SKASE Journal of Theoretical Linguistics, 2026; 23(1): 209–228
doi: 10.33542/JTL2026-1-11

# Evaluating Corpus Interface Usability for Linguistic Research: The Case of the Kazakh National Corpus

Timur Akishev
KIMEP University, Almaty, Kazakhstan

*This study examines the web interface of the Kazakh National Corpus (KNC) to assess its usability for general corpus-linguistic research. Drawing on established literature, we determine whether the corpus interface supports core analytical techniques, including frequency lists, concordances, keyword lists, and collocations and n-grams. We assess the presence, accessibility, functionality, and overall usability of these tools, identifying areas for improvement. To complement this theory-driven evaluation, we compare the KNC interface with those of two established national corpora, the British National Corpus (BNC) and the Russian National Corpus (RNC), to extract practical insights from their interface design and application of these analytical techniques. This two-level approach allows us to highlight both the theoretical requirements and practical considerations for enhancing the usability of the corpus interface for the purposes of international-standards linguistic research.*

**Keywords**: *linguistic research, corpus linguistics, analytical techniques, corpus tools, corpus interface*

## 1 Introduction

In the modern digital world, language plays an important role as not only the means of communication but also the subject of research at the level of big data. Large-scale studies are conducted worldwide on a multitude of languages in order to understand the specificities of their structure, meaning, and use. The role of language data analysis-focused fields like corpus linguistics, computational linguistics, and natural language processing (NLP) is becoming increasingly important, especially in the era of AI-driven research (Levy et al. 2025). These fields, particularly corpus linguistics, are concerned with analysis of large digital or digitized collections of language data, or corpora. Corpora, classified into multiple types, developed for multiple purposes, and assembled from different types of data, play a key role in determining the characteristics of language in use (Stefanowitsch 2020: 22–23). Unlike general linguistics, which relies on theory-driven, intuitive interpretation of linguistic and sociolinguistic phenomena, corpus linguistics deals with practical analysis of 'living' language data (McEnery & Brookes 2024; Troiani et al. 2024), which includes spoken and written texts and many other configurations of data types that reflect the real-life, contextual use of language.

Historically, one of the key objectives of corpus linguistic research has been the development of large-scale, 'national' corpora for different languages. National corpora are officially created and maintained by national institutions to represent the standard language of a country, often for purposes such as language planning, education, lexicography, and research. Cultural representation and preservation are also important aspects in the design of such large-scale corpora, especially taking into account their role in capturing linguistic variation and supporting the analysis of how language differs and changes across time, regions, and cultural contexts (McEnery & Hardie 2012: 94). In terms of its structure, a national corpus is per se a

general or reference corpus. It is not really a distinct type of corpus in linguistic terms, but rather a designation based on its purpose, scope, and institutional status (Xiao 2008: 383).

The first steps in developing national corpora were taken by the creators of the British National Corpus. A number of other national corpora followed, including the American National Corpus, the Czech National Corpus, the Hungarian National Corpus, the Russian National Corpus, and so forth (Xiao 2008: 384–388). There have also been two significant attempts to develop a large-scale corpus for Kazakh, an agglutinating Turkic language spoken primarily in Kazakhstan and neighboring regions of Central Asia, Russia, and China. These attempts include the Kazakh Language Corpus (Makhambetov et al. 2013), whose website is no longer available, and the Almaty Corpus of the Kazakh Language (2016), which is still available and somewhat queryable. A Kazakh corpus is also available via Sketch Engine (n.d.), based on texts collected from the web, but it is not clear how usable it is for research purposes.

The first large-scale National Corpus of the Kazakh Language (2025) (NCKL; henceforth the Kazakh National Corpus or KNC[1]) was developed at the Akhmet Baitursynuly Institute of Linguistics (n.d.), the national research center focusing primarily on Kazakh and other Turkic languages and operating under the country's Ministry of Science and Higher Education. Since its release to the general public around 2020, the platform has undergone a number of updates that primarily focused on expanding subcorpora. While the KNC grows as a dataset, the publicly available interface, including analytical tools, does not appear to be under further development, while the existing documentation shows little information about past interface changes. The interface currently functions as an access and analysis tool to the corpus data and is presented as part of a research and educational resource, which means that it is already intended for scholarly use. As with many publicly released corpus platforms, potential issues in interface design and analytical functionality may limit researchers' ability to interact effectively with the linguistic data. Addressing these issues requires rigorous research and improvement.

The main purpose of this paper is to evaluate the usability of the web interface of the Kazakh National Corpus for standard corpus-linguistic research by examining whether the interface supports core analytical techniques. We also compare its data analysis and visualization functionality to that of established national corpora for other languages to determine potential areas for improvement. Our goal is not to criticize, but to provide constructive feedback that can help improve the corpus interface. Our suggestions are informed by prominent literature in corpus linguistics, examples of other national corpora and their web interfaces, and our experience in working with corpora and corpus analysis tools.

## 2 Literature review

Despite an increasing number of linguistic projects focusing on Kazakh, it is still considered a low-resource, underresearched language, especially in terms of corpus linguistics and NLP (Joshi et al. 2020). Such languages naturally have a number of issues when it comes to corpus

[1] The official English abbreviation of the name of this corpus is "NCKL." However, some web pages of the online corpus platform, as well as several recent scholarly publications, also use the shorter form "Kazakh National Corpus" and the abbreviation "KNC." In this article, these two forms are used for brevity, while their correspondence to the official name is made explicit here to avoid confusion.

design and maintenance. These issues typically include data scarcity (Agić & Vulić 2019), limited resources (Sekeres et al. 2024), and the lack of progress in the development of tools (Ahmadi et al. 2023).

Recently launched corpora, especially those for low-resource languages, often face a number of issues that typically need to be addressed after launch. Apart from the obvious post-launch evaluations of the size, balance, and representativeness aspects, there are concerns about maintaining the quality and consistency of annotation (Voutilainen 2012), updating metadata and enriching it with NLP tools (Ecker et al. 2024), and improving the accessibility, customizability, and user-friendliness of the corpus interface (Soehn et al. 2008; Machálek 2020), including the development and maintenance of built-in tools to support open research practices corresponding to international standards (Hartmann 2024).

This study draws insights from Evison (2010) and Brezina & Gablasova (2018) to examine the analytical techniques integrated into the main web interfaces of the corpora. The use of guidelines from corpus linguists and developers of multiple corpora ensures that this study and its recommendations are grounded in research. It is important to note that the analytical techniques discussed in this paper are not exclusive to the guidelines developed by these authors. On the contrary, they have long been central to corpus design, and a large number of corpora have been developed on their basis to improve the understanding of language use. What this study borrows, specifically, is the authors' approach to evaluating how these core analytical techniques are implemented in a corpus interface.

*Frequency lists* are a fundamental analytical technique that enables corpus users and researchers to analyze the frequency information of specific linguistic items encountered within a corpus. Frequency analysis can be used to identify dominant lexical items within a dataset (Lijffijt et al. 2011), assess the lexical diversity of a text or corpus (Gregori-Signes & Clavel-Arroitia 2015), and compare word usage patterns across languages, corpora, and subcorpora (Taylor & del Fante 2020), among other purposes. Because frequency information can be sorted and filtered using a web interface of a corpus in various ways, it can serve as a solid foundation for both quantitative and qualitative research projects. However, it should be emphasized that the generation of frequency lists often serves as a foundational step for more complex analytical procedures.

*Concordances*, also known as KWIC (Key Word In Context), are one of the most basic techniques employed in corpus data analysis and visualization. They are used to display the target word or phrase centered, with a fixed number of words shown to its left and right, providing immediate context. This tool can be used to obtain information about the usage and meaning of words and phrases within authentic examples (Hunston 2002: 39). Analyzing how words occur in context contributes to improving the understanding of grammar, vocabulary, and stylistic preferences (Rauscher et al. 2013; Wulff & Baker 2021). While simple in nature, the generation of concordances often underpins more sophisticated forms of linguistic analysis. Concordances can also be a powerful tool in corpus-driven discourse analysis studies (Liu et al. 2022; Nonnenmacher & Naismith 2023). Aside from their applications in linguistics, concordances are widely used in language teaching and translation studies (Jones & Waller 2015; Zanettin 2023).

*Keyword lists* are ranked inventories of lexical items that are identified through frequency analysis as being unusually frequent in a given dataset relative to a reference corpus. This process requires quantitative comparisons between the target corpus, which is a specific dataset that a researcher is analyzing, and a reference corpus, which is a larger, more general

dataset (Rayson & Potts 2021; Moreno-Ortiz 2024). Depending on the research goals and the analytical platform, such comparisons may rely not only on statistical significance measures, but also effect-size measures. By highlighting words that are prominent in relation to a reference corpus, keyword lists help reveal specific patterns and themes in discourse analysis as well as in genre and register studies (Baker 2004). In addition, keyword lists are useful for language instructors aiming to teach specific vocabulary characteristics, translators seeking to simplify their texts (Rayson 2019), and other scholars interested in exploring or comparing domain-specific terminologies (Gries 2021).

*Collocations and n-grams* are two closely related corpus analysis techniques used to investigate how words combine in natural language. Collocations can be used to produce a list of collocates, i.e., co-occurrences of lexical items that have a tendency to be recurrent within a certain span and may be identified using a range of statistical or distributional measures focusing on the strength of association between words (Evert 2008: 1214–1215). N-grams, by contrast, focus on contiguous sequences of words of a specified length (*n*), and are identified on the basis of their frequency of occurrence in a corpus without focusing on association measures between individual words (Lyse & Andersen 2012). Both methods can be useful not only in corpus linguistics but also in language learning research (Gablasova et al. 2017) and general corpus-based exploration of register- and genre-specific language characteristics (Gries et al. 2011).

In the context of this study, these analytical techniques are examined within the web interface of the Kazakh National Corpus, which can serve as a valuable resource for exploring linguistic patterns. Several studies by Kazakhstani linguists, including researchers at the Akhmet Baitursynuly Institute of Linguistics, have examined the Kazakh National Corpus and its interface. These studies have addressed metadata and markup issues (Zhubanov 2015; Slyambekov & Sadyk 2024), the development of various subcorpora within the corpus (Amanbayeva et al. 2025; Fazylzhanova et al. 2025), and its applicability in linguistic research (Pirmanova et al. 2024). However, when they include data analysis, such studies typically rely on specialized standalone software tools to process corpus-elicited data (Ormanova et al. 2025) rather than using the corpus platform-internal, web-based analytical techniques. Importantly, in the present study, the web platforms are not evaluated as substitutes for these standalone corpus analysis applications, but rather as platforms designed to provide direct access to corpus data independently, accompanied by a set of analytical techniques for exploratory purposes.

## 3 Data and methods

The KNC contains 90 million word usages across 20 purpose-specific subcorpora, each hosted on a separate web page. The subcorpora vary in content and size, which reflects the developers' goal of achieving representativeness across different genres, text types, authors, and other domains. The platform contains such subcorpora as Spoken, Educational, Historical, Proverbs and Sayings, Parallel, and Terminological, among others. The web pages hosting these subcorpora provide various interface design and analytical functionality options that could also be evaluated for usability in research. This study, however, focuses only on the main part of the corpus and its interface and analytical functionality.

In order to determine whether the KNC interface is suitable for general corpus-linguistic research, we have compiled a list of core analytical techniques typically employed to work with

large-scale general corpora and evaluated the extent to which this platform supports them. This list of techniques includes (1) frequency lists, (2) concordances, (3) keyword lists, and (4) collocations and n-grams. It is based on a synthesis of ideas from seminal works by Evison (2010) and Brezina & Gablasova (2018). These authors established that these fundamental analytical techniques can be employed either within a corpus using its own web interface features or outside of the corpus using specialized software applications that work well with exported data. The current study aims only to examine whether these techniques are supported by the web interface, excluding discussion of any software applications.

In order to compare the interface of the Kazakh National Corpus to those of established national corpora for other languages in terms of data analysis and visualization functionality, we selected two such corpora, the British National Corpus and the Russian National Corpus, to analyze their features corresponding to the above-mentioned core techniques. These platforms were chosen as benchmarks for this study because they represent widely applied corpus resources that exemplify established interface design and analytical functionality solutions for large-scale national corpora. The British National Corpus (BNC) was selected due to its status as a well-established national corpus. Its availability through multiple platforms, especially BNCweb (CQP-Edition), enables effective querying, data collection, analysis, and visualization (Hoffmann and Evert n.d.). The Russian National Corpus (RNC) was selected due to its comprehensive structure, flexible search mechanisms, and a range of useful features. It provides a robust platform for linguistic research (Savchuk et al. 2024). The main objective of this comparison is to develop recommendations for the developers of the KNC corpus platform based on the insights obtained from the other two platforms.

Importantly, our review focuses only on the main corpus interfaces. Both the KNC and RNC have comprehensive websites that contain multiple pages, including these main corpus interfaces, links to subcorpora, about pages, documentation, and other materials. These websites allow users to explore corpus data, metadata, and design. Most of these features are not available through BNCweb, which is not structured as a comprehensive website with a main corpus interface, additional subcorpora, and other supporting resources. Rather, it is designed as a simplified queryable platform for direct research purposes. This structural difference reflects the fact that the BNC, unlike the KNC and RNC, is hosted by several platforms and has multiple versions. Due to such fragmentation and the absence of a unified website for the BNC, we selected BNCweb as the most user-friendly resource that is diverse in terms of its analytical features. While the BNC is also available via English-Corpora.org (Davies 2004), Sketch Engine (n.d.), and the Oxford Text Archive (OTA) (2007), we excluded these platforms because they require subscription or additional tools. Similarly, we excluded other widely used and updated large-scale corpus platforms, such as the Corpus of Contemporary American English (COCA) (Davies 2008), as they operate under subscription-based access models. In this context, while BNCweb may not represent the most recent or complete version of the BNC, it remains the most accessible and reliable version for the purposes of the current study.

The differing structures and aims of these corpus platforms reflect differences in interface design and intended modes of use, rather than differences in corpus design. Comprehensive national corpora typically aim to represent the entirety of a language in relation to a certain culture, with representativeness and coverage being primary corpus design goals. BNCweb, by contrast, is neither a comprehensive resource nor intended to capture a language as a reflection of a culture. Rather, it functions as an access and analysis interface to an existing

corpus. While such resources may differ in scope and design, the present review deliberately sets corpus-design considerations aside to focus exclusively on interface- and technology-related aspects. To ensure a rigorous evaluation of the usability of the KNC interface for linguistic research, we adopt a two-level approach. Combining theory-driven analysis and practical comparisons across different corpus web interfaces, this approach allows us not only to identify the missing or underdeveloped analytical features in this corpus interface, but also to explain how existing corpus interfaces can serve as models for further interface-related corpus development research.

Our review of the KNC interface was guided by the following criteria and questions, which together constitute the evaluation framework adopted in this study (see Table 1). This framework was developed specifically for this study and is informed by the body of literature in corpus linguistics referenced throughout this paper. Importantly, the same framework was also used to examine BNCweb and the RNC platform, but they were analyzed more generally,[2] with the primary aim of identifying insights from their overall structure, interface design, and analytical functionality.

Table 1: Criteria and guiding questions for reviewing the main KNC interface

| # | Criterion | Guiding question | Description |
|---|---|---|---|
| 1 | Presence or absence | Is this analytical technique available within the main KNC interface? | We determined whether each analytical technique could be located. |
| 2 | Accessibility (visibility and ease of access) | How easily can a corpus user find and use this corpus interface feature without relying on supporting documentation? | We navigated the main corpus interface, and if the techniques and features enabling them were present, we tested how easy they were to find and work with. |
| 3 | Functionality (operational status) | Does the feature work as expected? How well does it work? | We tested each available technique and its associated interface feature to observe whether they operated correctly and efficiently. |
| 4 | Overall usability for research purposes | To what extent do the corpus interface features representing the core analytical techniques support linguistic research? | We synthesized our findings related to the above criteria to determine whether the corpus interface is suitable for linguistic research purposes. |

Regarding the potential shelf-life of this evaluation, it is important to mention that this study focuses on core and stable interface- and analytical feature-related decisions that are typically changed slowly over time, rather than on temporary, cosmetic interface update

[2] For the sake of brevity and comparability, however, the findings for all three platforms are nevertheless presented in a single summary table in the Results and Discussion section.

aspects. Therefore, the observations made in this study may remain relevant over time and be useful not only for future major updates to the KNC interface and analytical tool functionality, but also for other publicly released corpus platforms that may contain related issues.

## 4 Results and discussion

### 4.1 *Introduction*

In most corpus interfaces, analytical techniques typically become available on the query results page, i.e., once a corpus search has been completed. In the RNC interface, for example, they are presented as separate tabs, which users can click to access specific information about a word obtained through different analyses. In BNCweb, similar tools also appear on the query results page, but they can be found and selected using a drop-down menu, which makes them less immediately visible or easy to find. In the KNC interface, by contrast, such tools are not automatically activated upon completing a query. They are often hard to locate, underdeveloped, non-functional, or entirely absent.

In order to provide a systematic evaluation, the results are organized based on a fixed set of interface design and data analysis criteria that were applied to each core analytical technique within each corpus platform: (a) availability within the main interface, (b) accessibility (visibility and ease of access), (c) functionality (operational status and analytical depth), and (d) overall usability for corpus-linguistic research. Each analytical technique is examined according to these criteria in the KNC interface and subsequently compared with its implementation in BNCweb and the RNC interface.

Although the study focuses on the characteristics of the built-in analytical techniques within the main corpus interfaces, it is also important to mention that the KNC platform features a number of tools that are not directly integrated with the corpus-internal data (see Figure 1). These include *Transliteration* for converting Cyrillic Kazakh text into Latin script; *tbisozdik*, a dictionary offering thesaurus-like information; *Orthoepic Converter* for generating orthoepic transcriptions; *Word Frequency*, which generates frequency lists from user-uploaded documents; *Keyboard* for typing Kazakh in Latin script, available as a downloadable application; and several other tools. Some of these tools allow users to paste text or upload text documents directly into the interface, while others require downloading and installation. It is unclear whether this input should be derived from the corpus itself, as there is no built-in functionality that would support exporting query results in text or table format. Importantly, as of December 2025, several of these tools are not operational and appear to function only via the Kazakh-language version of the website.

SKASE Journal of Theoretical Linguistics, 2026; 23(1): 209–228
doi: 10.33542/JTL2026-1-11

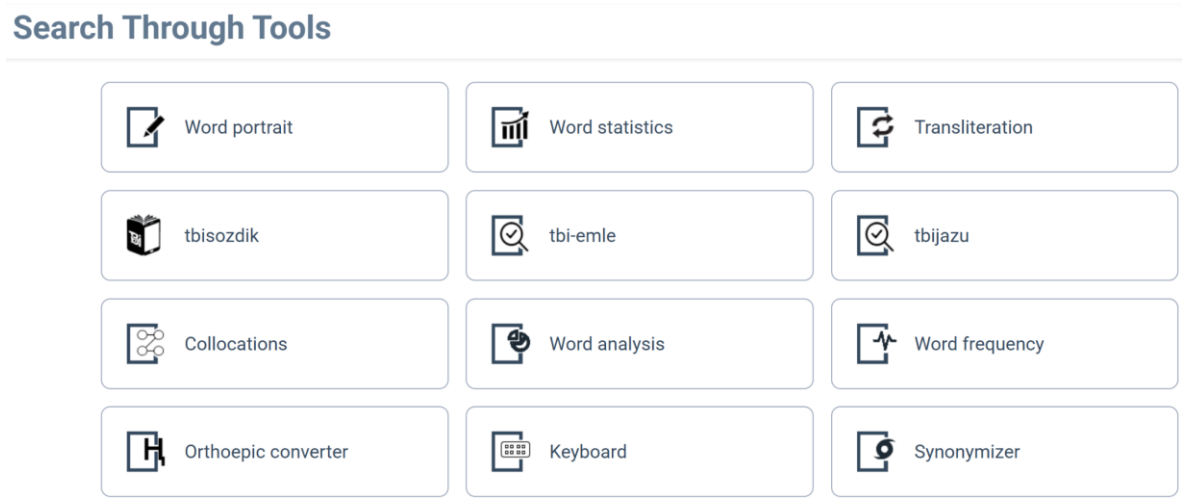


Figure 1: Corpus analysis features of the KNC interface (website accessed June 21, 2025)

It is important to note that the above tools, although web-based, demonstrate functionality similar to popular corpus compilation and analysis software like *AntConc* (Anthony 2024) or *#LancsBox X* (Brezina & Platt 2025) that allow users to upload text documents into their interface to learn about and visualize the data within the documents, and to compile small and large-size corpora on the basis of a collection of such documents. We decided not to include these software corpus tools into our analysis, as our focus is limited to only web-based corpus platforms that process pre-compiled corpus data.

The following discussion focuses on two main objectives: (1) evaluating whether the main KNC interface supports the core analytical techniques by examining their presence or absence, accessibility, functionality, and usability for research purposes; and (2) reviewing the web interface of the RNC and BNCweb to draw useful suggestions for improving the KNC platform based on the characteristics of the analytical techniques available through their main (the RNC) and default (BNCweb) interfaces.

### 4.2 *Scrutiny of the core analytical techniques available within the main KNC interface*

*Word frequency lists*

The main KNC interface does not provide a frequency lists feature on the search results page, so users cannot see which words are most frequent or how many occurrences of a word can be encountered in the corpus or its subcorpora. Instead, there is only a 'Word Statistics' feature that can be accessed by clicking a word in the results, which shows raw frequencies by style, year, author, gender, and topic (see Figure 2). While some of these categories, like style and year, are clearly useful, others seem less relevant. There is no supporting documentation explaining these corpus design choices. Therefore, because the frequency technique is hidden behind additional clicks and lacks corpus-wide and subcorpus-specific distribution information, it is very limited in form and scope and not fully aligned with standard corpus linguistic research practices. In order to make the corpus interface more usable for research purposes, this analytical technique and its associated features should be integrated directly into the search results interface. It should show frequency data comprehensively, so that researchers can access and use this information more intuitively and efficiently.

SKASE Journal of Theoretical Linguistics, 2026; 23(1): 209–228
doi: 10.33542/JTL2026-1-11

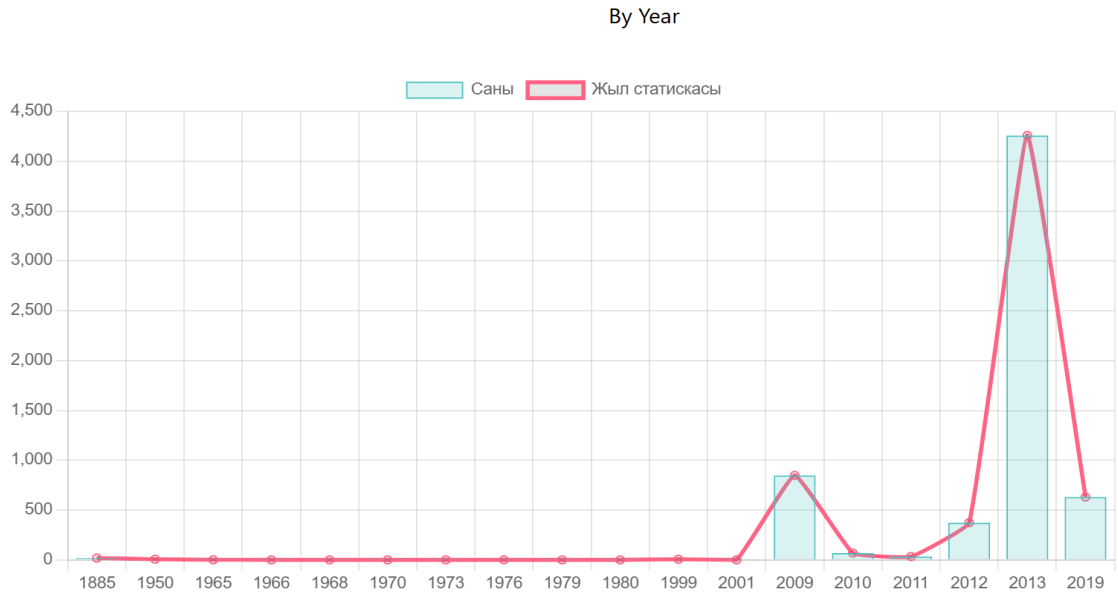


Figure 2: Frequency distribution of *söz* 'word' by year (available via *Word Statistics*)

*Concordances*

This analytical technique is offered in the KNC interface in a very limited form. It is available but consists only of highlighting the search term in red within the results, which are presented as separate paragraphs labeled by author. Because there is no clear KWIC view option, it may not be obvious to corpus users how to view and analyze the concordances efficiently. Furthermore, there is no option to adjust the size of the left and right contexts. The results cannot be aligned or sorted to support systematic analysis. However, it is a positive feature that the query interface provides some options for users to interact with the search results (see Figure 3). Overall, the concordance feature appears to be very limited, and researchers cannot reliably study words in context because of its current limitations. Adding more advanced concordance settings, such as sorting, aligning, filtering, adjusting context size, and export options, would create more opportunities for systematic analysis, making this analytical technique and its associated web interface features more suitable for linguistic research.

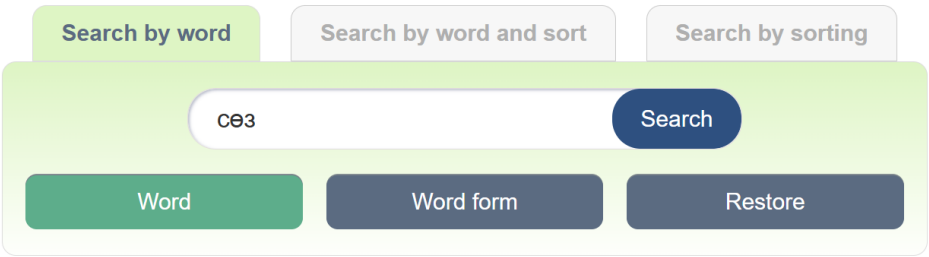


Figure 3: The query interface of the main part of the KNC platform

*Keyword lists*

In the KNC interface, this analytical technique is currently unavailable. The platform does not support the extraction of keyword lists or comparison functions. As a result, corpus users

cannot identify and compare keywords or statistically prominent words in relation to a reference corpus or across different subcorpora of the KNC. It is also not possible to highlight specific vocabulary characteristic of particular genres or registers in the corpus. This technique and its associated web interface features, if implemented, would work best for cross-corpus or cross-subcorpus comparisons. This tool would facilitate more corpus linguistic research aimed at identifying the distinctive lexical patterns, which would help researchers analyze variation more efficiently.

*Collocations and n-grams*

These techniques are currently unavailable in the main KNC interface. The tool labeled 'Collocations' does not retrieve collocates (see Figure 4). Instead, it performs simple searches within user-provided text, with unclear methods and no documentation to explain its use. This tool should not only be considered limited in accessibility and functionality, but also separate from the corpus platform itself, since it does not rely on the corpus-internal data but rather on text that users are invited to paste into its interface. As for n-grams, there are also no features supporting their extraction and analysis. Overall, the usability of these tools for research is minimal. If implemented transparently, they could facilitate corpus-linguistic research on recurring patterns and lexical combinations.

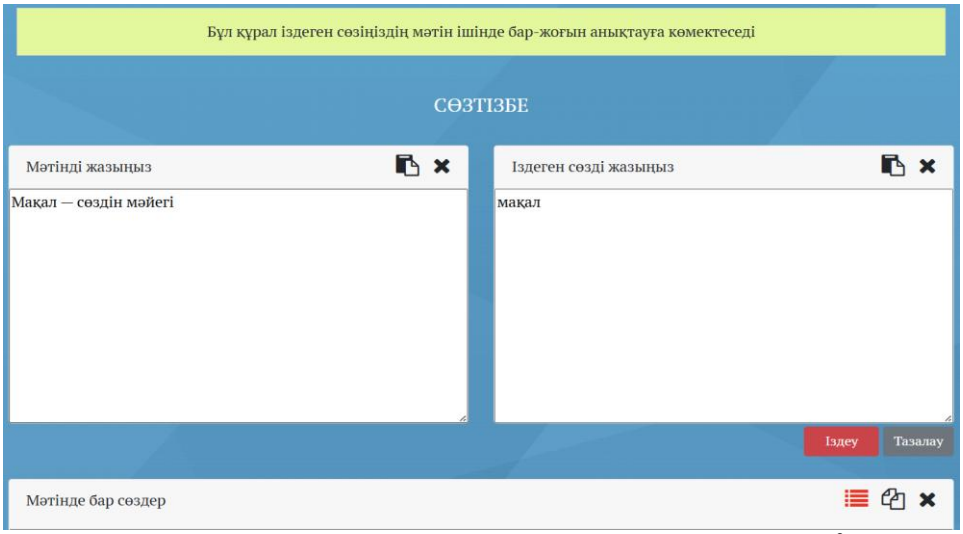


Figure 4: The Collocations feature of the KNC platform[3]

4.3 *Drawing insights from BNCweb and the RNC interface: A cross-corpus comparison*

While the previous subsections focus on reporting the observed characteristics of the analytical techniques available through corpus interfaces, this subsection adopts a more comparative and interpretive perspective, relying on the findings obtain in relation to BNCweb and the RNC interface as reference points to contextualize the findings related to the KNC.

[3] Note that the detailed description of this tool is provided in the text above. This figure is provided here only for visualization purposes. It has text in Kazakh only because there are no other language options provided on the web page hosting this corpus tool.

By conducting a similar review of the RNC interface and BNCweb, we have found that most of the above-mentioned analytical techniques and their associated features are available within their online interfaces. However, for reasons of space, we are not providing full, detailed overviews of the two corpus platforms in a similar way as we have for the KNC. Instead, we briefly summarize some aspects of the analytical techniques that they provide, and then compare all three corpus platforms in table format. Importantly, it was not our goal to determine whether one corpus platform can be used for more research projects than the others. The three resources under study contain different data obtained from different sources and are not comparable in terms of the recency of their materials or how regularly the materials are updated or corpus updates are performed. It should be understood that such considerations are beyond the scope of this study.

The RNC interface supports all the core analytical techniques under study, and BNCweb fully supports most of them. These techniques can be used to collect, analyze, and visualize corpus data to answer a wide range of research questions. However, in both corpus platforms, some features are not always easy to find and may require additional effort to learn to use for general corpus querying and research.

The frequency lists technique is fully integrated into the main RNC interface, allowing users to generate frequency lists and explore the distribution of a search term across multiple categories (author, gender, domain, text type, theme, genre, style, text category). Data visualization tools such as the *Graph* (or distribution by date) feature produce visual representations of frequency data over time. All such information is based on transparent statistical analysis, with relevant statistical measures and scores provided in the output. In BNCweb, the frequency technique is provided with a detailed frequency breakdown and distribution information by multiple categories. It is possible to compare frequency lists, especially across the spoken and written components of the corpus platform. As for concordances, the RNC's main interface separates concordance results from KWIC. Although both display instances of use of a search term in context, they serve different purposes. Various settings allow corpus users to adjust concordance and KWIC views, with some sorting options available exclusively for KWIC. In contrast, BNCweb integrates KWIC directly with concordancing results, and there is no distinction between these aspects. Regarding keyword lists, BNCweb provides a fully functional keyword analysis feature. This feature allows users to directly compare items and lists of items pertaining to the spoken and written components. The RNC platform, by contrast, does not have an easily accessible and intuitive keyword lists function within its main interface. This technique and its associated interface features may either be available under a different name or require an alternative method of access. The process of generating keyword lists may also be more manual within the RNC's interface than in BNCweb. However, it is worth mentioning that more intuitive and advanced keyword list generation is generally performed using standalone desktop software like *AntConc* or *#LancsBox X*, rather than relying solely on the web interface of a corpus. Regarding collocations and n-grams, the RNC platform supports both analytical techniques in a research-ready form, allowing corpus users to perform the procedures directly within its main interface. While the n-grams function is available as a series of clickable tabs based on the size of the n-gram, the collocations tool is slightly more difficult to locate within the corpus interface. Still, both analytical techniques and their features are presented in a straightforward manner, with supporting documentation, tips, and examples of use. By contrast, BNCweb includes a fully developed collocations feature, but it does not provide an n-grams tool within its interface.

The following table synthesizes the results of the systematic evaluation by summarizing the availability and research readiness of each analytical technique across the three platforms.

Table 2: Comparative evaluation of analytical techniques across corpus web interfaces

| # | Corpus platform | Frequency lists | Concordances | Keyword lists | Collocations and n-grams | Additional data visualization |
|---|---|---|---|---|---|---|
| 1 | KNC | Very limited | Very limited | Unavailable | Unavailable | Available, but limited |
| 2 | BNCweb | Available and research-ready | Available and research-ready | Available and research-ready | BNCweb has only collocations. This tool is research-ready. | Available, but limited |
| 3 | RNC | Available and research-ready | Available and research-ready | Available, but not intuitive or easily accessible. | Both are available and research-ready. | Available and research-ready |

4.4 *Implications and recommendations for improving the KNC interface*

The following categorized list of suggestions is based on our review of the corpus platforms. These suggestions do not all pertain to the analytical techniques and their associated web interface features. Some of them arose during the corpus platform review procedures and also reflect more general aspects pertaining to enhancing the search and annotation systems and ensuring the transparency of corpus interface design and development of data analysis functionality.

*Analytical techniques and their features*
(1) Introduce new analytical techniques that match those found in established corpus platforms;
(2) Redesign existing techniques and their features so they can work with the corpus data itself, reducing dependence on desktop corpus tools;
(3) Enhance data export options to make it easier for researchers to extract the data they need;

*Annotation and querying*
(4) Ensure that the corpus and its subcorpora are consistently structured, annotated, and tagged to support uninterrupted querying and advanced analytical procedures;
(5) Integrate more advanced linguistic settings into the corpus query tool to make corpus querying more flexible;

*Corpus platform navigation and data presentation*
(6) Improve navigation between subcorpora and ensure consistency in how analytical tools and interface features work across the platform;
(7) Fix broken links and ensure that all interface aspects operate as intended;
(8) Improve how data is categorized and presented to make interacting with it more convenient;

*Open research practices and community engagement*
(9) Maintain a log of corpus platform updates so external linguists can understand internal procedures and offer feedback;
(10) Share information about the corpus and corpus platform with Kazakhstani and international researchers to encourage collaboration;
(11) Ground all further corpus and corpus platform developments and updates in current theory and practice in the field of corpus linguistics.

These suggestions are important to implement for a variety of reasons, including but not limited to (1) making the corpus platform and its interface more usable for high-quality and international-standards linguistic research (Artetxe et al. 2022); (2) improving the ability of the corpus platform to accurately represent the language-in-use data in a more detailed way (Hurtado Bodell et al. 2022); (3) enabling more sophisticated forms of linguistic analysis, including topics like syntactic parsing, code-switching, and sociolinguistic variation (Rybka et al. 2015; Çetinoğlu 2016; Baker & Heritage 2022; Akishev & Kravtsova 2024; Miletic et al. 2024); (4) facilitating cross-linguistic comparisons based on corpus data (Stave et al. 2022); and (5) increasing the usability of the corpus platform for interdisciplinary projects including natural language processing, education, and language planning (Ilao 2017; Almujaiwel 2018; Alsop et al. 2020). It is our understanding that these improvements can contribute to transforming the corpus platform from a simple data repository to a high-level research resource.

## 5 Conclusion

This study has focused on the discussion of the usability of the web interface of the Kazakh National Corpus for research purposes. Our analysis has highlighted several ways in which the interface and architecture of the corpus could be significantly improved. The analytical techniques discussed here are widely recognized as essential components of any corpus platform. To provide more research opportunities, the KNC platform should evolve into a highly queryable resource whose interface and architecture would contain extensive built-in features for researchers to engage with and explore corpus data efficiently and intuitively. Additionally, further research is needed to address potential issues related to the internal organization of the corpus, data quality, data preprocessing and annotation, and, importantly, the deployment of the corpus on a modern, accessible web platform.

One of the limitations of this study is the sole focus on the core analytical techniques, especially in terms of the comparison between the selected corpus platforms. We decided to step away from the discussion of differences in terms of content, size, source of data, number and types of subcorpora, and other aspects of corpus compilation, corpus annotation, and corpus architecture. While we acknowledge that it would be possible to provide additional suggestions based on these aspects, they are simply beyond the scope of this paper. Further research can focus on these dimensions of the architecture and usability of the KNC platform.

The limitation of our corpus review process is the decision to focus solely on the main corpus interfaces. This decision applies to the KNC and RNC platforms, but it is less applicable to BNCweb, whose interface separates Spoken and Written parts but lacks a comprehensive collection of subcorpora. Including subcorpora in our analysis would be problematic, as the KNC and RNC have different types and numbers of subcorpora, and BNCweb does not offer subcorpora in the same way. Additionally, it was difficult to determine which version of the BNC should be considered the most relevant and reliable. The BNC is also hosted by other platforms, including English-Corpora.org, Sketch Engine, and the OTA. Each of these platforms offers a different version of the BNC, different levels of accessibility, and different functionalities. Furthermore, identifying the most recent or complete version of the BNC is difficult. Therefore, we chose to analyze only the web-accessible, default (BNCweb) and main (the KNC and RNC) interfaces of the corpora. We understand that further analysis of subcorpora and platform-based features may offer valuable insights and represent an important direction for future work.

In closing, it is important to highlight the scarcity of sources that specifically focus on improving the functionality and applicability for research purposes, especially in the case of recently launched corpus platforms. It is possible to assume that such considerations are discussed internally during corpus design, without the need to document them in research papers. Consequently, developers of corpora for low-resource languages do not always have access to specific guidelines on corpus compilation, architecture, and interface design. There is a substantial amount of research on corpus building, but it is mostly fragmented and outdated, offering few generalizable ideas for corpus developers to use when designing and maintaining their corpora. As a result, corpora of different types are still developed by linguists, educational institutions, and government bodies around the world without following any unified guidelines. Developing and sharing such guidelines would be an important first step in ensuring that corpora and corpus interfaces for all languages, including the low-resource ones, are based on the international standards of corpus building and usable for high-quality linguistic research.

## Acknowledgments

The author thanks two anonymous reviewers whose detailed feedback helped significantly improve this paper.

## References

Agić, Željko & Vulić, Ivan. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3204–3210. Association for Computational Linguistics. https://doi.org/10.18653/v1/p19-1310

Ahmadi, Sina & Azin, Zahra & Belelli, Sara & Anastasopoulos, Antonios. 2023. Approaches to corpus creation for low-resource language technology: The case of Southern Kurdish and Laki. In *Proceedings of the Second Workshop on NLP Applications to Field Linguistics*, 52–63. Association for Computational Linguistics. https://aclanthology.org/2023.fieldmatters-1.7/

Akhmet Baitursynuly Institute of Linguistics. 2025. *National Corpus of the Kazakh Language*. (https://qazcorpus.kz/indexen.php) (Accessed 2025-06-23)

Akishev, Timur & Kravtsova, Yekaterina. 2024. A corpus-based analysis of intrasentential code-switching. In *Proceedings of the 5th International Conference "Language. Text. Society"*, 3–4. https://www.hse.ru/data/2024/10/18/1940216934/LTS-2024%20Proceedings.pdf

Al-Farabi Kazakh National University. 2016. *Almaty Corpus of the Kazakh Language*. http://web-corpora.net/KazakhCorpus/search/ (Accessed 2025-06-23)

Almujaiwel, Sultan. 2018. Integrating NLP with corpus linguistics and vice versa. In *Proceedings of the International Conference on Learning and Optimization Algorithms: Theory and Applications*, 1–6. New York: Association for Computing Machinery.

Alsop, Siân & King, Virginia C. & Giaimo, Genie & Xu, Xiaoyu. 2020. Uses of corpus linguistics in higher education research: An adjustable lens. In Huisman, Jeroen & Tight, Malcolm (eds.), *Theory and method in higher education research*, 21–40. Bingley: Emerald Publishing Limited.

Amanbayeva, Aisaule & Zhumabayeva, Zhanar & Bazarbayeva, Zeinep & Fazylzhanova, Anar & Ospangaziyeva, Nazgul. 2025. National Corpus of the Kazakh Language: Prosodic features of the poetic discourse. *Forum for Linguistic Studies* 7(1). 505–519. https://doi.org/10.30564/fls.v7i1.7428

Anthony, Laurence. 2024. *AntConc* (Version 4.3.1) [computer software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software/AntConc

Artetxe, Mikel & Aldabe, Itziar & Agerri, Rodrigo & Perez-de-Viñaspre, Olatz & Soroa, Aitor. 2022. Does corpus quality really matter for low-resource languages? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 7383–7390. Abu Dhabi: Association for Computational Linguistics. https://aclanthology.org/2022.emnlp-main.499/

Baker, Paul. 2004. Querying keywords: Questions of difference, frequency, and sense in keywords analysis. *Journal of English Linguistics* 32(4). 346–359. https://doi.org/10.1177/0075424204269894

Baker, Paul & Heritage, Frazer. 2022. How to use corpus linguistics in sociolinguistics: A case study of modal verb use, age and change over time. In O'Keeffe, Anne & McCarthy, Michael J. (eds.), *The Routledge handbook of corpus linguistics*, 562–575. London: Routledge.

BNC Consortium. 2007. *British National Corpus, World edition*. Oxford Text Archive. http://hdl.handle.net/20.500.14106/2552 (Accessed 2025-06-23)

Brezina, Vaclav & Gablasova, Dana. 2018. The corpus method. In Culpeper, Jonathan V. & Kerswill, Paul & Wodak, Ruth & McEnery, Tony & Katamba, Francis (eds.), *English language: Description, variation and context* (2nd ed.), 595–609. Basingstoke: Palgrave Macmillan.

Brezina, Vaclav & Platt, William. 2025. *#LancsBox X* [computer software]. Lancaster University. http://lancsbox.lancs.ac.uk

Çetinoğlu, Özlem. 2016. A Turkish-German code-switching corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 4215–4220. European Language Resources Association (ELRA).

Davies, Mark. 2004. *British National Corpus*. Oxford University Press. https://www.english-corpora.org/bnc/

Davies, Mark. 2008. *The Corpus of Contemporary American English (COCA)*. Available online at https://www.english-corpora.org/coca/

Ecker, Jennifer & Fischer, Stefan & Schwarz, Pia & Trippel, Thorsten & Werthmann, Antonina & Wilm, Rebecca. 2024. Unlocking the corpus: Enriching metadata with state-of-the-art NLP methodology and linked data. In *Proceedings of the CLARIN Annual Conference 2024*. https://www.clarin.eu/sites/default/files/CLARIN2024_ConferenceProceedings_final.pdf

Evert, Stefan. 2008. Corpora and collocations. In Lüdeling, Anke & Kytö, Merja (eds.), *Corpus linguistics: An international handbook* vol. 2, 1212–1248. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110213881.2.1212

Evison, Jane. 2010. What are the basics of analysing a corpus? In O'Keeffe, Anne & McCarthy, Michael J. (eds.), *The Routledge handbook of corpus linguistics*, 122–135. London: Routledge.

Fazylzhanova, Anar & Seitbekova, Ainur & Kobdenova, Gulzhihan & Seidamat, Asel & Ayazbayev, Galymzhhan. 2025. The issues of developing the historical subcorpus of the National Corpus of the Kazakh Language. *Lodz Papers in Pragmatics* 21(1). 169–191. https://doi.org/10.1515/lpp-2024-0038

Gablasova, Dana & Brezina, Vaclav & McEnery, Tony. 2017. Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning* 67. 155–179. https://doi.org/10.1111/lang.12225

Gregori-Signes, Carmen & Clavel-Arroitia, Begoña. 2015. Analysing lexical density and lexical diversity in university students' written discourse. *Procedia-Social and Behavioral Sciences* 198. 546–556. https://doi.org/10.1016/j.sbspro.2015.07.477

Gries, Stefan. Th. & Newman, John & Shaoul, Cyrus. 2011. N-grams and the clustering of registers. *Empirical Language Research Journal* 5(11).

Gries, Stefan Th. 2021. A new approach to (key) keywords analysis: Using frequency, and now also dispersion. *Research in Corpus Linguistics* 9(2). 1–33. https://doi.org/10.32714/ricl.09.02.02

Hartmann, Stefan. 2024. Open Corpus Linguistics–or How to overcome common problems in dealing with corpus data by adopting open research practices. In Kaunisto, Mark & Schilk, Marco (eds.), *Challenges in Corpus Linguistics: Rethinking corpus compilation and analysis*, 89–105. Amsterdam: John Benjamins Publishing Company.

Hoffmann, Sebastian & Evert, Stefan. (n.d.). *BNCweb (CQP-edition)*. Lancaster University. http://bncweb.lancs.ac.uk/ (Accessed 2025-06-23)

Hurtado Bodell, Miriam & Magnusson, Måns & Mützel, Sophie. 2022. From documents to data: A framework for total corpus quality. *Socius* 8. https://doi.org/10.1177/23780231221135523

Hunston, Susan. 2002. Methods in corpus linguistics: Interpreting concordance lines. In Susan Hunston (ed.), *Corpora in Applied Linguistics*, 38–66. Cambridge: Cambridge University Press.

Ilao, Joel. 2017. Corpus linguistic analysis for language planning. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, 12–12. The National University (Philippines).

Jones, Christian & Waller, Daniel. 2015. *Corpus linguistics for grammar: A guide for research*. London: Routledge.

Joshi, Pratik & Santy, Sebastin & Budhiraja, Amar & Bali, Kalika & Choudhury, Monojit. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6282–6293. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.560

Levy, Roger & Kim, Yoon & & Fox, Danny. 2025. The science of language in the era of generative AI. *An MIT Exploration of Generative AI*. https://doi.org/10.21428/e4baedd9.f6a0052d

Lijffijt, Jefrey & Papapetrou, Panagiotis & Puolamäki, Kai & Mannila, Heikki. 2011. Analyzing word frequencies in large text corpora using inter-arrival times and bootstrapping. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 341–357. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-23783-6_22

Liu, Yanhong & Zhang, Lawrence J. & Yang, Li. 2022. A corpus linguistics approach to the representation of Western religious beliefs in ten series of Chinese university English language teaching textbooks. *Frontiers in Psychology* 12. https://doi.org/10.3389/fpsyg.2021.789660

Lyse, Gunn. I. & Andersen, Gisle. 2012. Collocations and statistical analysis of n-grams: Multiword expressions in newspaper text. In Gisle Andersen (ed.), *Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian*, 79–110. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/scl.49.05lys

Machálek, Tomáš. 2020. KonText: Advanced and flexible corpus query interface. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 7003–7008. European Language Resources Association. https://aclanthology.org/2020.lrec-1.865/

Makhambetov, Olzhas & Makazhanov, Aibek & Yessenbayev, Zhandos & Matkarimov, Bakhyt & Sabyrgaliyev, Islam & Sharafudinov, Anuar. 2013. Assembling the Kazakh Language Corpus. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1022–1031. Association for Computational Linguistics. https://aclanthology.org/D13-1104/

McEnery, Tony & Hardie, Andrew 2012. *Corpus linguistics: Method, theory and practice*. Cambridge, UK: Cambridge University Press.

McEnery, Tony & Brookes, Gavin. 2024. Corpus linguistics and the social sciences. *Corpus Linguistics and Linguistic Theory* 20(3). 591–613. https://doi.org/10.1515/cllt-2024-0036

Miletic, Filip & Przewozny-Desriaux, Anne & Tanguy, Ludovic. 2024. Modeling fine-grained sociolinguistic variation: The promises and pitfalls of Twitter corpora and neural word embeddings. In Kaunisto, Mark & Schilk, Marco (eds.), *Challenges in Corpus Linguistics: Rethinking corpus compilation and analysis*, 142–170. Amsterdam: John Benjamins Publishing Company.

Moreno-Ortiz, Antonio. 2024. Keywords. In Antonio Moreno-Ortiz (ed.), *Making Sense of Large Social Media Corpora*, 59–102. Cham: Palgrave Macmillan. https://doi.org/10.1007/978-3-031-52719-7_4

Nonnenmacher, Sean & Naismith, Ben. 2023. 'Her dreadful plight': A corpus-assisted analysis of the indexical and stance properties of *poor thing*. *Journal of Corpora and Discourse Studies* 6(1). 61–90. https://doi.org/10.18573/jcads.90

Ormanova, Assel. B. & Anafinova, Madina L. & Ospanova, Dana Zh. & Tleshova, Zhibek K. 2025. A corpus-based approach in vocabulary research: Defining the Word of the Year 2023 in Kazakh. *Theory and Practice in Language Studies* 15(3). 677–687. https://doi.org/10.17507/tpls.1503.02

Pirmanova, Kunsulu K. & Tokmyrzayev, Darkhan O. & Pirmanova, A. K. 2024. Application of the National Corpus of the Kazakh Language in linguistic research. *Journal of Ecohumanism* 3(7). 2806–2814. https://doi.org/10.62754/joe.v3i7.4418

Rauscher, Janneke & Swiezinski, Leonard & Riedl, Martin & Biemann, Chris. 2013. Exploring cities in crime: significant concordance and co-occurrence in quantitative literary analysis. In *Proceedings of the Workshop on Computational Linguistics for Literature*, 61–71. Association for Computational Linguistics. https://aclanthology.org/W13-1409/

Rayson, Paul. 2019. Corpus analysis of key words. In Carol A. Chapelle (ed.), *The concise encyclopedia of applied linguistics*, 320–326. Hoboken: John Wiley & Sons.

Rayson, Paul & Potts, Amanda. 2021. Analysing keyword lists. In Paquot, Magali & Gries, Stefan Th. (eds.), *A practical handbook of corpus linguistics*, 119–139. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_6

Rybka, Roman & Sboev, Alexander & Moloshnikov, Ivan & Gudovskikh, Dmitry. 2015. Morpho-syntactic parsing based on neural networks and corpus data. In *2015 Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference*, 89–95. IEEE.

Savchuk, Svetlana O. & Arkhangelskiy, Timofey A. & Bonch-Osmolovskaya, Anastasiya A. & Donina, Ol'ga V. & Kuznetsova, Yuliya N. & Lyashevskaya, Ol'ga N. & Orekhov, Boris V. & Podryadchikova, Mariya V. 2024. *Natsional'nyy korpus russkogo yazyka 2.0: Novye vozmozhnosti i perspektivy razvitiya* (Russian National Corpus 2.0: New possibilities and development prospects). *Voprosy Jazykoznanija* 2. 7–34. https://edgccjournal.org/0373-658X/article/view/650588

Sekeres, Hedwig G. & Heeringa, Wilbert & de Vries, Wietse & Zwagers, Oscar Yde & Wieling, Martijn & Jensma, Goffe Th. 2024. Developing infrastructure for low-resource language corpus

building. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, 72–78. Torino, Italia: ELRA and ICCL. https://aclanthology.org/2024.sigul-1.10/

Sketch Engine. n.d. *British National Corpus*. https://www.sketchengine.eu/british-national-corpus-bnc/ (Accessed 2025-06-23)

Sketch Engine. n.d. kkWaC: *Kazakh corpus from the web*. https://www.sketchengine.eu/kkwac-kazakh-corpus/ (Accessed 2025-06-23)

Slyambekov, Qymbat. B. & Sadyk, Ainagul M. 2024. The National Corpus of the Kazakh Language: The semantic markup of verbs. *Tiltanym* 1. 189–196. https://doi.org/10.55491/2411-6076-2024-1-189-196

Soehn, Jan-Philipp & Zinsmeister, Heike & Rehm, Georg. 2008. Requirements of a user-friendly, general-purpose corpus query interface. In *Proceedings of the LREC 2008 Workshop Sustainability of Language Resources and Tools for Natural Language Processing*, 27–32. http://nbn-resolving.de/urn:nbn:de:bsz:352-opus-96897

Stave, Matthew & Delafontaine, François & Pellegrino, François & Coupé, Christophe. 2022. How comparable are languages across linguistic corpora? Some methodological thoughts. In *ALT 2022 - 14th Conference of the Association for Linguistic Typology.* Austin: United States. https://hal.science/hal-03900052

Stefanowitsch, Anatol. 2020. *Corpus linguistics: A guide to the methodology*. Language Science Press.

Taylor, Charlotte & del Fante, Dario. 2020. Comparing across languages in corpus and discourse analysis: some issues and approaches. *Meta* 65(1). 29–50. https://doi.org/10.7202/1073635ar

Troiani, Giorgia & Du Bois, John W. & Filchenko, Andrey. 2024. Corpus as a slice of life: Representing naturally occurring language and its speakers. *Research in Corpus Linguistics* 12(2). 174–202.

The Institute of Linguistics named after Akhmet Baitursunuly. (n.d.). Official website. The Committee of Science of the Ministry of Science and Higher Education, Republic of Kazakhstan. https://tbi.kz/eng

Voutilainen, Atro. 2012. Improving corpus annotation productivity: A method and experiment with interactive tagging. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, 2097–2102. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/393_Paper.pdf

Wulff, Stefanie & Baker, Paul. 2021. Analyzing concordances. In Paquot, Magali & Gries, Stefan Th. (eds.), *A practical handbook of corpus linguistics*, 161–179. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_8

Xiao, Richard. 2008. Well-known and influential corpora. In Lüdeling Anke & Kytö, Merja (eds.), *Corpus linguistics: An international handbook* vol. 1, 383–457. Berlin: Mouton de Gruyter.

Zanettin, Federico. 2023. Concordancing. In Chan Sin-wai (ed.), *Routledge encyclopedia of translation technology*, 498–511. New York: Routledge.

Zhubanov, Askar. K. 2015. National Corpus of the Kazakh Language and metamarking problems. *Tiltanym* 1. 21–29.

*Timur Akishev*
*Department of Linguistics and Cognitive Science*
*College of Human Sciences and Education*
*KIMEP University*
*Almaty, Kazakhstan*
*e-mail: t.akishev@kimep.kz*

*In SKASE Journal of Theoretical Linguistics [online]. 2026, vol. 23, no. 1 [cit. 2026-06-30]. Available on web page http://www.skase.sk/Volumes/JTL61/11.pdf. ISSN 1336-782X*