Article
UDC 81’1
https://doi.org/10.31489/2024Ph4/56-66

Received: 17.07.2024
Accepted: 29.10.2024

A. Baimyrza[1*], K. Pirmanova[2], A. Serikbayeva[3]

[1,2,3]*Al-Farabi Kazakh National University, Almaty, Kazakhstan*
(e-mail: baimurzina.ainur@gmail.com)

Challenges and Prospects of Creating Abai Poetic Corpus

This scientific article explores the complexities and prospects of creating a poetic corpus of Abai Kunanbayev. The primary goal of the project is to digitize and annotate Abai's poetry, addressing challenges such as the accessibility of diverse sources, linguistic barriers, technical difficulties, and cultural interpretations. It presents the results obtained from the creation of the corpus, highlighting its significance in preserving Kazakh cultural heritage and advancing scholarly research. The discussion explores the implications of the corpus for linguistic studies, educational initiatives, and international collaborations in digital humanities. The conclusion reflects on the project's achievements, outlines future directions for expanding the corpus, and emphasizes its role in promoting and preserving Abai Kunanbayev literary legacy.

Keywords: poetic text, literary corpora, poetic corpus, linguistic analysis, digital context.

Introduction

The poetic corpus is one of the sub-corpora of the national corpus. The National Language Corpus is an extraordinary and innovative resource on language. Upon request, it delivers a wide range of comprehensive linguistic information within seconds, supported by statistical data. It broadens the empirical foundation and significantly enhances the explanatory power of linguistic research, enabling the formulation of entirely new questions and providing a more precise understanding of linguistic phenomena and their functioning in the language [1; 378].

The creation and analysis of literary corpora have become essential components of linguistic and literary research in the digital age. In this context, the development of technology for constructing poetry corpora offers significant opportunities for exploring the works of renowned poets and understanding their cultural and linguistic impact. This paper focuses on the technology for building a poetry corpus based on the oeuvre of Abai Kunanbayev, a prominent figure in Kazakh literature.

Abai Kunanbayev poetry holds a central place in Kazakh cultural heritage, serving as a source of inspiration and reflection on various aspects of life, society, and spirituality [2; 15]. However, despite the richness and depth of his works, there has been a lack of comprehensive digital resources for studying his poetry in depth. This gap highlights the necessity of the importance of developing a technology-enabled approach to curating and analyzing Abai poetic corpus.

The aim of this paper is to examine the methods and tools used in the construction of a poetry corpus centered around Abai's literary legacy. We delve into the intricacies of data collection, annotation, and analysis, shedding light on the challenges and opportunities inherent in this endeavor. Furthermore, we explore the implications of this technology for linguistic and literary studies, highlighting its potential to deepen our understanding of Abai's poetic style, thematic focus, and cultural significance.

Through this exploration, we seek to contribute to the growing body of research on digital humanities and corpus linguistics, while also providing scholars and enthusiasts with a valuable resource for studying the works of Abai Kunanbayev in a digital context.

The establishment of a corpus of Abai's poetic texts aims to achieve several key objectives:

1. Preservation of cultural heritage: Digitization and systematic organization ensure the preservation and accessibility of texts for future generations.

2. In-depth analysis: The corpus provides opportunities for comprehensive linguistic and literary analysis of texts using modern data processing methods.

3. Educational purposes: The corpus can serve as an important tool for teaching Kazakh literature, culture, and the history of Kazakhstan.

* Corresponding author’s e-mail: *baimurzina.ainur@gmail.com*

4. International collaboration: Digitized and annotated texts of Abai become accessible to researchers worldwide, promoting international scholarly dialogue and cultural exchange [3, 20–29].

However, the process of creating Abai's poetic corpus faces several challenges, such as access to source materials, linguistic barriers, technical issues (Natural Language Processing, Optical Character Recognition, Automatic tagging problems, etc.), and cultural aspects of text interpretation. This article discusses these challenges, explores prospects for corpus creation, proposes methods to overcome them, and discusses potential directions for the project's future development.

This study follows the IMRAD model (Introduction, Methods, Results, and Discussion). The "Methods" section describes the stages of text collection, digitization, and annotation, as well as the use of modern natural language processing (NLP) tools. The "Results" section presents key findings from the analysis of Abai's texts. In the "Discussion" section, major challenges are examined, and approaches to overcoming them are proposed. The conclusion summarizes the research findings and outlines future perspectives for the development of the Abai poetic corpus project.

Literature review

Research on the creation of literary corpora and their analysis is becoming increasingly common in modern linguistics and literary criticism. A particularly important development in this area is the development of technologies for creating poetry corpora, which provide researchers with access to large volumes of poetic texts for analysis and research.

The first publications in the field of automating the analysis of metrorhythmic characteristics of poetic texts emerged in the mid-1990s, with M. Hayward conducting a computer study of the metrics in poems by various poets [4; 1–11]. However, broader interest in this research area developed only in the late 2000s to early 2010s. A similar study, conducted by J. Kao and D. Jurafsky, aimed to identify stylistic features that differentiate professional poets from amateur poets [5; 8–15].

The SPARSAR system, described by R. Delmonte [Delmonte, 2013], offers an automated comprehensive analysis of poetic texts to examine their stylistic elements. A series of publications by V.B. Barakhnin and O. Yu. Kozhemyakina, the authors of this article, is dedicated to automating the complex analysis of Russian-language poetic texts, such as [6; 16–27]. Among other works in this area, including those related to the analysis of poetry in Turkic languages, is the article by A. Kurt and M. Kara [7; 10–12], which proposes an algorithm for recognizing and analyzing poems written in the "arud" system, a structure characteristic of Eastern (Arabic, Persian, Turkish) poetry.

It should be noted that the algorithms of analysis of the lower levels of poetic texts strongly depend on the characteristics of a particular language, as well as on the degree of development of computer technologies for this language. Thus, the peculiarities of phonetic analysis of texts in English are poor paradigm of word changes and the presence of a large number of network dictionaries of phonetic analysis. In German — quite simple and strict rules of morphological changes of words and phonetic characteristics of word forms. In Russian — the absence of any general rules of morphological changes of words and their phonetic characteristics (primarily emphasis). In Turkish and other Turkic languages, including Kazakh — almost deterministic rules of the formation of the word forms and of the change of their phonetic characteristics following from the law of synharmonism which is characteristic to Turkic languages.

In the context of Kazakh literature, works such as Akhmetov's [8, 24–30] study of the poetic legacy of Abay Kunanbayev play an important role in understanding the cultural significance and literary influence of a given poet.

However, these studies pay insufficient attention to the use of modern technologies for the creation and analysis of poetry corpora. Our research is aimed at filling this gap by developing a technology for creating a poetic corpus using the example of the works of Abay Kunanbayev. We aim to combine methods from modern linguistics [9, 25-26] and computer science to create a valuable resource for researchers of Kazakh literature and language.

In achieving the intended goal and completing the tasks, the preliminary research began with a familiarization of scientific works in two directions: the first — the results of scientific work on automating the process of analyzing poetic texts [10; 68–75], [11; 1–15], [12; 5–18], the second — research devoted to the work of Abai Kunanbayev [2, 13]. In this studies, the structural and rhythmic organization of the texts was not given enough importance, and most of the attention was paid to the thematic and contextual analysis of the poems. For example, many Kazakh researchers focus on the cultural and ideological load of poetry. But the aspects of form and style are less studied. Comparing the results of this study with previous works it

shows that in the majority of studies aspects of corpus linguistics are not considered sufficiently and are focused on biographical and literary analysis. Since this is the first author's poetic corpus in the Kazakh language, familiarization work has been carried out with foreign poetic corpuses such as the Shakespeare Corpus, the poetic sub-corpus of the National corpus of Russian language.

Methods

For the creation of Abai Kunanbayev poetic corpus, a comprehensive approach was employed, encompassing several stages and the use of various tools and technologies. This section describes the primary methods used for collecting, digitizing, annotating, and analyzing texts.

Text Collection and Digitization

The first step in creating the corpus involved collecting all available editions of Abai's poetry, including:

1) Printed materials: Collection of various editions of Abai's poems published over different years.

2) Manuscripts and archive materials: Search and collection of manuscripts and other archival materials stored in libraries and archives across Kazakhstan.

3) Electronic sources: Compilation of texts from electronic libraries and online resources. Collected all the common famous poems of Abai. 225 poems were digitized and annotated for this project. After collecting all available materials, they underwent digitization, which included scanning printed and manuscript texts and converting them into electronic format using Optical Character Recognition (OCR) technology.

The adaptation of OCR tools to the Kazakh language posed several challenges, primarily due to the unique characteristics of the Kazakh script, which uses Cyrillic but includes specific letters not found in Russian [ә, ғ, қ, ң, ө, ұ, ү, і]. Existing OCR systems are generally optimized for languages with larger data sets, such as Russian or English, which meant that their application to Kazakh texts resulted in lower accuracy rates. To overcome this, we had to fine-tune the OCR models by incorporating Kazakh-specific characters and using training data consisting of historical Kazakh texts, including those written in Abai's Old Kazakh.

Linguistic Annotation

Following digitization, linguistic annotation of the texts was performed, involving several stages. The first stage is morphological annotation, which uses tools such as MORPHEUS for automatic recognition and tagging of parts of speech in the texts. The second stage is syntactic annotation, where tools like TreeTagger are applied to determine the syntactic structure of sentences. The final stage is thematic classification, involving the manual tagging of texts into thematic categories such as philosophical reflections, social critique, love lyrics, and religious motifs.

Use of Software Tools

Modern Natural Language Processing (NLP) tools and technologies were employed for text processing and analysis. For automatic morphological annotation of the texts, tools like MORPHEUS and similar programs were used for morphological analysis. Tools for syntactic analysis, including TreeTagger and other similar tools, were employed for the automatic detection of sentence syntactic structures. Additionally, custom scripts and existing programs were developed and utilized for the automatic classification of texts based on themes.

In terms of NLP, the primary challenge was adapting existing tools for morphological and syntactic analysis to handle the complex morphology of Kazakh, which is an agglutinative language. Kazakh words are formed by adding multiple affixes to a root word, resulting in a wide variety of word forms. This is in stark contrast to languages like English, where inflection is limited. Many widely-used NLP tools were not designed with agglutinative languages in mind, leading to inaccuracies in part-of-speech tagging, lemmatization, and syntactic parsing.

To address these issues, we utilized and adapted specific NLP tools such as MORPHEUS, which we customized to better handle Kazakh morphology. This included expanding the tool's lexicon with Kazakh vocabulary, adjusting its rules for inflection and word formation, and incorporating a larger set of annotated Kazakh texts for training. We also had to develop custom rules for handling unique linguistic phenomena in Abai's Old Kazakh, such as archaic word forms and syntactic structures not commonly found in modern Ka-

zakh. The process involved collaboration with linguists specializing in Kazakh language history to ensure accuracy.

Database Creation

Based on annotated texts, a relational database was created to facilitate complex queries and text analysis by performing the following steps:

- Database structure design: Defining the main tables and fields of the database, including tables for storing texts, annotations, and metadata [14; 3–14].
- Data import: Loading digitized and annotated texts into the database.
- Development of search and analysis interfaces: Creating web interfaces and tools for convenient access to the database and for conducting text analysis.

Validation and Quality Assurance

To ensure high quality and accuracy of annotations and analysis, the following procedures were implemented. Firstly, a manual annotation review was conducted, involving selective manual verification of automatic annotations and markings to assess their accuracy and correct errors. Secondly, database testing was performed, which included functional testing of the database and tools for search and analysis using various datasets. Lastly, cross-validation was carried out by comparing results of automatic annotation and marking with those obtained manually by experts to evaluate the accuracy of automatic methods.

Application of these methods resulted in the creation of a high-quality and versatile poetic corpus of Abai Kunanbayev, providing opportunities for in-depth analysis and research into his literary works.

Results

The creation of Abai Kunanbayev poetic corpus has yielded significant results, encompassing thematic and linguistic diversity in his poetry and opening new avenues for study and interpretation of his works. This section outlines the main findings obtained during the development and analysis of the corpus.

Thematic Diversity

Analysis of Abai Kunanbayev poetic corpus has revealed a wide spectrum of themes addressed in his works. Key thematic categories include: 1) philosophical reflections — Abai often delved into questions of life's meaning, morality, ethics, and human existence. His poems contain profound philosophical reflections and thoughts on spiritual values. 2) Social critique: Abai's poetry includes sharp criticism of social issues of his time, such as injustice, corruption, inequality, and ignorance. 3) Love lyrics: some of Abai's works are dedicated to the theme of love, where he expresses his personal feelings and reflections on the nature of love and human relationships. 4) Religious motifs: his verses frequently feature religious themes and motifs that reflect his views on spirituality and faith [15; 64–73].

Linguistic Complexity

Textual analysis has demonstrated a high degree of lexical and syntactic diversity. Key observations include: 1) exical richness: Abai's texts employ a wide range of lexical units, including archaisms and borrowings from other languages, reflecting the cultural and historical diversity of the Kazakh people. 2) Syntactic complexity: The sentence structure in Abai's verses often exhibits complexity and variety, including complex sentences and the use of various stylistic devices [16; 33–36].

Translation and Interpretation

The creation of the corpus has underscored the need for accurate and nuanced translations of Abai's works into other languages. Major findings include: 1) translation challenges: Translating Abai's poetry into other languages requires a deep understanding of the context, culture, and linguistic nuances of the original texts. 2) Need for multilingual annotations: For international use of the corpus, it is crucial to develop multilingual annotations and commentaries that help researchers and readers from different countries better understand and interpret Abai's texts.

Social Significance

The Abai poetic corpus has become a vital tool for studying Kazakh culture and history. Key achievements include: 1) educational use: The corpus is utilized in educational institutions for teaching Kazakh literature and history, contributing to the preservation and popularization of cultural heritage. 2)

Research opportunities: The corpus offers scholars and researchers access to a valuable data source for interdisciplinary studies, including linguistic analysis, literary research, and socio-cultural studies. 3) Cultural heritage: Digitization and systematization of Abai's poetic texts contribute to the preservation of his legacy and make it accessible to a wide audience of readers and researchers [17; 2–8].

Technical Achievements

The development of Abai Kunanbayev poetic corpus involved several technical achievements: 1) Database creation: Development of a relational database for storing and managing digitized and annotated texts, enabling complex queries and analysis; 2) Analysis tools: Development and use of software tools for morphological and syntactic analysis of texts, as well as thematic classification and search capabilities.

For example, morphological identification methods were defined:

Automatic tagging: using computer programs that automatically analyze text and assign tags based on predefined rules and dictionaries.

Manual tagging: text analysis and tagging by linguists. This method provides high accuracy, especially in case of complex or mixed language constructions.

Hybrid approach: combining automatic and manual labeling to improve process accuracy and efficiency.

Conventional signs for morphological designation:

The following (international) symbols may be used for the Kazakh language:

Noun: N
Adjective: ADJ
Verb: V
Adverb: ADV
Pronoun: PRON
Number noun: NUM
Auxiliary word: PREP
Conjugation: CONJ
Case (for example, name case, income case): NOM, ACC, etc.
Tense (eg present, past): PRES, PAST, etc.
Singular, plural: SG, PL

As an example of the morphological markup model for Abai's line “Zhelsiz tunde zharyq ai” (translated as “Bright moon on a windless night”), the following symbols can be used:

Zhelsiz (windless): Part of speech: Adjective (ADJ) Case: Nominative (NOM) Number: Singular (SG)
Tunde (night): Part of speech: Adverb (ADV) Case: In the prepositional (LOC) Number: Singular (SG)
Zharyq (bright): Part of speech: Adjective (ADJ) Case: Nominative (NOM) Number: Singular (SG)
Ai (moon): Part of speech: Noun (N) Case: Nominative (NOM) Number: Singular (SG).

Thus, the morphological markup of this line looks like this: Zhelsiz (ADJ, NOM, SG) tunde (ADV, LOC, SG) zharyq (ADJ, NOM, SG) ai (N, NOM, SG).

This markup allows for a detailed analysis of the morphological features of the poetic language used in Abay's poem and promotes a deep understanding of the text from both linguistic and literary points of view.

These conventions are applied to each word or word form in the poem to indicate their morphological characteristics. This allows for a detailed analysis of the text and the study of its linguistic features.

Web interfaces: Creation of user-friendly web interfaces for accessing the database and conducting text analysis, making the corpus accessible to a broad range of users (Fig. 1).

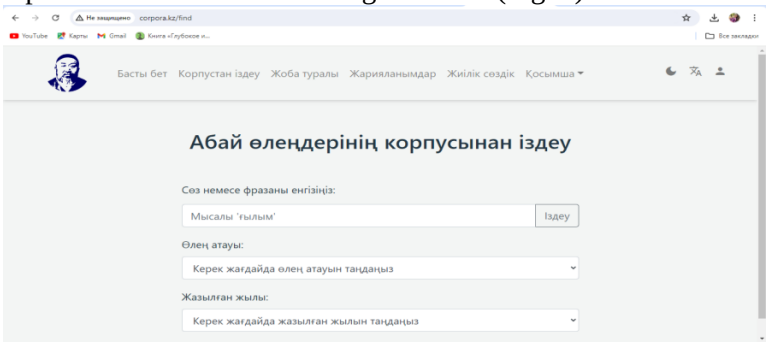

Figure 1. A searching page of Abai Poetic Corpus at www.corpora.kz

Abai Kunanbayev poetic corpus is a collection of poetic texts collected and organized for research, data processing, or other purposes. It provides a valuable source of information for analysis and research in various fields. Abai poetic corpus is useful for analyzing the structure, style, theme, and cultural significance of his poetic works based on established texts (Fig. 2).

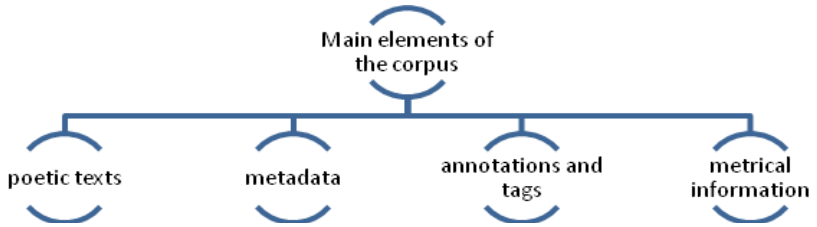

Figure 2. Main elements of Abai Poetic Corpus

Main elements of the corpus: poetic texts: Abai's poems written at different times; metadata: information about each poetic text that may be useful for analysis (title of work, year of composition, genre, cultural context, and other details); annotations and tags: additional tags or annotations that describe text features, themes, styles used, and other aspects that may be useful in classifying and analyzing texts; metrical information: provides information about the poem's metric, rhyme, and other formal aspects as needed.

The creation of Abai Kunanbayev poetic corpus represents a significant contribution to the development of Kazakh philology and cultural studies. The results obtained open new perspectives for the study of his works and contribute to the preservation and popularization of Kazakhstan's cultural heritage.

Discussion

The creation of Abai Kunanbayev poetic corpus represents a complex and multi-stage process that entails overcoming various technical, linguistic, and cultural barriers. This section discusses the main challenges faced by the corpus developers, as well as the prospects and directions for further project development.

Challenges in Corpus Creation

One of the primary challenges is the availability of source materials: 1) Rare and unique sources: Many of Abai's texts exist as rare manuscripts or old editions that are difficult to locate and digitize. Some are held in private collections or archival funds with limited access. 2) Quality of sources: The condition of some sources is suboptimal. Due to aging paper and ink, texts may be damaged or illegible, complicating the digitization and subsequent processing [18; 20–24].

The second problem is the linguistic barriers. Abai's poetry is written in Old Kazakh, which presents specific challenges. Firstly, the use of archaic and dialectal forms requires specialized knowledge and experience for accurate interpretation and annotation. Secondly, the multilingual nature of Abai's texts, which include borrowings from other languages such as Arabic, Persian, and Russian, complicates automated analysis and necessitates expertise in these languages.

Technical aspects of creating a poetic corpus also pose significant challenges. The first of them is NLP tool limitations: Modern natural language processing (NLP) tools may not always accurately process poetic texts, especially with archaic language forms and complex syntactic structures. The second is OCR quality: Optical character recognition (OCR) technology may not provide high-quality recognition for old and damaged texts, requiring additional manual verification and correction.

Interpreting Abai's works requires a deep understanding of the context of his time and culture. Historical context: Researchers must consider the historical and social context in which Abai lived and created in order to interpret his works accurately. Cultural nuances: Abai's poetry is rich in cultural and national nuances, which may be challenging for international audiences to grasp without additional commentary and explanations.

The cultural and historical aspects are revealed through macro- and micropoetic analyzes. For example, when analyzing Abai’s verse «Qalyn elim, qazagym, qairan zhurtym», we can see the cultural and historical peculiarity of the people of that time. But the interpretation of each reader, as well as the researcher, may

differ. Here is one of the translations (by Dorian Rottenberg) and interpretations of two couplets of this verse:

The first couplet:

Qalyn elim, qazagym, qairan zhurtym,
Ustarasyz auzyna tusti murtyn.
Zhaksy menen zhamandy aiyrmadyn,
Biri qan, biri mai bop eki urtyn.

«Qalyn elim, qazagym, qairan zhurtym» （Oh, my luckless Kazakh, my unfortunate kin） — words expressing Abai's love and concern for his people. The phrase *«qalyn elim»* shows the large number of people, the word *«qazagym»* means national unity, and the phrase *«qairan zhurtym»* shows the difficult situation of the people. «*Ustarasyz auzyna tusti murtyn»*（An unkempt moustache hides your mouth and chin) — this metaphor refers to the ignorance and shortcomings of the people. An unshaven mustache is a symbol of unpreparedness, indifference. «*Zhaqsy menen zhamandy aiyrmadyn»* （Your looks are not bad and your numbers are vast) — people's inability to distinguish between right and wrong, loss of moral values. «*Biri qan, biri mai bop eki urtyn»* **(**Your looks are not bad and your numbers are vast) — describes the hypocrisy of the people, distrust of each other.

The second couplet:

Ozimdiki dei almai oz malyndy,
Kundiz kulkin buzyldy, tuned uiqyn.
Korseqyzar keledi bailauy zhoq,
Bir kun tyrtyn etedi, bir kun bultyn.

«Ozimdiki dei almai oz malyndy» (Unable to manage your property) — people's inability to use their property properly. «*Kundiz kulkin buzyldy, tuned uiqyn»* （Day and night, care and worry are all you see) — difficulty of life, lack of peace. «*Korseqyzar keledi bailauy zhok»* （Constant ia nought but inconstancy） – *people's lack of control, lack of discipline. «Bir kun tyrtyn etedi, bir kun bultyn»* (Now haughty, now wearing a look of offence) — instability, one day fun, one day sad life.

Prospects and Directions for Further Development

Despite numerous challenges, the project to create Abai Kunanbayev's poetic corpus holds substantial promise. This section explores potential directions for its further development. Firstly, there is the expansion of the corpus, which includes the inclusion of new texts that may be discovered in private collections or archival funds. Moreover, it involves the digitization and annotation of these new texts, including morphological and syntactic tagging [19; 118–124].

The second direction is the improvement of tools and methods. This direction includes the development of new tools, such as creating and implementing more advanced tools for morphological and syntactic analysis of texts, specifically adapted for working with archaic and dialectal language forms. Additionally, it involves the enhancement of OCR quality by utilizing state-of-the-art machine learning technologies for processing old and damaged texts.

The next direction is international collaboration, which can develop along two vectors. The first vector is multilingual annotations, focusing on developing multilingual annotations and commentaries to aid international researchers and readers in better understanding and interpreting Abai's texts. The second vector is network projects, involving participation in international network projects and conferences dedicated to digital humanities and literary studies to exchange knowledge and expertise.

And also one of the directions is the educational programs. Firstly, the integration into curricula: Using the corpus in educational institutions to teach Kazakh literature, culture, and history. Secondly, the development of educational materials: creating educational materials and guides based on Abai Kunanbayev poetic corpus for schools and universities.

Significance and Contribution

The creation of Abai Kunanbayev's poetic corpus holds significant importance for the preservation and study of Kazakh cultural heritage. The project contributes to the preservation of cultural heritage by digitizing and systematizing Abai's poetic texts, ensuring their preservation and accessibility for future generations. Additionally, it advances scholarly research by providing scholars and researchers with a unique data source for interdisciplinary studies. Furthermore, the project promotes the popularity of Abai's work

both within Kazakhstan and internationally, drawing attention from the global scholarly and reader community.

Conclusion

The creation of Abai Kunanbayev poetic corpus represents a significant contribution to the development of Kazakh philology and cultural studies, as well as to the preservation and study of national cultural heritage. Despite numerous challenges, the project holds great potential and opens up new opportunities for researchers and educators alike.

This corpus is a rich source of data for research in lexicography and grammar. Research in the field of semantics is closely related to research in lexicography. By observing the environment of a particular linguistic unit in a corpus, it is possible to establish certain semantic features that characterize this unit.

Theoretical linguists can use the corpus as an experimental basis to test hypotheses and prove their theories. Applied linguists (teachers, translators, etc.) use it when teaching languages and to solve their professional problems. Computational linguists represent a special class of users: they can identify and use statistical and linguistic patterns present in texts to create computer models of language. Other language specialists (literary scholars, editors) can also, in some cases, get answers to their questions by turning to this corpus. Social scientists (historians, sociologists) can also study their subjects through language, using text parameters such as period, author, or genre. Literary scholars use corpora for stylometric research. Finally, the corpus is used to develop and configure various automated systems (machine translation, speech recognition, information retrieval).

Acknowledgments

This research is funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No.BR21882334 "Kazakh poetic corpus development: morphological and poetic designation of Abai's poems").

References

1 Сулейменова Э.Д. Макросоциолингвистика [Текст] / Э.Д. Сулейменова. — Алматы: Қазақ университеті, 2011. — 406 с.

2 Нұрқатов А. Абайдың ақындық дәстүрі / А. Нұрқатов. — Алматы: Жазушы. — 1966. — 344 б.

3 Kudaibergenova D.T. Misunderstanding Abai and the legacy of the canon: «Neponyatnii» and «Neponyatii» Abai in contemporary Kazakhstan / D.T. Kudaibergenova // Journal of Eurasian Studies. — 2018, — 9(1). — 20–29. https://doi.org/10.1016/j.euras.2017.12.007

4 Hayward M. Analysis of a corpus of poetry by a connectionist model of poetic meter / M. Hayward // Poetics. — 1996. — Vol. 24, Issue 1. — P. 1–11.

5 Kao J. A computational analysis of style, affect, and imagery in contemporary poetry / J. Kao, D. Jurafsky // Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature. — 2012. — P. 8–17.

6 Barakhnin V.B. Usage of modern computer technologies in the learning process of the philologists of complex analysis of Russian poetic texts / V.B. Barakhnin, O. Yu. Kozhemyakina, A.V. Zabaykin // International Conference «Education Environment for the Information Age» (EEIA-2016). (SHS Web of Conference). — 2016. — Vol. 29.

7 Kurt A. An algorithm for the detection and analysis of arud meter in Diwan poetry / A. Kurt, M. Kara M. // Turkish Journal of Electrical Engineering & Computer Sciences. — 2012. — 20 (6).

8 Ахметов З.А. Поэтика эпопеи «Путь Абая» в свете истории ее создания [Текст] / З.А. Ахметов. — Алма-Ата: Наука, 1984. — 256 с.

9 Madieva G.B. Innovations in the study of Kazakh poetry: from theory to practice in the development of the Kazakh poetic corpus / G.B. Madieva // International scientific and practical conference: "Modern technologies of computer linguistics — CTCL.2024". — Uzbekistan, Tashkent, 04.26.2024.

10 Rakhimova D. Semantic analysis of the Kazakh language based on the approach of neural networks / D. Rakhimova, A. Turganbayeva // News of the National Academy of Sciences of the Republic of Kazakhstan. Physico-mathematical series. — 2020. — Vol. 5, No. 333. — P. 68–75, http://doi.org/10.32014/2020.2518-1726.84

11 Akhmed-Zaki D. Development of the information system for the Kazakh language preprocessing / D. Akhmed-Zaki, M. Mansurova, G. Madiyeva, N. Kadyrbek, M. Kyrgyzbayeva // Cogent Engineering. — 2021. — Vol. 8, Issue 1. — 1896418. — DOI: 10.1080/23311916.2021.1896418.

12 Барахнин В.Б. Автоматизация комплексного анализа русского поэтического текста: модели и алгоритмы / В.Б. Барахнин, О.Ю. Кожемякина, А.В. Забайкин, В.Д. Хаятова // Вест. Новосиб. гос. ун-та. Сер. Информационные технологии. — 2015. — Т. 13. — Вып. 3. — С. 5–18.

13 Негимов С. Абай ғибратнамасы / С. Негимов. — Астана, 2023.

14 Akar A. Methods of transferring full word-formation affixes in the national corpus of the Kazakh language / A. Akar, K.K. Pirmanova, N.Zh. Shaimerdenova, I.M. Shalabaieva // Tiltanym. — 2023. — No 3(91). — P. 3–14. https://doi.org/10.55491/2411-6076-2023-3-3-14

15 Яковлев А.А. Корпус как универсальный информант (об экспериментальном изучении семантики и языковой картины мира) / А.А. Яковлев // Вестн. ТвГУ. Сер. Филология. — 2017. — № 2. — С. 64–73.

16 Kopotev M.V. Principles of creating Helsinki annotated corpus of Russian texts (HANKO) on the Internet / M.V. Kopotev, A. Mustajoki // Scientific and Technical Information. — 2023. — Ser. 2, No. 6. — P. 33–36.

17 Verbitskaya L.A. Some creation problems national corpus of the Russian language / L.A. Verbitskaya, N.N. Kazansky, V.B. Kasevich // Scientific and Technical Information. — 2003. — Ser. 2, No. 6. — P. 2–8.

18 Гаспаров М.Л. Статьи о лингвистике стиха / М.Л. Гаспаров, Т.В. Скулачева. — М.: Языки славянской культуры, 2004. — 283 с.

19 Осипов Г.А. О перспективах корпусных исследований и практике их применения в лингвистике / Г.А. Осипов // Вестн. Адыгей. гос. ун-та. Сер. Филология и искусствоведение. — 2023. — Вып. 1(312). — С. 118–124.

20 Barakhnin V. Automated determination of the type of genre and stylistic coloring of Russian texts / V. Barakhnin, O. Kozhemyakina, I. Pastushkov // Seminar on Systems Analysis (ITM Web of Conf.). — 2017. — Vol. 10.

21 Delmonte R. Computing poetry style / R. Delmonte. — 1st International Workshop ESSEM. — 2013.

22 Doszhan G. Problems of Creation of the All-Turkic National Corpus / G. Doszhan // Proceedings of the International Conference on Information, Business and Education Technology (ICIBET 2013). — 2013. — 03. — 610–615, doi:10.2991/icibet.2013.131

23 Гришина Е.А. Поэтический корпус в рамках Национального корпуса русского языка: общая структура и перспективы использования / Е.А. Гришина, К.М. Корчагин, В.А. Плунгян, Д.В. Сичинава // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. — СПб.: Нестор-История, 2009. — С. 71–113.

24 Корпусная лингвистика: учебный словарь ключевых терминов и понятий / сост.: Г.Б. Мадиева, С. Бектемирова, Н. Исмайлова. — Алматы: Қазақ университеті, 2017. — 27 с.

25 Мадиева Г.Б. Ұлттық корпустардың жасалу мәселелері [Электрондық ресурс] / Г.Б. Мадиева, С.Б. Бектемирова, П.Т. Медетбекова, А.А. Молдасанова / ПМУ Хабаршысы. Филологиялық сериясы. — 2018. — №3. — Б. 222–230. — Қолжетімділігі: https://vestnik-philological.tou.edu.kz/storage/journals/100.pdf

26 McEnery T. Corpus Linguistics / T. McEnery, A. Wilson. — Edinburgh: Edinburgh University Press, 1999. — 256 p.

27 Орехов Б.В. «Проблеск» Ф.И. Тютчева в ретроспективе Корпуса. Очерк корпусной поэтики / Б.В. Орехов // В кн.: Корпусный анализ русского стиха: сб. науч. ст. — М.: Изд. центр «Азбуковник», 2014. — Вып. 2. — С. 304–318.

28 Сичинава Д. В. Поэтический подкорпус Национального корпуса русского языка: несколько примеров поиска стиховедческой информации // Славянский стих. — М.: Рукописные памятники Древней Руси, 2012. — С. 482–491.

29 Smith J. Methods in Corpus Linguistics / J. Smith. — Cambridge University Press, 2010.

30 Stefanowitsch A. Collostructions: investigating the interaction between words and constructions / A. Stefanowitsch, S.Th. Gries // International journal of corpus linguistics. — 2003. — 8 (2). — P. 209–243.

31 Абай шығармаларының поэтикалық корпусы. — [Электрондық ресурс]. — Қолжетімділігі: http://corpora.kz/

32 Национальный корпус русского языка. — [Электронный ресурс]. — Режим доступа: https://ruscorpora.ru/corpus/poetic.

33 Open Source Shakespeare. — [Electronic resource]. — Access mode: https://www.opensourceshakespeare.org/

А. Баймырза, К. Пірманова, А. Серікбаева

Абайдың поэтикалық корпусының мәселелері мен болашағы

Мақалада Абай Құнанбаевтың поэтикалық корпусын құрудың қиындықтары мен болашағы қарастырылған. Жобаның негізгі мақсаты — түрлі дереккөздердің қолжетімділігі, тілдік кедергілер, техникалық мәселелер мен мәдени интерпретациялар сияқты мәселелерді қарастыра отырып, Абай поэзиясын цифрландыру және аннотациялау. Авторлар Абай поэзиясын цифрлау мен аннотациялауда қолданылатын әдіс-тәсілдерді, оның ішінде тілдік талдау мен тақырыптық классификацияны зерделеген. Корпусты құру барысында алынған нәтижелер көрсетіліп, оның қазақтың мәдени мұрасын сақтаудағы және ғылыми зерттеулерді дамытудағы маңыздылығы көрсетілді.

Кілт сөздер: поэтикалық мәтін, әдеби корпус, поэтикалық корпус, лингвистикалық талдау, цифрлық контекст.

А. Баймырза, К. Пирманова, А. Серикбаева

Проблемы и перспективы создания поэтического корпуса Абая

В статье исследованы сложности и перспективы создания поэтического корпуса Абая Кунанбаева. Основная цель проекта — оцифровать и аннотировать поэзию Абая, решая такие проблемы, как доступность различных источников, языковые барьеры, технические проблемы и культурные интерпретации. Авторы рассматривают методы, используемые при оцифровке и аннотировании поэзии Абая, включая лингвистический анализ и тематическую классификацию. В работе представлены результаты, полученные при создании корпуса, при этом подчеркивается его значение в сохранении казахского культурного наследия и развитии научных исследований.

Ключевые слова: поэтический текст, литературный корпус, поэтический корпус, лингвистический анализ, цифровой контекст.

References

1 Suleimenova, E.D. (2011). *Makrosotsiolingvistika [Macrosociolinguistics]*. Almaty: Qazaq University [in Russian].

2 Nurkatov, A. (1966). *Abaidyn aqyndyq dasturi [Abai's poetic tradition]*. Almaty: Zhazushy [in Kazakh].

3 Kudaibergenova, D.T. (2018). Misunderstanding Abai and the legacy of the canon: «Neponyatnii» and «Neponyatii» Abai in contemporary Kazakhstan. *Journal of Eurasian Studies, 9*(1), 20–29. https://doi.org/10.1016/j.euras.2017.12.007.

4 Hayward, M. (1996). Analysis of a corpus of poetry by a connectionist model of poetic meter. *Poetics, 24* (1), 1–11.

5 Kao, J., & Jurafsky, D. (2012). A computational analysis of style, affect, and imagery in contemporary poetry. *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, 8–17.

6 Barakhnin, V., Kozhemyakina, O.Yu., & Zabaykin, A. (2016). Usage of modern computer technologies in the learning process of the philologists of complex analysis of Russian poetic texts. *International Conference «Education Environment for the Information Age» (EEIA-2016). (SHS Web of Conference), 29.*

7 Kurt, A., & Kara, M. (2012). An algorithm for the detection and analysis of arud meter in Diwan poetry. *Turkish Journal of Electrical Engineering & Computer Sciences, 20* (6).

8 Akhmetov, Z.A. (1984). *Poetika epopei «Put Abaia» v svete istorii ee sozdaniia [Poetics of the epic Abai's Path in the history of its creation]*. Alma-Ata: Nauka [in Russian].

9 Madieva, G.B. (2024). Innovations in the study of Kazakh poetry: from theory to practice in the development of the Kazakh poetic corpus. *International scientific and practical conference: "Modern technologies of computer linguistics — CTCL.2024".* Uzbekistan, Tashkent.

10 Rakhimova, D., & Turganbayeva, A. (2020). Semantic analysis of the Kazakh language based on the approach of neural networks. *News of the National Academy of Sciences of the Republic of Kazakhstan. Physico-mathematical series, 5* (333), 68–75. http://doi.org/10.32014/2020.2518-1726.84.

11 Akhmed-Zaki, D., Mansurova, M., Madiyeva, G., Kadyrbek, N., & Kyrgyzbayeva, M. (2021). Development of the information system for the Kazakh language preprocessing. *Cogent Engineering, 8* (1), 1896418. DOI: 10.1080/23311916.2021.1896418.

12 Barakhnin, V.B., Kozhemyakina, O.Yu., Zabaikin, A.V., & Khayatova, V.D. (2015). Avtomatizatsiia kompleksnogo analiza russkogo poeticheskogo teksta: modeli i algoritmy [Automation of the Complex Analysis of Russian Poetic Text: Models and Algorithms]. *Vestnik Novosibirskogo gosudarstvennogo universiteta. Seriia Informatsionnye tekhnologii — Bulletin of the Novosibirsk State University Series Information technologies, 13* (3), 5–18 [in Russian].

13 Negimov, S. (2023). *Abai gibratnamasy [Abay's edification].* Astana [in Kazakh].

14 Akar, A., Pirmanova, K.K., Shaimerdenova, N.Zh., & Shalabaieva, I.M. (2023). Methods of transferring full word-formation affixes in the national corpus of the Kazakh language. *Tiltanym — Linguistics, 3,* 3–14. https://doi.org/10.55491/2411-6076-2023-3-3-14

15 Yakovlev, A.A. (2017). Korpus kak universalnyi informant (ob eksperimentalnom izuchenii semantiki i yazykovoi kartiny mira) [Corpus as a Universal Respondent (on the Experimental study of the Semantics and the Wordviev]. *Vestnik Tverskogo universiteta. Seriia Filologiia — Bulletin of Tver State University. Series Philology,* 2, 64–73 [in Russian].

16 Kopotev, M.V., & Mustajoki, A. (2023). Principles of creating Helsinki annotated corpus of Russian texts (HANKO) on the Internet. *Scientific and Technical Information, 2* (6), 33–36.

17 Verbitskaya, L.A., Kazansky, N.N., & Kasevich, V.B. (2003). Some creation problems national corpus of the Russian language. *Scientific and technical information, 6,* 2–8.

18 Gasparov, M.L., & Skulacheva, T.V. (2004). *Stati o lingvistike [Articles about linguistics of poems]*. Moscow: Yazyki slavianskoi kultury [in Russian].

19 Osipov, G.A. (2023). O perspektivakh korpusnykh issledovanii i praktike ikh primeneniia v lingvistike [On the prospects of corpus research and the practice of its application in linguistics]. *Vestnik Adygeiskogo gosudarstvennogo universiteta. Seriia*

Filologiia i iskusstvovedenie — Bulletin of the Adyghe State University, Series Philology and Art Criticisms, 1 (312), 118–124 [in Russian].

20 Barakhnin, V., Kozhemyakina, O., & Pastushkov, I. (2017). Automated determination of the type of genre and stylistic coloring of Russian texts. *Seminar on Systems Analysis (ITM Web of Conf.), 10*. DOI: https://doi.org/10.1051/itmconf/20171002001.

21 Delmonte, R. (2013). *Computing poetry style.* 1st International Workshop ESSEM.

22 Doszhan, G. (2013). Problems of Creation of the All-Turkic National Corpus. *Proceedings of the International Conference on Information, Business and Education Technology (ICIBET 2013)*, 03, 610–615. doi:10.2991/icibet.2013.131.

23 Grishina, Ye.A., Korchagin K.M., Plungyan, V.A., & Sichinava, D.V. (2009). Poeticheskii korpus v ramkakh Natsionalnogo korpusa russkogo yazyka: obshchaia struktura i perspektivy ispolzovaniia [Poetic corpus within the National Corpus of the Russian Language: general structure and prospects for use]. *Natsionalnyi korpus russkogo yazyka: 2006–2008. Novye resultaty i perspektivy — The National corpus of the Russian Language: 2006–2008. New results and prospects*. Saint Petersburg: Nestor-Istoriia, 71–113 [in Russian].

24 Madieva, G.B., Bektemirova, S.B., & Ismailova, N.A. (Comp.) (2017). *Korpusnaia lingvistika: uchebnyi slovar kliuchevykh terminov i poniatii [Dictionary of Corpus Linguistics: educational dictionary of key terms and concepts. Educational dictionary]*. Almaty: Qazaq universiteti [in Russian].

25 Madyeva, G.B., Bektemirova, S.B., Medetbekova, P.T., & Moldasanova, A.A. (2018). Ulttyq korpustardyn zhasalu maseleleri [Problems of creating national corpora]. *Pavlodar memlekettik universitetinin khabarshysy. Filologialyq seriasy — Bulletin of Pavlodar Pedagogical University. Philological Series, 3,* 222–231. https://vestnik-philological.tou.edu.kz/storage/journals/100.pdf. [in Kazakh].

26 McEnery, T., & Wilson, A. (1999). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

27 Orekhov, B.V. (2014). «Problesk» F.I. Tiutcheva v retrospektive Korpusa. Ocherk korpusnoi poetiki [F.I. Tyutchev`s "Glimpse" in the retrospective of the Corpsus. An essay on corpus poetics]. *V knige: Korpusnyi analiz russkogo stikha: sbornik nauchykh statei — In the book: Corpus analysis of Russian verse: A collection of scientific articles, 2*, 304–318. Moscow: Izdatelskii tsentr «Azbukovnik» [in Russian].

28 Sichinava, D.V. (2012). Poeticheskii podkorpus Natsionalnogo korpusa russkogo yazyka: neskolko primerov poiska stikhovedcheskoi informatsii [Poetic subcorpus of the National corpus of the Russian language: a few examples of verse study relevant information]. *Slavianskii stikh — Slavic verse*, 482–491. Moscow: Rukopisnye pamiatniki Drevnei Rusi [in Russian].

29 Smith, J. (2010). *Methods in Corpus Linguistics*. Cambridge University Press.

30 Stefanowitsch, A., & Gries, S.Th. (2003). Collostructions: investigating the interaction between words and constructions. *International Journal of Corpus Linguistics, 8* (2), 209–243.

31 Abai shygarmalarynyn poetikalyq korpusy [The Poetic Corpus of Abai`s works]. Retrieved from http://corpora.kz/ [in Kazakh].

32 Natsionalmyi korpus russkogo yazyka [National Corpus of Russian language]. Retrieved from https://ruscorpora.ru/corpus/poetic [in Russian].

33 Open Source Shakespeare. Retrieved from https://www.opensourceshakespeare.org/

Information about the authors

Baimyrza, Ainur Amangaliqyzy — PhD student in "Turkology and Language Theory", Al-Farabi Kazakh National University, Almaty, Kazakhstan; e-mail: baimurzina.ainur@gmail.com;

Pirmanova, Kunsulu Qambarbekqyzy — PhD student in "Turkology and Language Theory", Al-Farabi Kazakh National University, Almaty, Kazakhstan; e-mail: kunsulu.pirmanova@mail.ru;

Serikbayeva, Aizhan Duisenbaiqyzy — PhD student in "Turkology and Language Theory", Al-Farabi Kazakh National University, Almaty, Kazakhstan; e-mail: aizhan.duisenbaiqyzy@gmail.com.