Issue 16, 2024

УДК 004.9:81’322

INFORMATION SYSTEMS FOR WORKING WITH TEXT CORPORA: CLASSIFICATION AND COMPARATIVE ANALYSIS

Ivan Kozak[1], Nataliia Kunanets[2]

[1-2] Lviv Polytechnic National University,
Information Systems and Networks Department, Lviv, Ukraine
[1] E-mail: ivan.kozak.lp@gmail.com, ORCID: 0009-0007-4953-2816
[2] E-mail: Nataliia.E.Kunanets@lpnu.ua, ORCID: 0000□0003□3007□2462

The article examines information systems for working with text corpora, particularly their application for linguistic analysis and management of large text data. Information systems for supporting text corpora are analyzed, classified, and compared based on their historical development and functional capabilities. The main focus is comparing the two most common systems that can be distinguished by functionality as corpus managers: “AntConc” and “Sketch Engine”. These are evaluated based on key criteria: corpus creation, text processing, annotation, storage and export, data analysis and visualization, interface intuitiveness, support for the Ukrainian language, as well as the presence of an open license. The research aimed to conduct a comparative analysis of these systems using the analytic hierarchy process method to determine their strengths and weaknesses under different usage conditions. It was found that “Sketch Engine” provides advanced capabilities for creating and managing large corpora, annotating and visualizing data, making it a better choice for large research projects. At the same time, “AntConc” is a more accessible and efficient system for individual or small-scale research due to its simplicity, lack of licensing costs, and support for specific parameters for text analysis. The research findings can be useful for corpus and applied linguists when choosing systems for creating and working with text corpora. The conclusions will contribute to making decisions regarding the selection of appropriate tools based on specific research needs, workload, and budget constraints. In addition, the research results can be applied to improving existing and developing new information systems to support corpora in future scientific projects by the authors.

Keywords: corpus linguistics, corpus manager, AntConc, Sketch Engine, analytic hierarchy process method.

Introduction and problem statement

In the current context of digital technology development, information systems play a crucial role in organizing work with text corpora, facilitating the automation and optimization of large-scale data analysis. Such systems are indispensable tools for applied linguists as they assist in selecting linguistic material, preparing it for corpus inclusion, organizing texts into corpora and subcorpora, and managing them. Moreover, these systems enable deep analysis of linguistic material across various dimensions, extraction of relevant data, and dissemination for further use in scientific and educational purposes.

Despite significant progress in the development of information systems for working with text corpora, the problem of selecting the most optimal system for specific research tasks still persists. There is a wide range of such systems, each with its own features, purposes, functional capabilities, and limitations. Therefore, it is crucial to compare these systems to determine their suitability for researchers’ needs, which requires a clear classification and comparative analysis.

Analysis of recent studies and publications

Recent research in the field of information systems for working with corpora (including corpus managers, concordancers, tagging systems, etc.) demonstrates significant progress in the development of tools for processing and analyzing linguistic data–from basic functionalities such as KWIC (Key Word in Context) to advanced comprehensive information systems [2].

The latest generations of information systems are often implemented using cloud and web technologies to ensure efficient processing of large text volumes, as well as to provide user-friendly and intuitive interfaces [5]. Given the growing number of information systems for working with corpora, researchers [1] aim to classify these systems and identify their main differentiating characteristics.

The research focuses on improving corpus management capabilities, enhancing search functions for researchers [8], and developing data visualization tools [11]. Special attention is given to the use of generative artificial intelligence as an analyzer for corpus information systems [4].

Scholars are examining the potential applications of corpus systems in bilingual lexicography [13], specialized translation [6], foreign language learning [10], and cross-linguistic comparison of keyword frequency [9].

Formulation of the research problem and article aim

The object of this study is information systems designed for working with text corpora, while **the subject** is their classification and comparison based on functional capabilities and purpose.

The work aims to analyze and classify the most popular information systems for working with text corpora, specifically AntConc and Sketch Engine, as well as to compare their key characteristics using the analytic hierarchy process.

To achieve this aim, the following **research tasks** are defined:

1. To classify information systems for working with text corpora.
2. To describe and analyze in detail the two most common systems: AntConc and Sketch Engine.
3. To identify the main parameters by which these systems can be compared and conduct a comparative analysis, highlighting key advantages and disadvantages.
4. To apply the analytic hierarchy process for comparing AntConc and Sketch Engine.

Presentation of the results

Comparative analysis of information systems for text corpus management

Before performing the comparison of existing systems for text corpus management, it is essential to understand the approaches to their development and application, and consequently, the principles of their classification. There are several approaches to the classification of information systems for working with corpora (see Fig. 1).

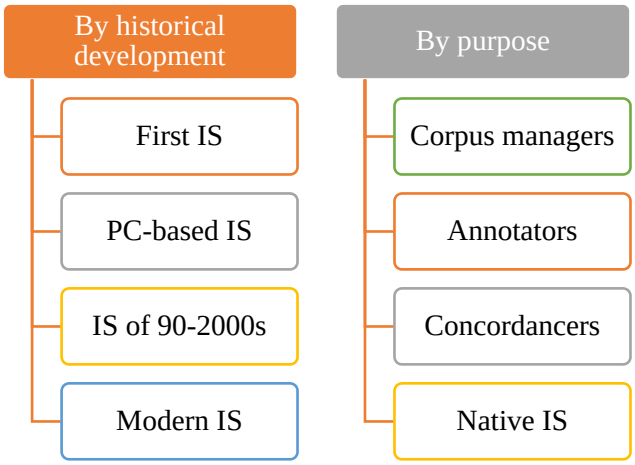

Fig. 1. Classification scheme of information systems for working with corpora

One of the most common classifications is based on the historical development and functional and non-functional capabilities of such programs [2], which allows us to identify four types of information systems:

1. The first information systems for working with corpora, developed in the 1960s and 1970s. A distinctive feature of these systems was the limitation to standard ASCII characters, restricting their functionality primarily to the English language and offering limited features–mostly quantitative word analysis in texts and KWIC (Key Word in Context).

2. The second generation of information systems, functionally differed little from the first, except for the ability to install them on personal computers. This advancement allowed for a significant breakthrough in corpus linguistics, enabling researchers to work on individual corpus projects.

3. The generation of information systems that began in the late 1990s to early 2000s. These systems positively differ from previous generations as they allow work not only with the English language, possess expanded functional capabilities, and utilize statistical approaches for information processing. However, they have certain limitations regarding the volume of processed texts, as they are mostly installed on personal computers or small on-premise servers.

4. The last and newest generation of information systems for working with text corpora emerged and began to rapidly develop with the growing popularity and accessibility of cloud environments. These information systems are mostly hosted in the cloud and feature user-friendly web interfaces, making them accessible and understandable for most researchers, educators, and even novice enthusiasts while having serious capabilities for processing large data sets. At the same time, Anthony [2] points out the disadvantages of such systems, including closed code, commercialization, and limited options for choosing a cloud or adapting product behavior.

Among the types mentioned above, we focus our attention on the last two generations of information systems, as they are relevant and currently in use. The most widely used representative of the third generation of information systems is AntConc, while Sketch Engine represents the fourth generation.

Another classification scheme that deserves attention is based on the primary purpose of such information systems. According to this scheme, we can identify the following types of information systems for working with linguistic corpora:

1. Corpus managers (or as noted by Abdullayeva [1] – “text compilers”) are information systems designed to assist users in creating and managing corpora. Typically, their main functionality includes creating a linguistic corpus from a set of texts, text archives, or web resources, as well as supporting it through expansion, modification, and the addition of file-level metadata, etc. At the same time, these systems are not limited to analytical or other functionalities, so they can combine the capabilities of the following types of information systems. In fact, this type is the most extensive in terms of offered functionalities. Among such systems, Sketch Engine, already mentioned above, stands out. It is also worth mentioning WebBootCat [5]–a web application that allows users to create corpora from web resources. Sketch Engine served as the querying tool for the corpus created by WebBootCat. WebBootCat was developed by the team responsible for the development of Sketch Engine. Consequently, as of today, all functionalities of WebBootCat have been transferred to Sketch Engine, which has become, to some extent, a “universal soldier” for corpus work.

2. Text annotators – the primary purpose of such systems is to add markup at the text level and below. The added metadata may include information about parts of speech (POS), lemmas, and tokens. It is important to note that annotation is not a mandatory requirement for corpus creation. Representatives of this type include TreeTagger, Dexter, and Elianto.

3. Concordancers – programs that allow users to obtain analytical data from a linguistic corpus. Such programs typically provide statistical information about the usage of a word or phrase, its context (KWIC), and can compare two corpora to highlight anomalous patterns, among other functionalities. In our opinion, the most advanced concordancer is AntConc. It is also worth mentioning WordSmith, ParaConc, and CasualConc.

It is also worth mentioning information systems developed for specific text corpora– for example, BNCweb, a specialized web application for analyzing the British National Corpus, developed by researchers at Lancaster University. Such information systems allow work only with a specific corpus and do not permit research based on one's own linguistic corpora; however, they are usually very effective due to their narrow specialization.

Synthesis of functional and non-functional characteristics to be used in the analysis

Before proceeding to the analysis of information systems for supporting corpus work, it is essential to define the functional and non-functional characteristics of the system that are important for the end user.

Among the functional characteristics, we can highlight those that have been substantiated in previous works by the authors [14], namely:

1. Creation of text corpora.
2. Text processing functionality.
3. Text markup.
4. Creation of an interface for manual or semi-automatic text annotation.
5. Function for saving and exporting annotated texts and the entire corpus.
6. Management of corpora and annotated data.
7. Function for searching and filtering corpora based on various parameters.
8. Data analysis and visualization.
9. Function for statistical comparison of different text corpora.

In addition to the aforementioned characteristics of the information system, we find it necessary to highlight certain additional requirements that, in our opinion, are important for corpus linguists, as well as those that allow for local use of the product:

1. *The presence of an intuitive user interface* should undoubtedly be taken into account, as a clear, intuitive, and user-friendly interface will expand the audience for the information system's use–allowing not only experienced corpus linguists but also novice linguists, language teachers, and even students to use it within the framework of Data-Driven Learning methodologies.
2. *Support for the Ukrainian language*, with a two-way aspect of integration of the Ukrainian language into the system. On the one hand, the information system should support the processing of texts in the Ukrainian language. Without this functional characteristic, we believe that the system would lose its primary purpose–advancing research specifically in the Ukrainian language: the creation of grammars and dictionaries, the development of machine translation, and support for generative AI. On the other hand, the system requires a user interface in the Ukrainian language, as although corpus linguists generally have sufficient knowledge of foreign languages, the lack of language support in the UI leads to a decrease in the user base for the system.
3. *The use of the system and the dissemination of source code based on an open license*, as noted by Lawrence Anthony [2], specifically the use of individual, small projects allows for the development of corpus linguistics. While the commercialization of the product allows for the acquisition of specific functional and non-functional characteristics of the system, it makes it less flexible in terms of individualizing parameters for corpus work, developing new features tailored to a specific corpus or language, and significantly reduces its use by small volunteer organizations or individual researchers. The extensive efforts to personalize the work of tools for corpus management are evidenced by the number of published open-source libraries, extensions, and standalone programs. Analyzing recent works on Ukrainian language corpora [7, 12, 17], we see researchers' attempts to utilize tools that are most suitable for creating a specific corpus while also allowing for maximum customization.

Among the main characteristics of this information system, the following should be highlighted:

1. Organization of corpora, as it allows for the uploading and storing of texts as linguistic corpora for further analysis and management.

2. Search procedures provide tools for effective searching of phrases, collocations, and word forms within the corpus.

3. Text analysis, which involves the ability to conduct various linguistic analyses, such as frequency analysis, collocation analysis, keyword-in-context searches, statistical conclusions, etc.

4. Data visualization, including graphs and charts, aids in understanding linguistic patterns and trends within the text.

5. Multilingual capabilities, as it supports various languages, including Ukrainian at a basic level, making it a useful tool for research in different linguistic environments.

6. Ease of use, defined by the presence of an intuitive interface that simplifies navigation and usage.

Among the drawbacks of this information system, the following should be noted:

1. Commercial nature of the project – although this information system is a powerful corpus manager, it is paid for all aspects of its use–both for individual users to access added corpora and for hosting the corpus itself. This essentially means a complete lack of open corpora for research. Additionally, since the project is commercial and users do not have access to the source code, it develops according to the plans set by the developing company rather than by the researchers themselves.

2. Limited ability to customize the behavior of the information system. While the information system has a fairly extensive list of settings in advanced configurations, users do not have access to the source code and cannot modify the system's behavior, for example, to refine word annotation. Users also cannot add their own "individual" annotations or adjust the annotations made by the system, thus relying entirely on its functionality, which is sufficient for basic analytical use but creates limitations for analyzing deeper or more specific linguistic phenomena.

Analysis of Sketch Engine

Sketch Engine [15] (Fig. 2) is a corpus manager that provides a wide range of tools for storing, organizing, and analyzing language corpora.

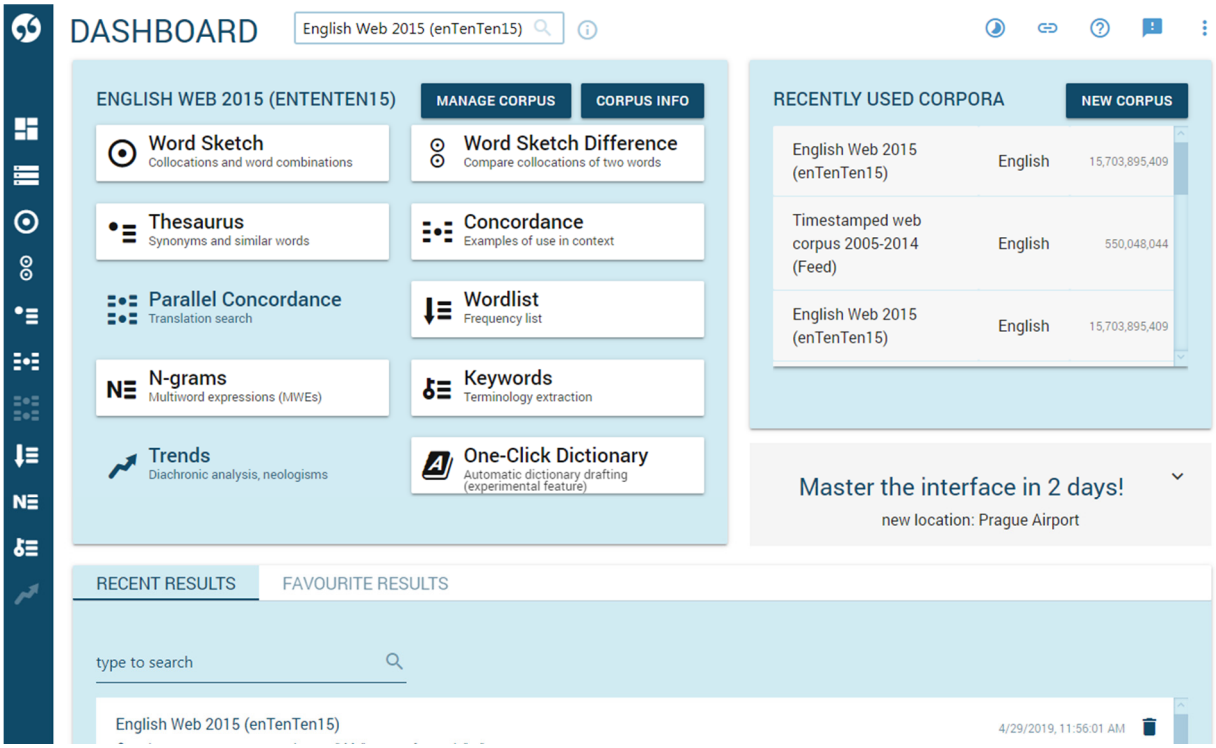

Fig. 2. Sketch Engine interface

3. Lack of choice regarding where the corpus will be stored can pose problems for text corpora with “sensitive” data, such as personal correspondence.

Lack of a user interface in Ukrainian.

Analysis of AntConc

AntConc [3] (Fig. 3) is a concordancer with elements of a corpus manager that has deep functionality for statistical analysis of corpora.

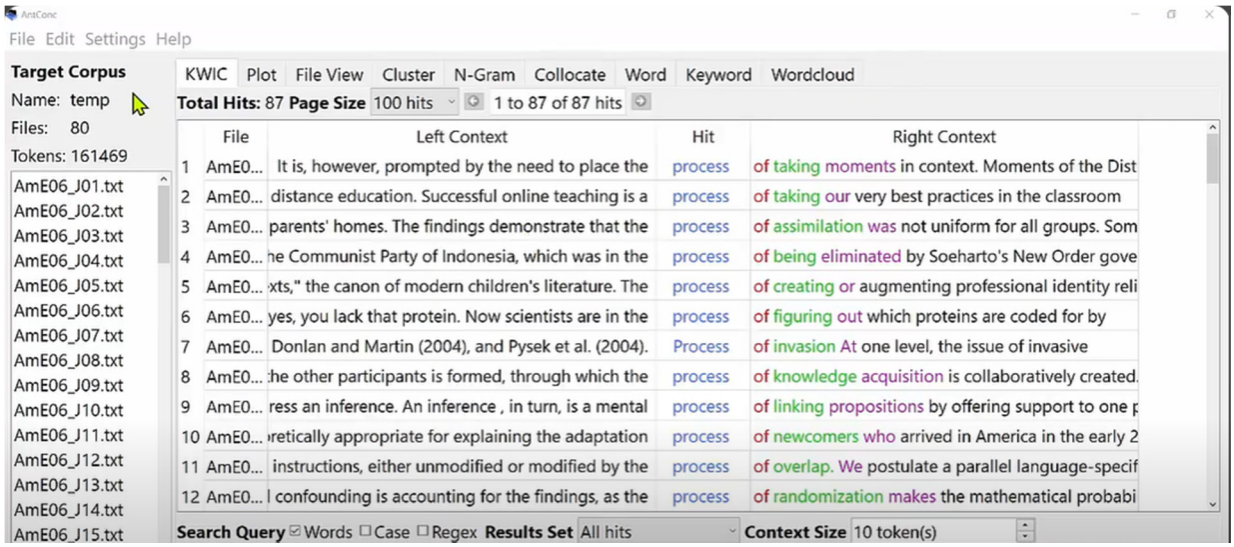

Fig. 3. AntConc Interface

Among the main characteristics of this information system, the following should be highlighted:

1. Corpus organization allows users to upload and store texts as linguistic corpora for further analysis, although its primary function is to analyze already created corpus databases.
2. Search procedures involve the availability of tools for effective searching of phrases, collocations, and word forms within the corpus.
3. Text analysis enables conducting in-depth linguistic analyses, including frequency analysis, collocation analysis, keyword-in-context searches, and comparison of two corpora. The system offers extensive customization options for the analytical engine.
4. Support for multiple languages means it accommodates various languages encoded in UTF-8 (though other encodings are also available if needed), including Ukrainian at a basic level, making it a useful tool for research in diverse linguistic environments.
5. Usability lies in the presence of a relatively simple interface, accompanied by comprehensive documentation and a series of instructional videos.
6. Installation capabilities for Windows, Mac, and Linux operating systems allow the system to be utilized for both small and larger projects.

Among the disadvantages of the information system, the following should be noted:

1. AntConc is not a corpus manager; it primarily relies on the user already having a prepared corpus database.
2. The system is mainly oriented towards individual use and lacks user authentication and authorization.
3. The system does not allow for the addition or editing of annotations.
4. There is no Ukrainian user interface.
5. The system’s interface is somewhat outdated.

Parallel comparison of Sketch Engine and AntConc

The structured comparative information of the reviewed information systems can be seen in Table 1.

Table 1

Comparative information of the Sketch Engine and AntConc

Критерій	**Sketch Engine**	**AntConc**
1	2	3
Corpus creation	**Advantages**: Creation of corpora from the file system and the WWW (searching, specific URL, and website).	**Advantages**: Creation of corpora from the file system.
	Disadvantages: Limitations of the search engine used; lack of the ability to bypass CAPTCHA systems.	**Disadvantages**: Lack of web scraping capability.
Text processing	**Advantages**: Incorrect symbols are not considered for analysis; tokenization and lemmatization are automatic; there is a possibility to add various metadata to each text; the ability to create bilingual and parallel corpora.	**Advantages**: Incorrect symbols are not considered for analysis; tokenization and lemmatization are automatic.
	Disadvantages: Lack of the ability to modify behavior for foreign words, tables, or images.	**Disadvantages**: Lack of ability to add metadata to the textual entity, modify behavior for foreign words, tables, or images.
Text annotation	**Advantages**: The ability to use markup to denote structures; automatic markup up to the sentence level.	**Disadvantages**: Lack of the ability to input markup.
	Disadvantages: Manual or external tools are required for markup deeper than the sentence level and for certain metadata details; lack of a customizable markup system.	
Interface for manual or semi-automatic text markup	**Disadvantages**: Absence of an interface for inputting markup and functionality for its verification.	**Disadvantages**: Absence of an interface for inputting markup and functionality for its verification.
Saving and exporting annotated texts and the entire corpus	**Advantages**: Ability to upload a user corpus in the format of a vertical file, plain text, and TMX (format for parallel corpora), as well as certain statistical data.	**Advantages**: Ability to export the corpus to a backup file; ability to export statistical data of the corpus.
	Disadvantages: Lack of ability to export previously uploaded corpora; absence of archiving capability; limitations on the selection of output file formats.	**Disadvantages**: Lack of ability to export vertical, XML, and plain text files, their archives, or the selection of output file formats.
Corpus and annotated data management	**Advantages**: Availability of extensive functionality for corpus management.	**Advantages**: Availability of functionality for corpus management.
	Disadvantages: Lack of capabilities for managing annotated data or their versions.	**Disadvantages**: Lack of ability to share the corpus or create subcorpora; absence of capabilities for managing annotated data or their versions.

Contuniation Table 1

1	2	3
Search and filtering of corpora by various parameters	**Advantages**: Ability to search by name, language, and category; sorting by word count.	**Advantages**: Ability to sort by name, status, and corpus size.
	Disadvantages: Lack of ability to filter by corpus metadata.	**Disadvantages**: Lack of search and filtering capabilities.
Data analysis and visualization	**Advantages**: A wide range of analytical tools and strong visual presentation of analytics.	**Advantages**: A wide range of analytical tools and extensive parameterization of these tools [11].
	Disadvantages: Somewhat limited possibilities for parameterizing analysis algorithms.	**Disadvantages**: Somewhat limited visual component.
Statistical comparison of different text corpora	**Advantages**: Ability to compare corpora based on tokens.	**Advantages**: Ability to compare data with a reference corpus based on keywords.
Support for the Ukrainian language (functionality)	**Advantages**: Support at the level of lemmas and stems.	**Advantages**: Support at the level of lemmas and stems.
	Disadvantages: Lack of the ability for customized markup and support for interlexical relationships; absence of Ukrainian in the language learning interface (SKELL).	**Disadvantages**: Lack of the ability for customized markup and support for interlexical relationships.
User-friendly interface	**Advantages**: The interface is user-friendly, featuring info buttons on the UI and a dedicated page for language learning (DDL).	**Advantages**: The user-friendly interface allows for use in DDL [10].
		Disadvantages: Somewhat outdated UI, lack of information buttons.
Support for the Ukrainian language (UI)	**Advantages**: Possibility of providing support through the translation of informational labels.	**Disadvantages**: Lack of support for the Ukrainian language.
	Disadvantages: Lack of support for the Ukrainian language.	
System use and open-source code distribution	**Advantages**: Availability of a version based on an open license.	**Advantages**: The system (supported on Linux, Windows, and macOS) is completely free for non-commercial use.
	Disadvantages: The free version is functionally limited, installation is only possible on CentOS, and there is no access to the source code.	**Disadvantages**: Lack of access to the source code.

Comparison of AntConc and Sketch Engine information systems based on Analytic Hierarchy Process

To conduct an expert evaluation for comparing Sketch Engine and AntConc using the Analytic Hierarchy Process (AHP), it is necessary to assess the importance of each criterion and sub-criterion for the user. Subsequently, these assessments will be used to construct comparison matrices for each criterion [16].

Based on the results of the previous discussion, the following criteria for evaluating information systems are identified:

1. Text corpus creation
2. Text processing
3. Text annotation
4. Saving and exporting
5. Data analysis and visualization
6. User interface intuitiveness
7. Support of the Ukrainian language

For each pair of proposed criteria, their importance for comparison is evaluated using the following scale:

- 1 – equally important;
- 3 – one criterion is somewhat more important than the other;
- 5 – one criterion is moderately more important;
- 7 – one criterion is significantly more important;
- 9 – one criterion is critically more important than the other.

The table below presents the pairwise comparison of expert assessments based on the proposed criteria (Table 2).

Table 2

Pairwise comparison of expert assessments based on criteria

Criterion	Text Corpus Creation	Text Processing	Text Annotation	Saving & Export	Data Analysis	UI Intuitiveness	Support of Ukrainian
Text corpus creation	1	3	5	3	5	7	5
Text processing	1/3	1	3	3	5	5	5
Text annotation	1/5	1/3	1	1/3	3	5	3
Saving & Export	1/3	1/3	3	1	3	5	3
Data analysis & visualization	1/5	1/5	1/3	1/3	1	3	3
UI intuitiveness	1/7	1/5	1/5	1/5	1/3	1	3
Support of Ukrainian	1/5	1/5	1/3	1/3	1/3	1/3	1

Normalization of the resulting matrix for each of its values is carried out according to the formula:

$$normalized_{value(i,j)} = \frac{A(i,j)}{\sum_{i=1}^{n} A(i,j)} \quad (1)$$

where the sum of each column is shown in Table 3.

Table 3

The sum of the columns of the expert assessment table

Criterion	Text Corpus	Text Processing	Text Annotation	Saving & Export	Data Analysis	UI Intuitiveness	Support of Ukrainian

	Creation						
∑	2.4095	5.2667	12.8667	8.2	17.6667	23.3333	23

Based on the normalized table, we calculate the weights using the formula:

$$weight_{(i)} = \frac{\sum_{j=1}^{n} normalized\_value(i,j)}{n} \quad (2)$$

Table 4

Normalized assessment table with priority weights

Criterion	Text Corpus Creation	Text Processing	Text Annotation	Saving & Export	Data Analysis	UI Intuitiveness	Support of Ukrainian	Weight
Text corpus creation	0.415	0.5696	0.3886	0.3659	0.283	0.3	0.2174	**0.3628**
Text processing	0.1383	0.1899	0.2332	0.3659	0.283	0.2143	0.2174	**0.2346**
Text annotation	0.083	0.0633	0.0777	0.0407	0.1698	0.2143	0.1304	**0.1113**
Saving & Export	0.1383	0.0633	0.2332	0.122	0.1698	0.2143	0.1304	**0.1530**
Data analysis & visualization	0.083	0.038	0.0259	0.0407	0.0566	0.1286	0.1304	**0.0719**
UI intuitiveness	0.0592	0.038	0.0155	0.0244	0.0189	0.0429	0.1304	**0.0470**
Support of Ukrainian	0.083	0.038	0.0259	0.0407	0.0189	0.0143	0.0435	**0.0378**

We will calculate the maximum eigenvalue necessary for further determining the consistency index (see Table 5).

Table 5

Calculation of weighted columns and weighted sum

Criterion	Text Corpus Creation	Text Processing	Text Annotation	Saving & Export	Data Analysis	UI Intuitiveness	Support of Ukrainian	Weight
Text corpus creation	0.3628	0.7038	0.5565	0.4590	0.3595	0.3290	0.1890	**2.9596**
Text processing	0.1209	0.2346	0.3339	0.4590	0.3595	0.2350	0.1890	**1.9319**
Text annotation	0.0726	0.0782	0.1113	0.0510	0.2157	0.2350	0.1134	**0.8772**
Saving & Export	0.1209	0.0782	0.3339	0.1530	0.2157	0.2350	0.1134	**1.2501**
Data analysis & visualization	0.0726	0.0469	0.0371	0.0510	0.0719	0.1410	0.1134	**0.5339**
UI intuitiveness	0.0518	0.0469	0.0223	0.0306	0.0240	0.0470	0.1134	**0.3360**
Support of Ukrainian	0.0726	0.0469	0.0371	0.0510	0.0240	0.0157	0.0378	**0.2850**

Thus, the maximum eigenvalue can be calculated as follows:

$$\lambda_{max} = \frac{\frac{2.95}{0.3628} + \frac{1.9319}{0.2346} + \frac{0.8772}{0.1113} + \frac{1.2501}{0.153} + \frac{0.5339}{0.0719} + \frac{0.336}{0.047} + \frac{0.2850}{0.0378}}{7} =$$

$$= \frac{8.1312 + 8.2349 + 7.8814 + 8.1706 + 7.4256 + 7.1489 + 7.5397}{7} == \frac{54.5323}{7} \approx 7.79 \quad (3)$$

The consistency index (CI) is calculated using the following formula:

$$CI = \frac{\lambda_{max-n}}{n-1} = \frac{7.79 - 7}{6} = 0.1317 \quad (4)$$

The consistency ratio (CR) is calculated using the formula:

$$CR = \frac{CI}{RI} \quad (5)$$

Considering that the random index (RI) value for a matrix with 7 criteria is 1.32, we obtain the following CR value:

$$CR = \frac{0.1317}{1.32} \approx 0.0997 \quad (6)$$

which is less than 0.1, thus we can consider that the proposed matrix of expert evaluations of criteria is consistent.

For each of the selected criteria, we will assess both systems using the following scale:

- 1 – the system has no advantages over the other;
- 3 – one system has slight advantages over the other;
- 5 – moderate significance advantage;
- 7 – the system has significant advantages;
- 9 – the system has critical advantages over the other.

The evaluation will be based on the comparative extraction of the two systems presented in Table 1. Comparison of the two systems for the "Corpus creation" criterion (Tables 6–7).

Table 6

Comparative assessment for the “Corpus creation” criterion

Corpus creation	Sketch Engine	AntConc
Sketch Engine	1	5
AntConc	1/5	1
Sum	**1.20**	**6.00**

Table 7

Normalized table with weights for the “Corpus creation” criterion

Corpus creation	Sketch Engine	AntConc	**Weight**
Sketch Engine	0.83	0.83	**0.83**
AntConc	0.17	0.17	**0.17**

Comparison of the two systems for the “Text processing” criterion (Tables 8–9)

Table 8

Comparative assessment for the “Text processing” criterion

Text processing	Sketch Engine	AntConc
Sketch Engine	1	3
AntConc	1/3	1
Sum	**1.33**	**4.00**

Table 9

Normalized table with weights for the “Text processing” criterion

Text processing	Sketch Engine	AntConc	**Weight**
Sketch Engine	0.75	0.75	**0.75**
AntConc	0.25	0.25	**0.25**

Comparison of the two systems for the “Text annotation” criterion (Tables 10–11)

Table 10

Comparative assessment for the “Text annotation” criterion

Text annotation	Sketch Engine	AntConc
Sketch Engine	1	7
AntConc	1/7	1
Sum	**1.14**	**8.00**

Table 11

Normalized table with weights for the “Text annotation” criterion

Text annotation	Sketch Engine	AntConc	**Weight**
Sketch Engine	0.88	0.88	**0.88**
AntConc	0.13	0.13	**0.13**

Comparison of the two systems for the "Saving & Export" criterion (Tables 12–13)

Table 12

Comparative assessment for the “Saving & Export” criterion

Saving & Export	Sketch Engine	AntConc
Sketch Engine	1	3
AntConc	1/3	1
Sum	**1.33**	**4.00**

Table 13

Normalized table with weights for the “Saving & Export” criterion

Saving & Export	Sketch Engine	AntConc	**Weight**
Sketch Engine	0.75	0.75	**0.75**
AntConc	0.25	0.25	**0.25**

Comparison of the two systems for the "Data analysis & visualization" criterion (Tables 14–15)

Table 14

Comparative assessment for the “Data analysis & visualization” criterion

Analysis & visualization	Sketch Engine	AntConc
Sketch Engine	1	1/3
AntConc	3	1
Sum	**4.00**	**1.33**

Table 15

Normalized table with weights for the “Data analysis & visualization” criterion

Analysis & visualization	Sketch Engine	AntConc	**Weight**
Sketch Engine	0.25	0.25	**0.25**
AntConc	0.75	0.75	**0.75**

Comparison of the two systems for the "UI intuitiveness" criterion (Tables 16–17)

Table 16

Comparative assessment for the “UI intuitiveness” criterion

UI intuitiveness	Sketch Engine	AntConc
Sketch Engine	1	3
AntConc	1/3	1
Sum	**1.33**	**4.00**

Table 17

Normalized table with weights for the “UI intuitiveness” criterion

UI intuitiveness	Sketch Engine	AntConc	**Weight**
Sketch Engine	0.75	0.75	**0.75**
AntConc	0.25	0.25	**0.25**

Comparison of the two systems for the “Support of Ukrainian” criterion (Tables 18–19).

Table 18

Comparative assessment for the “Support of Ukrainian” criterion

Support of Ukrainian	Sketch Engine	AntConc
Sketch Engine	1	3
AntConc	1/3	1
Sum	**1.33**	**4.00**

Table 19

Normalized table with weights for the “Support of Ukrainian” criterion

Support of Ukrainian	Sketch Engine	AntConc	**Weight**
Sketch Engine	0.75	0.75	**0.75**
AntConc	0.25	0.25	**0.25**

We will conduct priority weighting for both systems across all criteria and derive an overall priority for each system (Table 20).

Table 20

Overall priority of the AntConc and Sketch Engine systems

Criterion		**Corpus creation**	**Text processing**	**Text annotation**	**Saving & Export**	**Analysis & Visualization**	**UI intuitiveness**	**Support of Ukrainian**	**Overall priority**
Criteria weight		0.3541	0.2506	0.1160	0.1555	0.0700	0.0450	0.0302	
Sketch Engine	**Priority**	0.83	0.75	0.88	0.75	0.25	0.75	0.75	**0.7745**
	Weighted priority	**0.2939**	**0.1880**	**0.1021**	**0.1166**	**0.0175**	**0.0338**	**0.0227**	
AntConc	**Priority**	0.17	0.25	0.13	0.25	0.75	0.25	0.25	**0.2481**
	Weighted priority	**0.0602**	**0.0627**	**0.0151**	**0.0389**	**0.0525**	**0.0113**	**0.0076**	

Based on the defined criteria, their priorities, and expert evaluations, the Sketch Engine system can be considered a more optimal choice for a corpus manager. However, it should be noted that the evaluation criteria and their priority may vary significantly depending on the circumstances of using the information system (for example, considering the commercial aspect–usage by a profit-making organization or within a research project; scale–necessity for creating sub-corpora or lack thereof; analytical needs, etc.).

Conclusions

The research conducted on information systems for working with text corpora has shown that each of the analyzed platforms–Sketch Engine and AntConc–has its unique advantages and disadvantages, making them optimal for different usage scenarios.

Sketch Engine stands out with extensive capabilities for general, unconditional application in linguistic research. The system supports the creation and management of corpora, text annotation, and offers tools for data visualization, making it a versatile solution for large projects and research teams. Additionally, Sketch Engine provides a high level of automation and flexibility in working with multilingual corpora, which is a significant advantage in large-scale research.

However, AntConc also possesses several important advantages, especially in cases involving individual or small research projects. This system, while not as powerful compared to Sketch Engine, can be the optimal choice under certain circumstances, particularly with limited budgets, as AntConc is free software. Its user-friendly interface and support for various specific parameters for text analysis enable researchers to work effectively on narrow tasks that may not require the extensive features offered by Sketch Engine. AntConc is also a good option for beginners and those working with small text corpora or needing quick and simple analysis of linguistic material without complex preprocessing.

The results of this study may be beneficial for corpus and applied linguists in selecting appropriate information systems for creating and analyzing text corpora. The described advantages and disadvantages of each system, along with the comparative analysis based on the analytic hierarchy method, will help researchers determine which platform best meets their needs depending on the project's scale, budget, and specificity of linguistic tasks. Furthermore, the results may serve as a foundation for further scientific work, particularly for improving information systems or developing new methodologies for analyzing textual data. Based on the conducted analysis, requirements for the information system for language corpus processing have been formulated, which will be developed by the authors of the article.

In future research, the authors plan to delve deeper into additional aspects of using these systems, as well as the possibilities of integrating these tools with other software for comprehensive analysis of linguistic material.

Список літератури

1. Abdullayeva, O. (2020). Programs used to create the language corpus and their principles. *ACADEMICIA: An International Multidisciplinary Research Journal*. 10. 1778. https://doi.org/10.5958/2249-7137.2020.00749.1.

2. Anthony, L. (2013). A critical look at software tools in corpus linguistics. *Linguistic Research*. 30. 141–161. https://doi.org/10.17250/khisli.30.2.201308.001.

3. Anthony, L. (2023). AntConc (Version 4.2.4) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software.

4. Anthony, L. (2023). Corpus AI: Integrating Large Language Models (LLMs) into a Corpus Analysis Toolkit. *Presentation given at the 49th Annual Conference of the Japan Association for English Corpus Studies*, Kansai University, Osaka, Japan. URL: https://osf.io/srtyd/.

5. Baroni, M., Kilgarriff, A., Pomikálek, J., & Rychlý, P. (2006). WebBootCaT: a Web Tool for Instant Corpora. *Proceedings of the 12th EURALEX International Congress*. URL: https://www.researchgate.net/publication/242220785_WebBootCaT_a_web_tool_for_instant_corpora

6. Bayón, Candelas. (2024). Specialized terminology, instrumental competence, and corpus management tools / Terminología especializada, competencia instrumental y herramientas de gestión de corpus. *FITISPos International Journal*. 11. 220-–238. https://doi.org/10.37536/FITISPos-IJ.2024.11.1.402.

7. Chaplynskyi, D. (2023). Introducing UberText 2.0: A Corpus of Modern Ukrainian at Scale. *Proceedings of the Second Ukrainian Natural Language Processing Workshop*, 1–10, Dubrovnik. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.unlp-1.1

8. Jusoh W. et al. (2024). Exploring corpus linguistics via computational tool analysis: key finding review. *Indonesian Journal of Electrical Engineering and Computer Science*. 34. 1052. http://doi.org/10.11591/ijeecs.v34.i2.pp1052-1062.

9. Капранов, Я. В. (2021). Корпусний менеджер AntConc та його можливості для визначення частоти ключових слів різних мов. *Інженерія знань як фактор міжкультурної кооперації України з Японією, КНР і Республікою Корея: матеріали II міжнародної науковопрактичної відеоконференції*, 1-2 грудня 2021 року (с. 100-102). Видавничий центр КНЛУ. URL: http://rep.knlu.edu.ua/xmlui/bitstream/handle/787878787/2980/Капранов%20Я.%20В.%20Корпусний%20менеджер%20AntConc%20та%20його%20можливості%20для%20визначення%20частоти%20ключових%20слів%20різних%20мов.pdf.

10. Khairas, Eri. (2019). Using Antconc Software As English Learning Media: The Students' Perception. *Epigram*. 16. 189–194. http://dx.doi.org/10.32722/epi.v16i2.2234.

11. Kocincová, Lucia & Jakubíček, Miloš & Kovář, Vojtěch & Baisa, Vít. (2015). Interactive Visualizations of Corpus Data in Sketch Engine. *Proceedings of the Workshop on Innovative Corpus Query and Visualization Tools*

at NODALIDA 2015. 17–22. URL: https://www.researchgate.net/publication/280566089_Interactive_Visualizations_of_Corpus_Data_in_Sketch_Engine.

12. Коциба Н. (2013). Лабораторія української. URL https://mova.institute/.

13. Kovář, Vojtěch & Baisa, Vít & Jakubíček, Miloš. (2016). Sketch Engine for Bilingual Lexicography. *International Journal of Lexicography*. Volume 29, Issue 3, September 2016, Pages 339–352, https://doi.org/10.1093/ijl/ecw029.

14. Козак І., Кунанець Н. (2024). Проблеми розроблення текстових корпусів засобами інформаційних систем і шляхи їх вирішення. *Науковий вісник НЛТУ України*, 34(2), 101-108. https://doi.org/10.36930/40340213

15. Lexical Computing CZ s.r.o. (2023). SketchEngine [Computer Software]. Available from https://www.sketchengine.eu/.

16. Mu, E., Pereyra-Rojas, M. (2017). Understanding the Analytic Hierarchy Process. In: Practical Decision Making. *SpringerBriefs in Operations Research*. Springer, Cham. https://doi.org/10.1007/978-3-319-33861-3_2.

17. Шведова М., Валденфельс Р., Старко В. (2019). Генеральний регіонально анотований корпус української мови (ГРАК). URL: https://uacorpus.org/Kyiv/ua.

References

1. Abdullayeva, O. (2020). Programs used to create the language corpus and their principles. *ACADEMICIA: An International Multidisciplinary Research Journal*. 10. 1778. https://doi.org/10.5958/2249-7137.2020.00749.1.

2. Anthony, L. (2013). A critical look at software tools in corpus linguistics. *Linguistic Research*. 30. 141–161. https://doi.org/10.17250/khisli.30.2.201308.001.

3. Anthony, L. (2023). AntConc (Version 4.2.4) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software.

4. Anthony, L. (2023). Corpus AI: Integrating Large Language Models (LLMs) into a Corpus Analysis Toolkit. *Presentation given at the 49th Annual Conference of the Japan Association for English Corpus Studies*, Kansai University, Osaka, Japan. URL: https://osf.io/srtyd/.

5. Baroni, M., Kilgarriff, A., Pomikálek, J., & Rychlý, P. (2006). WebBootCaT: a Web Tool for Instant Corpora. *Proceedings of the 12th EURALEX International Congress.* URL: https://www.researchgate.net/publication/242220785_WebBootCaT_a_web_tool_for_instant_corpora

6. Bayón, Candelas. (2024). Specialized terminology, instrumental competence, and corpus management tools / Terminología especializada, competencia instrumental y herramientas de gestión de corpus. *FITISPos International Journal*. 11. 220-–238. https://doi.org/10.37536/FITISPos-IJ.2024.11.1.402.

7. Chaplynskyi, D. (2023). Introducing UberText 2.0: A Corpus of Modern Ukrainian at Scale. *Proceedings of the Second Ukrainian Natural Language Processing Workshop*, 1–10, Dubrovnik. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.unlp-1.1

8. Jusoh W. et al. (2024). Exploring corpus linguistics via computational tool analysis: key finding review. *Indonesian Journal of Electrical Engineering and Computer Science*. 34. 1052. http://doi.org/10.11591/ijeecs.v34.i2.pp1052-1062.

9. Kapranov Y. (2021). The Antconc Corpus Manager and its Possibilities for Determining the Frequency of Key Words in Different Languages. *Knowledge Engineering as a Factor of Intercultural Cooperation between Ukraine, Japan, China, and the Republic of Korea: materials of the II International Scientific and Practical Videoconference*, December 1-2, 2021 (pp. 100–102). Publishing Center of Kyiv National Linguistic University. URL: http://rep.knlu.edu.ua/xmlui/bitstream/handle/787878787/2980/Капранов%20Я.%20В.%20Корпусний%20менеджер%20AntConc%20та%20його%20можливості%20для%20визначення%20частоти%20ключових%20слів%20різних%20мов.pdf.

10. Khairas, Eri. (2019). Using Antconc Software As English Learning Media: The Students' Perception. *Epigram*. 16. 189–194. http://dx.doi.org/10.32722/epi.v16i2.2234.

11. Kocincová, Lucia & Jakubíček, Miloš & Kovář, Vojtěch & Baisa, Vít. (2015). Interactive Visualizations of Corpus Data in Sketch Engine. *Proceedings of the Workshop on Innovative Corpus Query and Visualization Tools at NODALIDA 2015*. 17–22. URL: https://www.researchgate.net/publication/280566089_Interactive_Visualizations_of_Corpus_Data_in_Sketch_Engine.

12. Kotsyba, N. (2013). Laboratoriia ukraiinskoii. URL https://mova.institute/.

13. Kovář, Vojtěch & Baisa, Vít & Jakubíček, Miloš. (2016). Sketch Engine for Bilingual Lexicography. *International Journal of Lexicography*. Volume 29, Issue 3, September 2016, Pages 339–352, https://doi.org/10.1093/ijl/ecw029.

14. Kozak, I. & Kunanets, N. (2024). Challenges in creating text corpus using information systems and ways to solve them. *Naukovyi visnyk NLTU Ukrainy*, 34(2), 101–108. [In Ukrainian]. https://doi.org/10.36930/40340213

15. Lexical Computing CZ s.r.o. (2023). SketchEngine [Computer Software]. Available from https://www.sketchengine.eu/.

16. Mu, E., Pereyra-Rojas, M. (2017). Understanding the Analytic Hierarchy Process. In: Practical Decision Making. *SpringerBriefs in Operations Research*. Springer, Cham. https://doi.org/10.1007/978-3-319-33861-3_2.

17. Shvedova M., Valdenfels R. & Starko V. (2019). Heneralnyi rehionalno anotovanyi korpus ukrainskoi movy (HRAK). URL: https://uacorpus.org/Kyiv/ua.

ІНФОРМАЦІЙНІ СИСТЕМИ ДЛЯ РОБОТИ З ТЕКСТОВИМИ КОРПУСАМИ: КЛАСИФІКАЦІЯ ТА ПОРІВНЯЛЬНИЙ АНАЛІЗ

Іван Козак[1], Наталія Кунанець[2]

[1-2] Національний університет “Львівська політехніка”,
кафедра інформаційних систем та мереж, Львів, Україна
[1] E-mail: ivan.kozak.lp@gmail.com, ORCID: 0009-0007-4953-2816
[2] E-mail: Nataliia.E.Kunanets@lpnu.ua, ORCID: 0000□0003□3007□2462

У статті досліджено інформаційні системи для роботи з текстовими корпусами, зокрема їх застосування для лінгвістичного аналізу та управління великими текстовими даними. Проаналізовано інформаційні системи для підтримки текстових корпусів, проведено їх класифікацію та досліджено поступ функціональних можливостей. Основну увагу зосереджено на порівнянні двох найпоширеніших систем, котрі можна виділити за функціоналом як корпусні менеджери: “AntConc” і “Sketch Engine”. Оцінено їх за ключовими критеріями: створення корпусів текстів, опрацювання текстів, розмітка, збереження та експорт, аналіз і візуалізація даних, інтуїтивність інтерфейсу, підтримка української мови, а також наявність відкритої ліцензії. Метою дослідження було провести порівняльний аналіз цих систем з використанням методу аналізу ієрархій для визначення їх сильних та слабких сторін у різних умовах використання. Виявлено, що “Sketch Engine” забезпечує розширені можливості для створення й управління великими корпусами, розмітки та візуалізації даних, що робить його кращим вибором для великих дослідницьких проєктів. Водночас “AntConc” є більш доступною та ефективною системою для індивідуальних або малих досліджень завдяки простоті, відсутності ліцензійних витрат і підтримці специфічних параметрів для аналізу текстів. Результати дослідження можуть бути корисними для корпусних та прикладних лінгвістів під час вибору систем для створення і роботи з текстовими корпусами. Висновки сприятимуть ухваленню рішень щодо вибору відповідних інструментів залежно від конкретних потреб дослідження, обсягу роботи та бюджетних обмежень. Окрім того, результати дослідження можуть бути застосовані для вдосконалення існуючих та розробки нових інформаційних систем для забезпечення підтримки корпусів у подальших наукових проєктах авторів.

Ключові слова: корпусна лінгвістика, корпусний менеджер, AntConc, Sketch Engine, метод аналізу ієрархій