







ARTICLE

## The Creation of Concordance as an Effective Tool for Studying the Text: On the Example of A. Baitursynov's Concordance

Gulnur Baishukurova <sup>1</sup> , Akerke Irgebayeva <sup>1</sup> , Nurlykhan Aitova <sup>2\*</sup> , Dariga Kapassova <sup>3</sup> ,  
Samal Serikova <sup>1</sup> , Dana Ospanova <sup>4</sup> 

<sup>1</sup>Department of Russian Language and Literature, Abai Kazakh National Pedagogical University, Dostyk Avenue, 13, Almaty 050010, Kazakhstan

<sup>2</sup>Department of Kazakh Linguistics, L.N.Gumilyov Eurasian National University, Kazhymukan Street, 11, Astana 010000, Kazakhstan

<sup>3</sup>Department of Humanity Disciplines Nur-Mubarak University, Kazakhstan, Al-Farabi Avenue, 73, Almaty 050040, Kazakhstan

<sup>4</sup>National Scientific and Practical Center "Til-Qazyna" Named after Shaisultan Shayakhmetov, Sauran Street, 7, Astana 010000, Kazakhstan

### ABSTRACT

The active use of modern computer technologies in philology has led to the intensive development of corpus linguistics. A promising direction in this area is the development of national language corpora and the construction of concordances of various types, differing in the forms of presentation of the material, search capabilities, and technical functions. This research presents basic information about the national language corpora and concordances currently being developed in Kazakhstan. We note that the level of development of corpus linguistics in Kazakhstan differs from the level of development of language corpora in developed countries. Therefore, we point out objective factors that cause an insufficiently high rate of development of Kazakhstani dictionaries in electronic form, especially alphabetical-frequency concordances to the author's texts. This research is a part of our project to develop a Kazakh-Russian parallel corpus of Akhmet Baitursynov,

#### \*CORRESPONDING AUTHOR:

Nurlykhan Aitova, Department of Kazakh Linguistics, L.N.Gumilyov Eurasian National University, Kazhymukan Street, 11, Astana 010000, Kazakhstan; Email: [nurlykhan.an@gmail.com](mailto:nurlykhan.an@gmail.com)

#### ARTICLE INFO

Received: 11 July 2024 | Revised: 3 August 2024 | Accepted: 14 August 2024 | Published Online: 23 October 2024  
DOI: <https://doi.org/10.30564/fls.v6i5.6856>

#### CITATION

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>

#### COPYRIGHT

Copyright © 2024 by the author(s). Published by Bilingual Publishing Co. This is an open access article under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License (<https://creativecommons.org/licenses/by-nc/4.0/>).

which has about 3,500 words in usage. The article analyzes the structure of concordance and presents the stages, tasks of development, and functional description of the first concordance in Kazakh lexicography by Akhmet Baitursynov, a reformer and founder of Kazakh orthography, a Turkologist, author of the first textbooks. The article analyzes different approaches of scientists to the issue of systematization and typological characteristics of known concordances. Based on this analysis, we have proposed their generalized classification. The developed version of the concordance of A. Baitursynov's language is the foundation for creating a variety of dictionaries; it can become a precedent for translating Kazakh language to a digital platform.

**Keywords:** Concordance; Dictionary; Lexicography; Corpus of Language; Corpus Linguistics; A. Baitursynov

## 1. Introduction

Modern linguistics has significantly expanded and strengthened its toolkit with computer technologies in the development and use of dictionaries; therefore, researchers have the opportunity to analyze more complex linguistic phenomena. The use of the latest IT technologies opens huge opportunities in the study of lexical, grammatical, and syntactical language tools: quick search, selection, automated sorting and grouping of vocabulary, interactive dictionary browsing, instant click-through from the dictionary to the corpus of texts, provision of information in various formats. These possibilities are implemented in corpus linguistics, in particular, the construction of concordances. "A concordance is a collection of the occurrences of a word form, each in its textual environment. In its simplest form, it is an index. Each word form is indexed and a reference is given to the place of occurrence in a text"<sup>[1]</sup>. Since all the cases of the use of the word are recorded in concordance, it is a convenient lexicographic form reflecting the individual characteristics of the style in a particular work or individual author.

The possibility of automated creation of electronic dictionaries is a developing area of lexicographic practice. Different types of dictionaries are obtained from the same dictionary base: a frequency dictionary, a reverse alphabetical dictionary, a grammatical dictionary, and dictionaries of specific works by different projection and sorting. In addition, the researcher will always be able to get the desired dictionary projection on request.

The major priority in creating a language corpus is improving the quality of linguistic research and its empirical base by forming extensive arrays of texts. It is known that the largest national language buildings have been created in developed countries with advanced information technolo-

gies, such as Japan, Great Britain, Finland, etc. We have to admit that Kazakhstan is currently slightly inferior to the world in terms of high-tech development. In addition, Kazakhstan is characterized by a multilingual language area, and the Kazakh language has acquired the status of the state language relatively recently. In this regard, the experience of building a Kazakh language corpus differs significantly from the experience of developing language corpora in developed countries. At the same time, it should be noted that there is a lack of specialists-developers of corpus linguistics and IT specialists who speak the Kazakh language. These factors determine the insufficiently high level of development of corpus linguistics in Kazakhstan, including the development of electronic forms of dictionaries, especially alphabetical frequency concordances to the author's texts. The availability of electronic dictionaries will contribute to the improvement of the national corpus of the Kazakh language.

The development of corpus linguistics in Kazakhstan goes in two directions: 1) development of the national corpus of the Kazakh language and its sub corpus, 2) compilation of dictionaries of the language and works by individual authors, mainly explanatory and frequent. The second direction, namely, the study of the text array and the creation of concordances in the writer's (author's) lexicography, is becoming particularly relevant.

The purpose of the article is to present the stages of development of the first author's concordance in Kazakh linguistics, the tasks of creation, functional description, and advantages of Akhmet Baitursynov's concordance as an electronic dictionary of commenting and searching type<sup>[2]</sup>. This study is a part of our project to develop a Kazakh-Russian parallel corpus of A. Baitursynov<sup>[2]</sup>. As part of the project, we turn to the parallel corpus as a tool for proper semantic research. The concordance is intended to reflect all possi-

ble differences and peculiarities of word usage because it is not known in advance which of them may turn out to be significant and in demand. Concordance allows to analysis of not one, but several texts or corpus of electronic texts at once and provides information about the context of the use of language units. While comparing the original text and the translation, concordance is an indispensable tool for identifying more complex forms of requesting the desired units (for example, excluding certain words from the search), displaying information on the frequency of use of language units and their compatibility in a given corpus of texts; providing the opportunity to refer to a specific text in which an example was found; to offer various options for the output of information (in the form of completed sentences or with the omission of the desired lexical unit), etc.

Concordance based on bilingual parallel corpora as a scientific problem will be developed in our country for the first time. The project is promising because the development of concordance using a fundamentally new methodology and technology of preparation meets the modern level of philological science. The results of the research will undoubtedly be useful in translation activities, lexicography, teaching and learning language, analyzing the text, and creating a National corpus of Kazakh language.

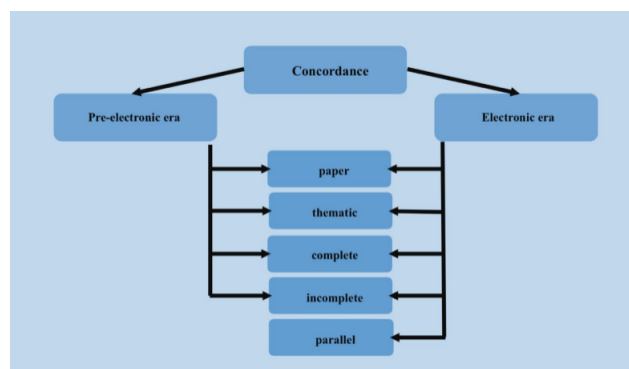
The development of A. Baitursynov’s concordance is primarily due to the need for scientific analysis of the vocabulary and grammar of the author’s language, as well as historical changes that took place in the language during the formation of Kazakh script. Various graphic systems have been used in the history of Kazakh writing: Runic, Arabic, Latin, and Cyrillic. After converting to Islam, for a long time, Turkic-speaking people used Arabic graphics. Therefore, most of the manuscripts of A. Baitursynov came down to us with different versions of Arabic graphics (*mөme жазу*) [*tóte jazý*] following the spelling norms of that time. In this regard, the issue of translating the texts of A. Baitursynov’s works from Arabic graphics into Cyrillic were acute. The basis of the concordance created was the texts of the scientists translated into Cyrillic.

The selected works of A. Baitursynov was used as the main textual source for the creation of the concordance, which included parts from fundamental scientific works and tutorials in the field of linguistics, literary studies, such as “*Оқу құралы*” [*“Oqý quraly”*], “*Тіл-құрал*” [*“Til-qural”*],

“*Әдебиет танытқыш*” [*“Ádebiet tanytqysh”*], selected articles, letters, reports, and speeches, as well as poems, translations of the poet. Currently, there are academic editions of A. Baitursynov’s works in three, six, and twelve volumes<sup>[3]</sup>. They present the works of the scientist in different graphical projections. The fact is that in some editions of his works (from 1914 to 1928), A. Baitursynov used different versions of Arabic graphics, which, in turn, led to a heterogeneity of the array of texts. Moreover, most of the scientist’s works have been edited and transliterated into Cyrillic several times. Thus, the heterogeneity of publications is caused by the author’s editorial variability of graphics. We did not set out to bring these graphics options into unity since this task is clearly beyond the scope of our study.

## 2. Literature Review

A review of the literature on theoretical problems of corpus linguistics has shown the heterogeneity of the criteria for classifying concordance. Some researchers note that concordances can be bilingual (these are concordances based on parallel texts) and thematic (these are lists of topics that the book covers, with the content of the essence of these topics), complete (which gives a list of all the words of a text indicating all the contexts of their use) and incomplete (when the dictionary is differentiated, and quoting contexts is selective). Having considered the existing classifications of concordances, we have generalized and systematized the types of concordances according to the chronological principle (**Figure 1**).



**Figure 1.** The classification of concordances.

The turning point in the differentiation of periods is the emergence of computer technology. According to this principle, concordances before the electronic and electronic

era are distinguished<sup>[4]</sup>.

The first concordances refer to the pre-electronic era, they were paper, thematic, and incomplete. It is known that the first concordance called “Concordantiae Morales” was compiled in the XIII century by the monk Anthony of Padua to the Latin version of the fifth century Bible “Vulgate”. At the same time, under the guidance of Hugo de Saint-Cher, an alphabetical index of words in the Vulgate Bible was compiled<sup>[5]</sup>.

In the pre-electronic era, the prerequisites for the transition to electronic-era buildings were created. The first concordances were considered as dictionaries and indexes because they referred from the word to the context of its use in the text. Electronic and paper dictionaries differ in the form of information presentation and the use of technical means. This difference leads to many others at the level of usage, presentation, content, search capabilities, and functions performed.

The starting point of the electronic era of concordance creation is considered to be the period of experiments on the development of algorithms, machine translation of texts, and, in general, natural language processing (1960–1970)<sup>[6]</sup>. Unlike concordances of the pre-electronic era, modern concordances are automatically developed using concordances. Concordances are computer programs and tools for building frequency lists of words, searching for lemmas, and word forms, and extracting keywords and terms. The concordancer is an authentic academic English language search tool that allows the user to learn vocabulary usage, frequency of use, phrases, and more<sup>[7]</sup>.

Famous concordances of the twentieth century are concordances to individual works, for example, to the Bible, the works of Chaucer, Shakespeare, Byron, Faulkner, O’Neill, Fitzgerald, etc. In Russian lexicography, the first experiments of concordance are dictionaries for poems by G.R. Derzhavin, works and translations by D.I. Fonvizin, author’s dictionaries by M. Lermontov, L.N. Tolstoy, F.M. Dostoevsky. The most complete and authoritative concordances are: “Dictionary of Pushkin’s Language”, concordance to the texts of M.V. Lomonosov<sup>[8]</sup>, “Dictionary-concordance of Publicity by F.M. Dostoevsky”, concordance of Osip Mandelstam’s poems, concordance of A.S. Aksakov’s works, etc.

The most famous corpora are British National Corpus, Bank of English (BoE), American National Corpus (OANC),

Finland Language Bank (Kielipankki), National Spanish Language Corpus (CORDE, CREA), National Bulgarian Language Corpus (BulNC), Czech National Corpus (ČNK), Turkish National Corpus (TNC), National Corpus of Russian Language (NCRL), etc.

The thematic concordance, as it was mentioned earlier, existed in the pre-electronic era. Examples of thematic concordances of the electronic era are Nave’s Topical Bible Concordance Online by Orville J. Nave, Authorized King James Bible, online dictionary of King James Version words (Online dictionary of King James Version words KJV Dictionary), etc.

Bilingual concordances (bitexts) appeared in the 80s and 90s of twentieth century<sup>[9, 10]</sup>. Nowadays, bilingual concordances have been developed for such languages as English, French, German, Czech, Greek, etc. Bilingual corpora based on the original texts and their translations are called parallel corpora. For example, there is a parallel corpus of texts of meetings of the Canadian Parliament (English/French). Further development of parallel buildings led to the creation of a multilingual corpus (Acquis Communautaire database of European legislation in 22 languages, MeLLange multilingual educational corpus, parallel concordance in Birmingham<sup>[11]</sup>). Thus, modern concordances are presented in all forms and forms: they can be paper (book) and electronic, complete or incomplete, thematic, multilingual (parallel). The types of pre-electronic and electronic time concordances considered in our study have common basic elements. However, they differ in the forms of presentation of the material, search capabilities, and technical functions. Concordance has wide research capabilities and special advantages compared to other forms of author’s dictionaries. Toldova notes that concordances are used to solve the following linguistic problems<sup>[12]</sup>:

- comparison of different possibilities of using the same word;
- keyword analysis;
- analysis of the frequency of words and phrases;
- search and research of phrases and idioms;
- search for translations, for example, terminology;
- create word lists (used in publishing).

The large-scale development of corpus research and the dynamics and success of the developing concordances are explained by the fact that they have become a powerful

information resource for use in various fields: lexicography, translation activities, methods of teaching language and literature, etc. It is a source of ready-made illustrative material; a base of modern lexicography; a tool for solving linguistic problems (building lists of words for various purposes; identifying and analyzing keywords; analyzing the frequency of words and phrases; comparative analysis of lexicons by different authors; the originality of the writer's individual author's style; identification of stable structures of various types).

There is no doubt that concordance has many advantages in comparison with traditional dictionaries: a word is analyzed based on context, which not only illustrates word usage, but shows its lexical environment sufficient to understand the meaning of the word, and also conveys the transformative effect of other lexical units on it. Unlike the dictionary, which focuses on the dictionary entry, concordance provides examples – contexts of word usage; concordances do not imply the establishment of the structure of the meanings of the registered words and do not necessarily include the interpretation of these meanings. If the dictionary is focused on the principle of representativeness, normativity, and invariance, then concordance is based on the principle of variability, on an exhaustive description.

The concordance is particularly important for the arrangement of contexts in accordance with their chronology. This principle of presentation of the material allows us to consider each context against the background of the general evolution of the vocabulary of the language and the dynamics of the use of a particular word in a single individual style.

Thus, concordance provides wide opportunities for conducting linguistic research in a real context; for solving various linguistic problems; and for preparing various historical and modern multilingual dictionaries. The concordance data is the basis for compiling grammar to learn the language, as well as for clarifying some of its aspects.

### 3. Methods

During the working process on A. Baitursynov's concordance, the main methods of text corpus analysis were applied, which significantly improved the study of corpus data both quantitatively and qualitatively. The method of computer text analysis allows you to establish linguistic pat-

terns based on empirical data. Due to the system of computer selection of a text array, it became possible to carry out, process, and display the results of text data, as well as search and select words and word forms, calculate frequency and compile statistics, and analyze grammatical markup. In this way, it is possible to determine how and in which contexts a certain word or grammatical form is used, with which frequency, and in which works. One of the main advantages of corpus linguistics is the possibility of conducting objective statistical analysis based on accurate empirical data, rather than on assumptions and hypotheses. In the process of analyzing the word form, the morphological descriptor plays a role in determining the grammatical categories of the desired word. Concordance is an up-to-date and productive way of studying the text, which performs the most important functions. Researchers have repeatedly drawn attention to the conditionality of the functional orientation of any dictionary edition by the basic principles of lexicography and user queries. Concordance performs various interrelated, interdependent functions that "constitute not just a sum, but a single organizational system of lexicographic modeling of the language"<sup>[13]</sup>. The available functions of corpus text analysis range from searching and compiling contexts to calculating the frequency of use and creating statistics. In addition, the analysis of the meta-markup of the text and the possibility of creating your subcorpus of texts are included. In studying the text, the search, heuristic, and analytical functions of concordance, as well as indexing and comparison functions are highlighted.

1. The search function allows you to quickly find the desired text fragment using a given word or phrase<sup>[14]</sup>. This concordance function is used in the selection of quotations and their reconciliation, when studying the features of the original text, and also allows you to search for phrases and idioms in the text, and search for terms and idioms in translation memory. The search for the desired word form can be performed in various contexts, depending on the specified parameters. In the practical aspect of research, this function is used to collect factual material and create a theoretical basis for research, such as dictionaries, grammar, and reference manuals. In addition, the concordance function is an ordered list of word forms indicating all occurrences in a given array of texts, which allows you to study the use of a language unit in a fixed-

length context and analyze the combination of the desired word forms;

2. The heuristic function of concordance distinguishes it from nominal pointers by the presence of context. Often, contexts allow a linguist or a usual reader to see a new interpretation of the text. This function involves solving several tasks related to various aspects of the study of the word: the specifics of the lexical composition of texts of various genres, the semantics of the word, its connections, the contextual environment of the word, the features of syntactic constructions, the identification of thematic groups of vocabulary, comparative observations, and other linguistic phenomena;
3. The analytical function allows you to analyze various linguistic indicators, such as lexemes, keywords, the frequency of their use in the text, etc. The analysis of word forms, grammatical categories, and phrases using a morphological descriptor allows you to conduct a unit study based on the grammatical categories of the word;
4. The indexing function makes it possible to create indexes and word lists when preparing a text for publication. The purpose of this function is to speed up the process of data extraction, modification, and sorting;
5. The comparison function is used when comparing various connotations and uses of a word in the text, as well as keywords of one corpus with another.

A. Baitursynov's concordance, which we are developing, performs two main functions – reference and heuristic. These functions enable the user to search for a word or word form on request with a quick transition to the necessary dictionary information. At the same time, the user receives additional linguistic information via the link. So, concordance is a highly informative form of dictionary description.

### 3.1. Theoretical Framework

In recent decades, the efforts of linguists in many countries have been aimed at creating national, or universal, integral corpora. Although the criteria for the representativeness of such a corpus are not yet completely clear: The corpus must have quantitative and qualitative parameters necessary and sufficient to build an adequate dictionary and grammar of the corresponding language on its basis. At present, Kazakh lexicography has been enriched by paper and electronic dic-

tionaries for individual works by one or another author, for a cycle of works or the entire work, for example, the corpus of selected works of Abai Kunanbayev, Mukhtar Auezov, Abish Kekilbayev, Mukhtar Magauin, Mukagali Makatayev has been created<sup>[15]</sup>. (The staff of the Institute of Linguistics of the Academy of Sciences of the Republic of Kazakhstan has developed the “Dictionary of Abai language”. The lexical fund of the dictionary of outstanding Kazakh poets and philosophers amounted to 6,030 words, and the total number of words in the texts was 50,000. The dictionary reflects the features of the Kazakh literary language of the 2nd half of the XIX century, in particular, the system of word formation and word usage, lexical and grammatical norms, stylistic means, and techniques. The “Dictionary of Abai language” became the first historical, literary, and explanatory dictionary of the writer's language. Moreover, it is a source for studying the history of the modern Kazakh literary language. Along with this, based on the above-mentioned Institute, within the framework of the National Corpus of Kazakh Language, the development of A. Baitursynov subcorpus began, the purpose of which is to highlight various aspects of his popular science, educational, and scientific works; to create an electronic database for studying the history of the development of the literary Kazakh language, ways to replenish the vocabulary and analyze functional styles. His works such as “Оқу құралы” [*Oqú quraly*], “Тіл-құрал” [*Til-qural*], “Тіл жұмсар” [*Til jumsar*], “Баяншы” [*Baianshy*], “Әлiнне” [*Álippe*], “Әлiнне астары” [*Álippe astray*], “Саят ашқыш” [*Sayát ashqysh*], “Әдебиет танытқыш” [*Ádebiet tanytqysh*], “Маса” [*Masa*], “Қырық мысал” [*Qyruq mysal*] and others served as material for the subcorpus. And in the historical subcorpus of NCRL of A. Baitursynov's work, the transliteration of the scientist's texts from Arabic graphics to Cyrillic graphics is presented. In 2019, the “Dictionary of Mukagali Language” was published, based on a five-volume collection of works of Kazakh poet Mukagali Makatayev and includes 13,348 words and 5,035 stable phrases<sup>[16]</sup>. Definitions are given for each of the linguistic units and examples of their use in the context of the poet's works are given. The authors of the dictionary define this dictionary as a consolidated (synthesized) dictionary consisting of explanatory and frequency dictionaries. The publication is supplemented with 15 appendices, which provide statistical data on the use of grammati-

cal forms and categories, pronouns, nouns, etc. The concept of the corpus was introduced into Kazakh linguistics<sup>[17]</sup>. In his opinion, the corpus is a large array of texts of all styles of Kazakh language, collected in an electronic database and intended for mass use through a computer program<sup>[18]</sup>. Work on the development of the National Corpus of Kazakh Language (NCKL) began at the Institute of Linguistics of the Academy of Sciences of the Republic of Kazakhstan in 2009 under the guidance of Professor Zhubanov. This corpus covers the lexical composition of all styles of the modern Kazakh language. It includes the following subcorpus: basic, oral, historical, dialectological, culturally representative, parallel, anosmatic, subcorpus of the advertising language, and subcorpus of A. Baitursynov. An alternative to NCKL is Kazakh language corpora (see links in **Table 1**), developed based on universities and research centers.

A variant of the National Corpus of Kazakh language is the Almaty Corpus of Kazakh Language (ACKL), a reference and information system based on an array of texts of the literary Kazakh language. A subcorpus of the National Corpus of Kazakh Language (SNCKL) is being developed in parallel with the above-mentioned corpora. The corpus is publicly available for a wide range of users who can use 4 types of searches: actual search, morphological search, parameter search, and distance search. The corpus indicates the frequency of the most common words, the forms of words, their semantics, word formation, and morphological features.

As we regard it, corpus research in Kazakh science has received a powerful impulse and is currently developing intensively. The existing dictionaries of the language and works of Kazakh writers are presented mainly in the form of explanatory and frequent ones, the dictionary entries of which traditionally include an interpretation and an exhaustive description of all the meanings of the word with a focus on normativity, semantic, and grammatical analysis. In contrast, the is being developing A. Baitursynov's concordance is aimed at creating a list of all cases of word usage regarding minimal contexts, identifying frequency and compatibility, and analyzing the evolution of an individual author's style. It follows from the above that in Kazakh literary lexicography, not only concordances of an individual author are in great demand, but also thematic, bi-, and multilingual ones that perform various functions of corpus analysis of the text.

## 4. Result and Discussion

A. Baitursynov's concordance is part of a project to create a Kazakh-Russian parallel corpus<sup>[2]</sup>, which will provide a wide range of users with a software and information environment for researching the literary and scientific heritage, language, and biography of Kazakh educators. The Concordance is based on a new methodology and technology of training that meets the modern level of philological science. It is based on a corpus of philologically verified author's texts, equipped with rich structural, philological, and grammatical markup. The project provides the creation of an open Internet resource, which will include:

1. the corpus of texts of A. Baitursynov, built based on his scientific, journalistic, and artistic works and translations;
2. alphabetical index of works in Kazakh and Russian languages;
3. a list of all the words and word forms found in the selected works of A. Baitursynov;
4. Kazakh-Russian dictionary of terms used in the works of A. Baitursynov.

### 4.1. The First Stage of A. Baitursynov's Concordance Development

The development of the concordance took place in several stages. The first stage is preliminary. During the initial processing of the text for subsequent processing by a special automated program, problems with additional information were solved, and material not directly related to the author's texts was excluded: the output data of the collection, notes, footnotes, remarks, etc. An alphabetical index of selected works by A. Baitursynov was compiled. This edition was used as a method to create a concordance between the authors. The choice of the publication was not accidental and was because the authors of this article, G. Baishukurova, and A. Irgebayeva, were translators from Kazakh into Russian, N. Aitova was a responsible editor and compiler of these works. "Selected works" of Akhmet Baitursynov are available in the electronic library on the website "Til alemi" ["Til alemi"]<sup>[3]</sup>. All the works included in the collection are given in accordance with the original, the spelling and style of the author are preserved, and therefore the collection is academic. At the moment, the volume of the concordance is about 3,500 words and word usage, which describe the words found in

**Table 1.** Kazakh language corpora.

Name of Corpus	Authors	Creation	The Number of Words Used	Composition and Structure	Link
National Corpus of Kazakh Language (NCKL)	A. Baitursynov Institute of Linguistics	2009– to present	There are about 40 million words in usage.	Near 31 million texts of different styles, has several subcorpuses	https://qazcorpus.kz
Almaty Kazakh Language Corpus (AKLC)	Al-Farabi Kazakh National University	2012– to present	There are about 40 million words in usage.	More than 100 classic works of Kazakh literature.	http://web-corpora.net/KazakhCorpus/search/?interface_language=ru
Subcorpus of National Corpus of Kazakh language	National Scientific and Practical Center “Til-Kazyna” named after Sh. Shayakhmetova	2021– to present	more than 23 million words usage.	near 15 million texts	https://new.qazcorpora.kz/

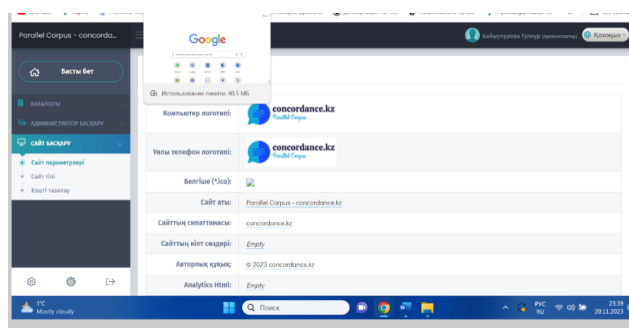
the works of A. Baitursynov. An important part of this stage is the analysis of the existing basic principles of concordance construction. These include:

1. The principle of representativeness (scaling). This principle implies a sufficient and proportional representation in the corpus of texts of various periods, genres, styles, etc.<sup>[19]</sup>. It is the representativeness of the corpus that determines the reliability of the results obtained on its material. Thus, representativeness can be considered as a problem of adequate reflection, adaptation, or integration of large arrays of texts or some other fragments of speech activity into a significantly smaller corpus of texts<sup>[20]</sup>;
2. The principle of balance. Balance is understood as a proportional representation in the corpus of texts of various periods, genres, styles, authors, etc. The achievement of this parameter is assumed by including texts with different genres and style affiliation: scientific, methodological, publicist, and artistic;
3. The principle of accessibility. To implement this principle, public access to the concordance online is required;
4. The principle of openness and continuous replenishment. The essence of this principle is that at any stage of concordance creation, the text array can be improved, supplemented, and edited in terms of volume, composition, and in terms of solving linguistic problems.

#### 4.2. The Second Stage of A. Baitursynov’s Concordance Development

In the second stage, the task was to determine the prin-

ciples of the corpus structure, to set the parameters for working with linguistic data (query to the corpus, output, sorting of results, analysis of quantitative metrics, distribution by year) to form a clear idea of the structure and design of the site. At this stage, work was carried out to create the main component of the concordance—the electronic corpus of the dictionary—an alphabetical list of words and word forms indicating the frequency of use. The alphabetical list was created automatically by the program, and access to it is possible only through the same program. First, a working version of the site was developed, which is presented in **Figure 2**.



**Figure 2.** Window 1: The main page of the original version of A. Baitursynov’s concordance.

The following options are listed in the site menu: catalog, administrative management, and site management. The catalog includes about 3,500 words and word forms, a list of words, a list of sentences, and books. Administrative



management indicates the status and terms of reference of developers (administrator, editor-in-chief, editor). Site management allows you to define parameters such as site settings, site language, and so on.

The list of words contains word forms in alphabetical order, indicating their frequency and place of use in sources. The list of sentences is a field for “manually” entering contexts of word usage from the writer’s works with the possibility of subsequent editing and addition (Figure 3).

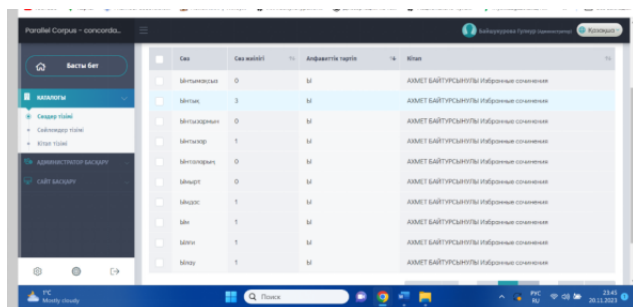


Figure 3. Window 2: List of words, sentences, list of books.

The list of books is an alphabetical index of the works of A. Baitursynov.

### 4.3. The Third Stage of A. Baitursynov’s Concordance Development

In the third stage, linguistic processing of the scientist’s electronic texts was performed. The created alphabetical list of words and word usage by A. Baitursynov is a completely reliable research tool that gives a sufficient idea of the vocabulary of the author’s works. Every word and every word form is consistent with the author’s megatext, in which they are available. Search and display of search results on the page have been added to the main page of the site by lemmatization. The works are displayed in alphabetical order. The main page of the site was developed using open and accessible BOOTSTRAP5 templates, HTML, CSS, JAVASCRIPT, and JQUERY, which allow you to work with the site the same way on your computer and your phone. Later, the design and interface of the site were updated and launched on the server in a publicly accessible operating mode under the domain name baitursynuly-corp.kz (Figure 4).

The frequency of words was calculated automatically. At the request of the desired language units, information about the frequency of their use is displayed on the screen. Here are examples of the most frequent lem-

mas in A. Baitursynov’s concordance: адам [adam] (person) (140), бастауыш [bastaýysh] (primary) (108), тіл [til] (language) (104), аз [az] (a little) (99), өз [óz] (yourself) (96), бағыныңқы [baǵynyńqy] (subordinate) (74) and others (Figure 5).



Figure 4. Window 3: The main page of the updated interface.

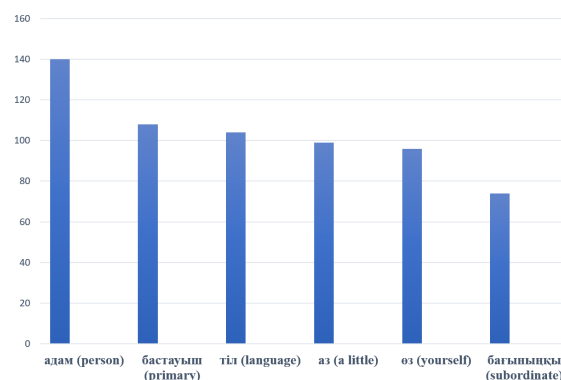


Figure 5. Frequency demonstration using the example of “adam” word (person).

In addition, all cases of word usage in different works of the author are displayed. In Figure 6 we use the example of the word “*bastay*” (to start): *bastaýysh*, *bastaýyshqa*, *bastaýyshpen*, *bastaýyshhtar*, *bastaýyshardyń*, *bastaýyshty*, *bastaýyshtyń*.

The structure of the dictionary entry of our concordance is:

1. Title word – word usage;
2. Second item grammatical marks (part of speech, rank, number, mood);
3. Frequency in contexts;
4. context with reference to the pages of the work

Here is an example of the use of the word “*бастай*” [“*bastay*”] (to start) in concordance:

*Бастай 2. Әдістің затына соңғы мағынада айтыулары дұрыс келеді. Үйткені – бұл әдісті*

қолданыушылар оқыу үйретіуді оқыудан бастау дұрыс емес, жазыудан бастау дұрыс дейді <2, 212> [Ádistiñ zatyna soñǵy maǵýnada aıyýlary durys keledi. Útkeni – bul ádisti qoldanıúshylar oqúú úiretıúdi oqúýdan bastaú durys emes, jazýdan bastaú durys deidi] // The last statement corresponds to the essence of the method, since those using this method assure it is better to start learning by writing, rather than starting with reading <2, 219> (Figure 7).

#	Сөз	Сөз жиілігі
301	Бастапқысы	1
302	Бастар	5
303	Бастарын	1
304	Бастаса	2
305	Бастау	2
306	Бастауыш	108
307	Бастаушы	1
308	Бастаушымен	1
309	Бастаушытыр	4
310	Бастаушыларын	1
311	Бастаушыты	1
312	Бастаушытың	3

Figure 6. Window 4 Lemma of the word “*bastay*” and its word forms.

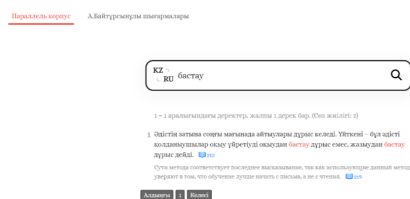


Figure 7. Window 5: An example of the contextual use of the word “*bastay*” [“*bastay*”] (to start) with a reference to the pages of the work.

From the above example, we can extract the following information: the word “*bastay*” [“*bastay*”] is low-frequency, used only in this context twice in the meaning of ‘to begin an action’, expressed by the main verb in the infinitive form. In the context, the syntactical function of this word is the subject with which the predicate agrees: the first grammatical basis is “*bastay durys emes*” [“*bastay durys emes*”], the second is “*bastay durys*” [“*bastay durys*”]. In combination with the additions “*oquдан*” [“*oquдан*”] and “*jazудан*” [“*jazудан*”], the syntagmatic connections of the analyzed word “*bastay*” [“*bastay*”] with the help of its repetition emphasize the author’s idea of the methodological importance of starting learning not from reading, but from writing. Now there is an example of a dictionary entry of

the high-frequency word “*адам*” [“*adam*”] (person): *Адам 140. Әріп қосып ежікпен үйренген адам оқыуды шапшаң жүргізіп оқығалмаған* [Árip qosyp ejikpen úirengen adam oqúdy shapshañ júrgizip oqıyalmaǵan]<140, 205> A person who learned to write by adding letters could not learn to read quickly <140, 213>. The word “*адам*” [“*adam*”] is a subject that agrees with the compound predicate “*оқығалмаған*” [“*oqıyalmaǵan*”], used in the meaning of ‘any person’. This word can be used in another syntactic function. For example,

*Адам 140. Ағаш көркі жапырақ, адам көркі шүберек* [Aǵash kórki japyraq, adam kórki shúberek]<140, 54> //A tree is decorated with leaves, a person is decorated with clothes <140, 58>.

Here we can see that the word “*адам*” [“*adam*”] is a complement that is consistent with the subject “*көркі*” [“*kórki*”] (adorns). According to the grammatical rules of the Kazakh language, the complement in the Genitive case (Ілік септік [Ilik septik]) answers such questions as *кімнің?* [kimniñ?] (who) and *ненің?* [neniñ?] (what) and is formed with the help of endings *-ның/-нің, -дың/-дің, -тың/-тің* [-nyñ/-niñ, -dyñ/-diñ, -tyñ/-tiñ]. In this example, the word “*адам*” [“*adam*”] is given in a reduced form without the ending *-ның* [-nyñ]. This use of the word “*адам*” [“*adam*”] is due to the excessive meaning of possessiveness in the phrase “*адам-ның көрк-і*” [“*adam-nyñ kórk-i*”]. The reduced possessive form “*адам*” [“*adam*”] coincides with the initial form of the word “*адам*” [“*adam*”] and causes homonymy thereby increasing the frequency index of the word.

Homonymy can also be explained by the fact that the word “*адам*” [“*adam*”] is used in the singular form but expresses the meaning of plurality. For example,

*Адам [“Adam”] (140). Хамза, Хамид, Ахмад, Салих – адам аттары.* Khamza [Hamza, Hamid, Ahmad, Salih-adam attary. Hamza]<140, 36> // Hamza, Khamid, Akhmad, Salih are people’s names <140, 39>.

The syntactic role of the word “*адам*” in this case is a compound predicate. This use of the word “*адам*” [“*adam*”] is due to the excessive meaning of multiplicity *-дар/-тар* [-dar/-tar] and possessiveness *-дың/-ы* [-dyñ/-y] in the phrase “*адам- дар-дың ат-тар-ы*” [“*adam - dar-dyñ at-tar-y*”].

Thus, at this stage, concordance helps the researcher to determine the contextual meaning of each word, its syntactic role, and syntagmatic connections. In general, concordance provides an opportunity to identify grammatical and typo-

logical features of the language and to explore the scientific and theoretical views of the scientist.

#### 4.4. The Fourth Stage of A. Baitursynov’s Concordance Development

At the fourth stage, a text markup format for concordance was developed. An electronic text consists, on the one hand, of linguistic elements of various levels (words, phrases, sentences) and, on the other hand, of structural segments of various types (headings, footnotes, remarks, poems, quotations, tables, pages). Markup is the placement of special markers (tags) in the text of a document that explicates the “hidden” elements of information present in the text<sup>[21]</sup>. The key properties of the format are openness, compactness, and replenishment. These properties allow you to combine all actions into a single algorithm, with the help of which, at the initial stage, untagged texts are entered, and at the final stage, a processed marked—up test array. As a result of using this algorithm, a concordance is automatically created. The text markup format was created taking into account the presence of meta-textual, structural, and linguistic information in it. Depending on the type of information, the following types of markups can be distinguished:

1. Metatext markup includes:
  - a text fragment;
  - the name of the text;
  - the genre of the text
    - date of publication;
    - source;
  - Source pages;
  - date of entry into the case;
    - Name, and surname of the person who contributed the text to the corpus;
    - Name, and surname of the person who edited the text in the corpus;
    - Notes (**Figure 8**).
2. Structural markup explicates the division of text into linguistic elements (tokens) of various types — sentences, words, punctuation marks, and numbers. Then its manual postprocessing is carried out (removal of homonymy, correction of parsing);
3. The grammatical designation indicates the necessary morphological characteristics of words or word forms. The

service window that opens via the link provides a list of morphological characteristics, divided into grammatical ones. With the help of special designations, various grammatical characteristics of the word are indicated: part of speech, number, gender, case, etc. Using these special designations, you can search for a word or search for a word with a modified form.

A	B	C	D	E	F	G	H	I
Мәтін	Мәтін атауы	Мәтін жанры	Жариялану мерзімі	Дереккөз	Мәтін беті	Корпусқа енгізу күні	Корпусқа мәтін енгізгеннің аты-жөні	Ескерту
1								
2	Сөздің оңай сарамы біту – саям біту болды	Әдебиет таптықшы	оқулық	1926	Тыңдамалы шығармалар	197	15.09.2023	Умарова А.Б.
3	Жаралы құры жомса кім зарығам! – Бұл – қызық. Төңірге зәр еткен қызық! Қызыл қызы азын шеккен қызы! Әйтеуір қызы азын күткір деп, оқпаған азын азын күткен қызы.	Тізі құрал	оқулық	1914	Тыңдамалы шығармалар	134	16.09.2023	Дребова А.Б.
4	Қара сөз көбінесе ғалымдардың, шешендердің сөз белгілемесіне қара сөз түрлерінен анық	Әдебиет таптықшы	мақала	1926	Тыңдамалы шығармалар	183	14.09.2023	Антона Н.Н.
5	Мұнда алдында сөйлемдер сөзін білгеннен кейін сөйлем жоғалып кеткен жоқ, тек	Тізі құрал	оқулық	1914	Тыңдамалы шығармалар	131	08.10.2023	Сречкова С.К.

Figure 8. Window 6: Example of metatext markup of the case.

Grammatical markup includes:

- lexeme (dictionary form);
- grammatical signs of the lexeme (part of speech, animation, discharge);
- grammatical features of the word form (number, case, declension, mood, tense, person, degrees of comparison) (**Figure 9**).

Ерге жар , **балаға** ана болмай ма қыз , Қыз жоқта қызық жоғын ойланбаймыз .

Ақан сері |

Тартынша |

Әбділда Т |

Лексема бала

Морфология 37 гла/бс

Семантика жалпы, дара, түбір, деректі

Figure 9. Window 7: Grammatical markup of the word form balağa (for a child).

Electronic concordance with a markup system opens up wide opportunities for the study of the author’s style, the evolution of scientific views of the scientist, and, in general, the dynamics of language in diachrony and synchrony. Concordance without markup in the form of a simple list of word forms is significantly inferior to structured marked concordance in terms of information content and application purpose.

Our research presents the main stages of A. Baitursynov’s concordance development (**Figure 10**). At the preliminary stage, the initial processing of the text was carried out. At the next stage, the principles of corpus arrangement (principles of representativeness, balance, accessibility, openness, and continuous replenishment), pa-

parameters of work with linguistic data, structure, and site design were determined, and an alphabetic-frequency list of words and word forms was created. At the third stage the dictionary article of A. Baitursynov's concordance includes such components as 1) word usage; 2) grammatical notes; 3) frequency of word usage; and 4) context with reference to the pages of the work. In the final stage, a format for marking up the texts for concordance was developed. Thus, all four stages of the creation of A. Baitursynov's concordance represents an integrated complex process of interrelated and interdependent technologies for creating a modern concordance.

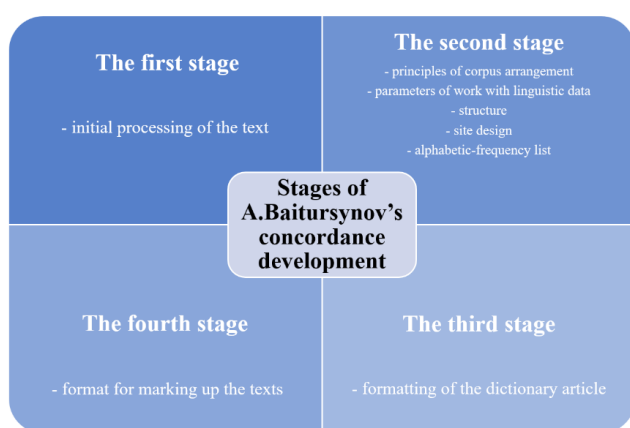


Figure 10. Stages of development of A. Baitursynov's concordance.

## 5. Conclusions

Our project of creating A. Baitursynov's concordance represents the first experience of developing an author's concordance in Kazakh linguistics. The electronic concordance is publicly available. Such studies serve as a source of illustrative material for studying the history of the development of Kazakh writing and the formation of the modern Kazakh literary language. A. Baitursynov's concordance is a valuable research tool that helps not only linguists and philologists, but also everyone interested in his work, to better understand and analyze his work. A. Baitursynov's concordance is an example of the effective use of this tool in linguistic research, and it continues to be relevant and valuable for many researchers of his work. The concordance collects and systematizes the words and expressions used in his works, which allows for a deeper study of his work and style, the peculiarities of changes in the spelling of the language, graphics, and grammatical system of Kazakh language; to study

the lexical and phraseological originality of A. Baitursynov's style; is to analyze the communicative strategies, educational materials, and teaching methods developed by him.

During the work on the project, the advantages of concordance were revealed: firstly, it is a source of ready illustrative material; secondly, it is the basis of modern lexicography; thirdly, it is a tool for solving linguistic problems (building lists of words for various purposes; identification and analysis of keywords; analysis of the frequency of words and phrases; comparative analysis of lexicons of different authors identification of stable structures of various types). To sum up, we can say that the development of concordances in Kazakhstan requires further development. The proof of this is their small number since concordances of the works of many prominent Kazakh writers, poets, scientists, and figures have not yet been created. For the intensive development of this area, it is necessary to implement several comprehensive measures: training qualified personnel in the field of corpus linguistics, creating a developed material and technical base and software, and conducting research work of philologists together with IT specialists. Thus, the concordance we are developing, which retains the usual parameters, but is created based on new approaches, has its specific features, which allows us to talk about it as an updated lexicographic form. A. Baitursynov's concordance is currently creating a platform and opportunities for conducting various types of scientific research and using it as a search and reference tool. For the next stage, it is planned to replenish the concordance with texts from the complete works of the scientist, improve the search algorithm for grammatical queries and meta-information, and create parallel corpus translations of his works into English, Russian, Turkish, and other languages on the basis of A. Baitursynov's concordance. A. Baitursynov's concordance will certainly become an integral part of the national corpus of the Kazakh language.

## Author Contributions

Conceptualization, supervision, project administration—G.B., methodology, investigation, writing and editing—A.I. and N.A., visualization and data curation—D.K., formal analysis and validation—S.S. and D.O. All authors have read and agreed to the published version of the manuscript.

## Funding

The research work was carried out within the framework of grant financing of the Ministry of Foreign Affairs of the Republic of Kazakhstan of the project «AP19676988 Concordance of A. Baitursynov. Kazakh-Russian parallel corpus» under the contract №352/23-25 from 3rd August 2023.

## Institutional Review Board Statement

Not applicable.

## Informed Consent Statement

Not applicable.

## Data Availability Statement

Access to the research materials can be accessed through these links <https://qazcorpus.kz>, [http://web-corpora.net/KazakhCorpus/search/?interface\\_language=ru](http://web-corpora.net/KazakhCorpus/search/?interface_language=ru), <https://new.qazcorpora.kz/>.

## Acknowledgments

We thank you for the grant financing from the Ministry of Foreign Affairs of the Republic of Kazakhstan.

## Conflicts of Interest

The authors declare no conflict of interest.

## References

- [1] Sinclair, J., 1991. *Corpus. Concordance and Collocation*, 2nd ed. Oxford University Press: Oxford, UK. pp. 1–179.
- [2] Baishukurova, G.J., Irgebayeva, A.B., Aitova, N., 2024. Concordance of A. Baitursynov. Kazakh-Russian parallel corpus. Almaty. Available from: <https://baitursynuly-corp.kz/> (cited 12 July 2024).
- [3] Baitursynuly, A., Baekeeva, A., Toleubaeva, A., et al., 2022. Selected works. Polygraph combinat: Astana, Kazakhstan. pp. 1–240. Available from: <https://abai.institute/assets/pdf/3247e39de00da723a31e09ae9a55bde1.pdf> (cited 12 July 2024).
- [4] Khrolenko, A.T., 2012. Automated concordance: Experience of creation and practice of use. *Philological Regionalism*. 2(8), 47–49. (in Russian). Available from: <https://cyberleninka.ru/article/n/avtomatizirovannyi-korporans-opyt-sozdaniya-i-praktika-ispolzovaniya> (cited 12 July 2024).
- [5] McCarthy, M., O’Keeffe, A., 2010. Historical perspective: What are corpora and how have they evolved? In: O’Keeffe, A., McCarthy, M. (Eds.). *The Routledge handbook of corpus linguistics*. Routledge: Abingdon, UK. pp. 3–13.
- [6] Johansson, S., 2008. Some aspects of the development of corpus linguistics in the 1970-s and 1980-s. In: Lüdeling, A., Kytö, M. (Eds.). *Corpus linguistics: An international handbook*. De Gruyter: Berlin, Germany. pp. 35–53.
- [7] Hamp-Lyons, L., 2011. English for academic purposes. In: Hinkel, E. (Ed.). *Handbook of research in second language teaching and learning*, Volume II. Routledge: New York, NY, USA. pp. 89–105.
- [8] Polyakov, A.E., Pilshchikov, I.A., Bergelson, M.B., 2009. Concordance to the texts of Lomonosov – concept and implementation. (in Russian). Available from: <http://feb-web.ru/feb/lomoconc/abc> (cited 12 July 2024).
- [9] Harris, B., 1988. Keywords. A history of debriefing in social psychology. In: Morawski, J.G. (Ed.). *The rise of experimentation in American psychology*. Yale University Press: New Haven, CT, USA. pp. 188–212.
- [10] Brown, P.F., Cocke, J., Pietra, S.A.D., et al., 1990. A statistical approach to machine translation. *Computational Linguistics*. 16(2), 79–85. Available from: <https://aclanthology.org/J90-2002.pdf> (cited 12 July 2024).
- [11] King, P., 1996. Trialling a Multilingual Parallel Concor-dancer. *Proceedings of the Second International Conference on Current Trends in Studies of Translation and Interpreting*; September 5–September 7, 1996; Budapest, Hungary. pp. 49–50.
- [12] Toldova, S.Y., 2011. Concordance. (in Russian). Available from: <http://www.lomonosov-fund.ru/enc/ru/encyclopedia:0127200:versions> (cited 14 December 2023).
- [13] Dubichinsky, V.V., 2009. The main functions of dictionaries. In: *Word and dictionary = Vocabulum et vocabularium: a collection of scientific articles*. Ya. Kupala Grodno State University: Grodno, Belarus. pp. 3–6. (in Russian)
- [14] Stolyarov, A.I., 2017. The dictionary-concordance and its application in the framework of corpus linguistics. *Humanitarian Scientific Research*. 2. (in Russian). Available from: <https://human.snauka.ru/2017/02/21074> (cited 13 January 2024).
- [15] Bektaev, K.B., Dzhubonov, A.X., 1979. Frequency dictionary of the novel M.O. Auezov “The way of Abai”.

- Gylym: Almaty, Kazakhstan. pp. 1–336. (in Russian)
- [16] Kaliyev, B.K., Tuimebayev, Z.K., Kurmanbayuly, S., et al., 2019. Dictionary of Mukagali language. Keremet Media: Almaty, Kazakhstan. (in Russian)
- [17] Dzhubanov, A.H., Khasanov, B., 1973. Statistical study of Kazakh text using a computer. Ata: Almaty, Kazakhstan.
- [18] Zhubanov, A.K., 2009. Corpus linguistics-The main direction of Kazakh linguistics. *Tiltanyim*. 2(34), 3–11. (in Kazakh)
- [19] Rykov, V.V., 2002. Texts Corpus as an Implementation of an Object-Oriented Paradigm. *Dialogue* 2024. (in Russian). Available from: <https://www.dialog-21.ru/digest/2002/articles/rykov/> (cited 25 December 2023).
- [20] Zakharov, V.P., Bogdanova, S.Y., 2011. Corpus linguistics: A textbook for students of humanitarian universities. Irkutsk State Linguistic University: Irkutsk, Russia. pp. 1–161. Available from: <https://www.iprbookshop.ru/21088.html> (cited 12 July 2024).
- [21] Aitova, N., Baishukurova, G., Irgebayeva, A., 2024. Corpus-based study of the works of A. Baitursynuly (based on the collections “Kyryk mysal” (“Forty examples”) and “Masa” (“Mosquito”)). *Bulletin of the Karaganda university. Philology series*. 11529(3), 14–23. DOI: <https://doi.org/10.31489/2024ph3/14-23>