American Journal Of Social Sciences And Humanity Research

VOLUME Vol.05 Issue08 2025
PAGE NO. 113-116
DOI 10.37547/ajsshr/Volume05Issue08-27

Functional Possibilities Of The Sketch Engine Platform

Nurullayeva Oydin
PhD Candidate, Tashkent State University of Uzbek Language and Literature, Uzbekistan

Received: 25 June 2025; **Accepted:** 19 July 2025; **Published:** 31 August 2025

Abstract: This article discusses the use of Sketch Engine for working with corpora and parallel corpora. It outlines the concept of a corpus, describes the program's main features (word frequency analysis, collocations, concordance, word sketches, parallel corpus creation) and interface, and examines applications in linguistic research, machine translation, and language learning.

Keywords: Corpus; parallel corpus; Sketch Engine; word frequency; collocations; concordance; word sketches; linguistic research; machine translation; language learning.

Introduction: A corpus is a large collection of written or spoken texts in a given language or multiple languages. Parallel corpora, in turn, consist of the same texts translated into different languages and are primarily employed for training machine translation systems. The Sketch Engine software provides significant support in this process, as it enables the creation, analysis, and comparison of corpora across various languages.

Sketch Engine is regarded as one of the most advanced and up-to-date tools in the field of corpus linguistics worldwide. With the help of this software, linguistic researchers can analyze large-scale textual data in different languages, investigate word frequency, identify lexical collocations, and assess translation quality within parallel corpora. Sketch Engine is widely used not only in linguistic research but also in language learning, machine translation, and many other domains. The tool proves highly beneficial not only for linguists but also for translators, language acquisition specialists, and academic researchers.

Developed in 2003 by Lexical Computing Ltd., Sketch Engine is a corpus management and analysis system specifically designed for linguists, translators, and specialists in language studies. It was created to facilitate linguistic research and provides a wide range of functionalities for corpus compilation, lexical analysis, grammatical analysis, and language learning.

The platform hosts corpora of multiple languages and domains with extensive textual resources. The following table illustrates selected statistics and features available in Sketch Engine.

Feature	Value
Number of corpora	1000+ (covering a wide range of languages and domains)
Number of languages	90+
Total size of corpora	30+ billion units (words, phrases, lexical analyses)
Word analysis	Morphological and syntactic analysis
Number of collocations	100,000+
Word Sketches	Available in 40+ languages, displaying common word combinations
Parallel corpora	Available in 100+ languages
Specialized corpora	Available for Uzbek, English, Russian, and many other languages
Specialized analysis tools	Word frequency, concordance, lexical analysis, n-gram analysis
Morphological analysis	Supported for 30+ languages
API integration	Yes, integration with other systems via API is supported

A corpus (plural: corpora) is a large collection of texts in one or more languages used for linguistic research. Corpora are primarily employed for the analysis of collocations, word frequency, and morphological as well as syntactic features. While early corpora served mainly as fundamental tools for language learning, today they are extensively applied across all branches of linguistics, particularly in semantic and pragmatic

analyses [1].

The feasibility of creating corpora and the factors determining their effective use are defined by the following considerations:

1. Sufficient size and representativeness – a corpus must be large enough and balanced, ensuring the typicality of data and enabling comprehensive reflection of the full spectrum of linguistic phenomena.

2. Contextual authenticity – various types of data are preserved in their natural (contextual) form, providing opportunities for thorough and objective examination.

3. Reusability – once compiled and prepared, a corpus (as a dataset) may be reused by different researchers for diverse purposes.

4. Stylistic diversity – since corpora contain materials from various functional styles, they allow for pragmatic investigations specific to speech registers.

5. Author-specific analysis – as an electronic collection of an individual author's works, a corpus can reveal the writer's linguistic poetics, stylistic features, lexical preferences, and statistical tendencies in word usage.

6. Semantic versatility – corpora enhance the possibility of identifying both the usual (out-of-context) and the occasional (context-dependent) meanings of particular lexical items.

The process of building a corpus with Sketch Engine is straightforward and highly intuitive. The platform enables users to collect texts, incorporate them into a dataset, and conduct linguistic analyses. Texts for corpus compilation may be drawn from a variety of sources, including books, websites, academic articles, and other written materials. Sketch Engine provides specialized tools for systematically gathering these texts and preparing them for analysis.

In text analysis, Sketch Engine allows researchers to identify word frequency, collocations, syntactic structures, and other linguistic parameters. The software is based on highly efficient algorithms that facilitate the identification of linguistic features of each word, including morphology and syntax [2].

The primary task of corpus analysis is to examine the distribution of words, their interrelations, and their contextual properties. Using Sketch Engine, corpus analysis provides the following opportunities:

- Word frequency – determining how often each word occurs in a given text.
- Collocations – identifying words that frequently co-occur, such as good morning or make a decision.
- Lexical analysis – detecting morphological and syntactic properties of individual words.

Furthermore, Sketch Engine supports the discovery and examination of lexical associations across texts. For instance, linguist J. Sinclair (2004) emphasized the remarkable suitability of Sketch Engine for language learning and linguistic research.

Corpora are highly valuable resources for language learning, translation, and a wide range of scientific investigations. They can be employed across all domains of linguistics, particularly in semantics, pragmatics, and syntax. Data extracted from corpora through Sketch Engine can lead to significant advances in both language acquisition and translation studies [4].

The software is also applied in the development of language learning methodologies, as it enables the analysis of grammar, vocabulary, and other linguistic components. Parallel corpora are collections of texts consisting of the same content translated into multiple languages. They are primarily employed in machine translation and language learning. Through parallel corpora, it becomes possible to investigate interlingual relations and evaluate translation quality. Sketch Engine demonstrates high efficiency in working with parallel corpora and provides the necessary tools to enhance translation accuracy [5].

The platform simplifies the process of creating parallel corpora. Users can merge texts available in multiple languages and analyze them within a unified framework. During this process, Sketch Engine aligns multilingual texts, identifies translations, and facilitates their comparison. With this functionality, researchers can accurately determine the context of translated words and explore the interconnections among languages.

Corpus construction in Sketch Engine represents a fundamental stage for linguistic research, language learning, translation studies, and other scholarly inquiries. In what follows, I will provide a detailed and precise explanation of this process, enabling researchers to independently construct their own corpora on the Sketch Engine platform.

1. Accessing the Sketch Engine Platform. To begin working with Sketch Engine, you must first log into the platform. If you do not yet have an account, you will need to create one. The following steps outline the login process:

1. Go to https://www.sketchengine.eu.

2. Create a new account or log in with your existing credentials.

If you are using the free version, access will be limited, and certain functionalities of Sketch Engine will remain

restricted.

3. Creating a New Corpus. On the main interface of Sketch Engine, locate the “Create corpus” option. The process of creating a corpus consists of several stages:

Step 1. Provide a Name and Description for the Corpus

- Corpus Name: Assign a name to your corpus. For example, Uz-English Phrase.
- Corpus Description: Write a brief description of the corpus, including its purpose and contents. For instance: This corpus consists of academic articles and is designed for word frequency analysis.

Step 2. Select Sources of Texts. To build a corpus, you must gather relevant texts. Sketch Engine allows texts to be imported in several ways:

1. Manual Uploading: If you already have prepared text files, you can upload them to the platform. These files must be in .txt format.
2. URL Uploading: Sketch Engine can also collect texts directly from websites. You can input URLs to extract data from specific websites or web pages.
3. External Sources (e.g., Yandex corpora): Where available, Sketch Engine integrates with other widely used resources (such as Yandex corpora), making it possible to obtain large volumes of text with ease.

Step 3. Select the Corpus Type. Sketch Engine supports different types of corpora:

- Plain Text: Each line is treated as a separate document.
- Segmented Text: The text must be pre-segmented, meaning each word is separated (e.g., tokenized).
- Annotated Text: The text contains morphological or syntactic tags. If you intend to conduct deeper linguistic analyses, you may opt for annotated corpora.

Step 1. Uploading Texts. To upload texts into Sketch Engine:

- Manual Uploading: If you are uploading files manually, select the files and click the “Upload” button.
- URL Uploading: If you choose to collect texts from websites, enter the URL and allow the automated system of Sketch Engine to gather the texts.

Once the corpus has been created, it can be analyzed to investigate various linguistic parameters. Sketch Engine provides the following key functionalities:

- Word Frequency: The Word Frequency function identifies words that occur repeatedly in the text. This feature lists the most frequently used words. For example, in English, words such as the, is, and science may appear with the highest frequency.
- Collocations: Collocations refer to words that are frequently used together, and they can be easily identified using Sketch Engine. In English, for instance, expressions such as make a decision or take a risk are recognized as collocations. Analyzing collocations allows linguists to gain deeper insights into lexical relationships.
- Concordance: Concordance tools display words within their textual context. For example, the function can retrieve all occurrences of the word computer and illustrate the contexts in which it appears.
- Word Sketches: Sketch Engine generates “word sketches” that summarize a word’s most common collocations, grammatical relations, and semantic properties. For instance, for the word computer, typical collocations such as to use, personal, or desktop are displayed.

After the analysis is complete, the created corpus may be saved and reused in other research. Sketch Engine also allows for corpus export, which makes sharing and cross-platform usage possible.

Using Sketch Engine, it is possible to identify word translations, conduct cross-linguistic analysis, and evaluate translation quality within parallel corpora. For instance, the platform allows users to compare different language versions of the same text, verify the accuracy of lexical translations, and test the performance of machine translation systems [6].

The process of creating and analyzing corpora and parallel corpora with Sketch Engine may be illustrated through the following examples:

1. Corpus Creation: Collecting academic articles in English and Russian and analyzing them with Sketch Engine.
2. Parallel Corpus Creation: Building a parallel corpus with Uzbek and English texts and using it to assess translation quality.

Such tools are highly beneficial for both language learning and translation studies. Sketch Engine is not only an effective platform for language acquisition but also a powerful resource for linguistic research and machine translation evaluation.

The application of Sketch Engine in linguistic research, language pedagogy, machine translation, and related academic domains is of considerable significance. When working with corpora and parallel corpora, the software provides users with opportunities not only for textual analysis but also for teaching languages and facilitating translation tasks. By employing these tools,

linguists and translators can better identify cross-linguistic relationships and achieve improvements in translation quality. Looking ahead, Sketch Engine is expected to expand further, offering new opportunities in the fields of linguistics and translation studies.

REFERENCES

Biber, D., Conrad, S., & Reppen, R. Corpus Linguistics: Investigating Language Structure and Use. Cambridge University Press.1998.

Kilgarriff, A., Rychlý, P., Smrž, P., & Tugwell, D. The Sketch Engine: Ten years on. Lexicography, 1(1), 2014. 7-36.

Rizvanov Q. Sketche Engine korpus menejeri va uning ayrim imkoniyatlari. “Kompyuter lingvistikasi: muammolar, yechim, istiqbollar” Xalqaro ilmiy-amaliy konferensiya. 2022. B, 144-153

Tiedemann, J. Parallel corpora for machine translation. Springer. 2018

Varga, D., & Schöpflin, G. The Creation of the JRC-Acquis Parallel Corpus. In Proceedings of the MT Summit X. 2015.

Zhang, Y., & Zhou, M. Machine translation quality evaluation based on parallel corpora. Springer. 2017