as providing a comprehensive and systematic analysis of language patterns.

VI. Conclusion

In the realm of language studies, corpus linguistics and computational linguistics have become indispensable tools for analyzing and understanding language. By utilizing large, structured collections of texts, researchers can uncover patterns and trends within language usage that would otherwise go unnoticed. One notable trend in recent years has been the practice of selecting the Word of the Year in various countries, which serves as a reflection of the social, cultural, and political developments of that particular year.

We propose a study on defining the Word of 2023 in Kazakh using a corpus-based approach. The software #LancsBox 6.0 extracts the list of words with high frequency indicators from online media platforms in Kazakhstan. The results showed that the conjunction ‘jáne” [and] possessed the highest frequency (FI 2,756). We extracted frequently used words with a conceptual meaning such as “Kazakhstan” (FI 670), “jańa” [new] (FI 663), “jumys” [work] (647), “adam” [person] (FI 468), “ulttyq” [national] (FI 442), “memlekettik” [state] (FI 409), “bilim” [education] (FI 332), etc. that could be candidates for WOTY 2023 in Kazakh. The announced by Collins Dictionary “artificial intelligence” estimated the highest frequency (FI 72) among other WOTYs in the research corpus. Thus, the paper aimed to define the Word of the Year 2023 in Kazakh and analyze the usage of WOTYs 2023 in other linguistic and cultural contexts in Kazakh through corpus tools.

We consider that the findings of our study are important in several respects. First, there were no similar works in domestic linguistic research that focused on defining the Word of the Year in Kazakh using a corpus-based approach. Second, further research can gain advantages by utilizing the results from the research corpus, which illustrates the linguistic landscape in Kazakhstan's media in 2023.

The limitations of this research are twofold. First, the words were chosen according to their frequency, meaning that the most common words were the only ones considered for analysis. This narrow selection process may have excluded important keywords or phrases that could have provided valuable insights into the topic. Second, the limited sample size of only 500 publications may not have provided a comprehensive representation of the topic. With such a small pool to draw from, the findings may not be generalizable or truly reflective of the broader research landscape.

Further research may be directed towards using voting. By analyzing the words that Kazakh people choose to vote for, researchers can gain valuable insights into the current cultural, educational, and political climate in the country. Additionally, further research in extracting WOTY could greatly benefit from collaborations between linguists, teachers, and Natural Language Processing (NLP) specialists. By combining linguistic expertise with advanced computational techniques, researchers can develop more efficient algorithms for analyzing vast amounts of language data to identify trends and patterns in general and WOTY in particular.

Acknowledgements

The authors extend their gratitude to the developers and mentors of the course "Corpus Linguistics: Method, Analysis, Interpretation" at Lancaster University for generously sharing their theoretical and practical materials, as well as the corpus tool program #LancsBox 6.0 (Brezina et al., n. d.).

References

[1] Abibullayeva, A., & Çetin, A. (2022). Keyword Extraction from Kazakh News Dataset with BERT. *El-Cezeri Journal of Science and Engineering*, *9*(4), 1193–1200. https://doi.org/10.31202/ecjse.1131826

[2] Aitova, N., & Ospanova, D. (2024). Verb-based emotive structures in the linguistic corpus base. Bulletin of Toraighyrov University. *Philology series*, *1*, 55-69. https://doi.org/10.48081/ELDM2166

[3] *Almaty Corpus of Kazakh Language*. (n. d.). Retrieved May 29, 2024, from http://web-corpora.net/.
[4] Brezina, V., Weill-Tessier, P., & McEnery, T. (n. d.). *LancsBox 6.0*. Retrieved November 10, 2021, from http://corpora.lancs.ac.uk/lancsbox.
[5] Brezina, V. (2018). *Statistics in Corpus Linguistics*. Cambridge University Press.
[6] Buriakovskaia, V. A., & Dmitrieva, O. A. (2017). Lingvokulturnye harakteristiki “Slova Goda” [Linguistic and cultural characteristics of the "Word of the Year"]. *İzvestia VGPU. Filologişeskie nauki*, *2*, 101-105.
[7] *Cambridge Dictionary*. (n. d.). Retrieved June 9, 2024, from https://dictionary.cambridge.org/editorial/word-of-the-year.
[8] *Collins Dictionary*. (n. d.). Retrieved June 9, 2024, from https://www.collinsdictionary.com/.
[9] Jubanov, A. Q., Janabekova, A. A., Toqmyrzaev, D., & Otegenova, B. J. (2020). *Qazaq auyzşa til mətinderiniñ jiilik sözdigi* [Frequency Dictionary of Kazakh oral language texts]. Eltanym baspasy.
[10] *Kazakh Speech Corpus 2*. (n. d.). Retrieved May 29, 2024, from https://issai.nu.edu.kz/.
[11] Makhambetov, O., Makazhanov, A., Yessenbayev, Zh., Matkarimov, B., Sabyrgaliyev, I., & Sharafudinov, A. (2013). Assembling the Kazakh Language Corpus [Conference session]. In *Proceedings of Empirical Methods in Natural Language Processing* (pp. 1022–1031). Seattle, WA, USA. https://doi.org/10.13140/2.1.5127.4882
[12] Martseva, T. A., Snisar, A. Yu., Kobenko, Yu. V., & Girfanova K. A. (2018). Neologisms in American electronic mass media. In A. Filchenko & Z. Anikina (Ed.), *Linguistic and Cultural Studies: Traditions and Innovations* (2nd ed., pp. 266–274). Cham.
[13] 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. https://doi.org/10.1515/cllt-2022-0060
[14] McEnery, T., & Hardie, A. (2012). *Corpus Linguistics: Method, theory and practice.* Cambridge University Press.
[15] Melnik, Yu. A. (2016). Sotsial'no-lingvisticheskie proekty "Slovo goda" - 2015: spetsifika i versii [Socio-linguistic projects "Word of the year" - 2015: specifics and versions]. *Bulletin of Omsk State Pedagogical University. Humanities research*, *2*(11), 56-57.
[16] *Merriam-Webster*. (n. d.). Retrieved June 9, 2024, from https://www.merriam-webster.com/.
[17] *Moscow Times*. (n. d.). Retrieved June 9, 2024, from https://www.themoscowtimes.com/.
[18] Myrzakhmetov, B., & Kozhirbayev, Zh. (2018). *Extended language modeling experiments for Kazakh*. CEUR Workshop Proceedings, 2303-2315. Retrieved May 10, 2024, from https://www.academia.edu/117671980/Extended_language_modeling_experiments_for_Kazakh.
[19] *National Corpus of the Kazakh Language*. (n. d.). Retrieved May 29, 2024, from https://v2.qazcorpus.kz/.
[20] *National Scientific and Practical Center "Til-Qazyna*. (n. d.). Retrieved June 9, 2024, from https://tilqazyna.kz/.
[21] Nikolaeva, E. V. (2017). “Slova goda” kak lingvokul’turnye koncepty [“Words of the Year” as linguocultural concepts]. *Philology, Theory & Practice*, *10*(1), 154–157.
[22] Nugumanova, A., & Mansurova, M. (2019). *Tabigi til matinderindegi terminderdi avtomatti turde tanu* [Automatic recognition of terms in natural language texts]. Oskemen.
[23] Ormanova, A. B., & Anafinova, M. L. (2022). A Linguistic Interference in Information Space Terms: A Corpus-Based Study in Kazakh. *Theory and Practice in Language Studies*, *12*(12), 2497-2507. https://doi.org/10.17507/tpls.1212.04
[24] Panda, A. K. (2024). Words of the Year 2023. *Journal of English Language Teaching*, *66*(2), 2-5.
[25] Peterson, M. (2013). Computer Games and Language Learning. *TESL Canada Journal*, *33*(1), 90-108.
[26] *Pushkin Institute of the Russian Language*. (n. d.). Retrieved June 9, 2024, from (https://slovogoda.pushkin.institute/
[27] Qajybek, E. Z., & Fazyljanova, A. M. (2016). *Jalpy bılım berudegı qazaq tılınıñ jiılık sözdıgı* [The frequency dictionary of the Kazakh language in general education]. Almaty.
[28] Rakhimova, D. R., Kassymova, D. T., & Isabaeva, D. N. (2021). Qazaq tiline arnalgan BERT modeli negizinde suraq-jauap juyesin zertteu jane azirleu [Research and development of a question-and-answer system based on the BERT model for the Kazakh language]. *Bulletin of the Abai KazNPU. Physical and mathematical sciences*, *4*(76), 119-127. https://doi.org/10.51889/2021-4.1728-7901.16
[29] Ryabova, M. Yu., & Sergeichik, T. S. (2019). Word of the Year as a Cultural Concept in Media Discourse. In Z. Anikina (Ed.) *The Abel Prize* (pp. 325–332). Springer Nature. https://doi.org/10.1007/978-3-030-11473-2
[30] Stepanov, Y. S. (2007). *Koncepty. Tonkaya Plyonka Civilizacii* [Concepts. A Thin Tape of Civilization]. Yazyki slavyanskikh kul’tur.
[31] Stubbs, M. (2001). *Words and Phrases: Corpus Studies of Lexical Semantics*. Blackwell.
[32] *Sub-Corpus of the National Corpus of the Kazakh Language*. (n. d.). Retrieved May 29, 2024, from https://qazcorpora.kz/.
[33] Zimmer, B. (n. d.). *A brief history of the Word of the Year*. Oxford Languages Blog. Retrieved May 15, 2024, from https://languages.oup.com/word-of-the-year/word-of-the-year-a-brief-history/.

Assel B. Ormanova, PhD, is currently an Assistant Professor in the Department of General Education Disciplines at Astana IT University in Kazakhstan. She has published her works in peer-reviewed international journals. She is also the co-author of textbooks in English. Her research focuses on corpus linguistics, bilingualism, terminology, specialized language, and first and second language acquisition. ORCID ID 0000-0003-3265-6111. Email: assel.ormanova@yandex.kz

Madina L. Anafinova, Candidate of Philological Sciences, is currently an Associate Professor at the Pedagogical Institute of Astana International University in Kazakhstan. She has published her research in peer-reviewed international journals. She is also the author of several books in English and a monograph published in 2024. Her research interests include lexicology, issues in terminology development, ESP, and sociolinguistics. She had a scientific internship at the University of Cambridge in the UK in 2013. ORCID ID 0000-0002-8523-0010. Email: mad-anafinova@mail.ru

Dana Zh. Ospanova, PhD, currently serves as the Head of the Terminology Department at the National Scientific and Practical Center "Til-Qazyna" named after Shaisultan Shayakhmetov in Kazakhstan. She has published her research in peer-reviewed international journals. She is also the author of books, vocabularies and a monograph published in 2024. Her research interests include corpus linguistics, emotiology, psycholinguistics, and linguodidactics. ORCID ID 0000-0001-8901-8060. Email: Dana.zhanabek@mail.ru

Zhibek K. Tleshova, candidate of Pedagogical Sciences, is currently an Associate Professor and Chair of the Department of General Education Disciplines at Astana IT University in Kazakhstan. She has published her research in peer-reviewed international journals. She had a scientific internship at the University of Cambridge in UK in 2012. She had grants from the Embassies of the United Kingdom, the United States and Erasmus + JM. She is also the author of academic publications and a monograph. Her research interests include intercultural communication, second language acquisition, and EFL. ORCID ID 0000-0001-5095-5436. Email: zhibek.tleshova@astanait.edu.kz