JOURNAL OF ADVANCED LINGUISTIC STUDIES
VOL. 12, NO. 2, JUL-DEC 2025 (ISSN 2231-4075)

A Comparative Analysis of Corpus Linguistics Platforms: From AntConc to Sketch Engine

BARNO KUTLIMURATOVA
Urgench State University, Uzbekistan

ABSTRACT

Corpus linguistics has evolved from small-scale text collections analyzed with standalone software to large, web-based systems that incorporate linguistic annotation and visualization features. Despite the rapid growth of tools and platforms, there still is a lack of systematic comparisons regarding their functionalities, usability, and applicability for linguistic and pedagogical research. This paper discusses a comparative analysis of major corpus linguistics platforms-AntConc, Sketch Engine, LancsBox, and NoSketch Engine-to evaluate strengths and limitations regarding accessibility, analytical capabilities, visualization tools, tagging support, and integration with external pipelines. Each platform was evaluated using a standardized test corpus and real-world linguistic tasks such as keyword extraction, collocation analysis, and part-of-speech-tagging. The findings indicate that although AntConc remains a powerful and transparent option for corpus-based instruction and smaller-scale research, Sketch Engine provides the most comprehensive and scalable solution, including advanced features like word sketches, grammatical relations, and API support. LancsBox excels in visual concordance and collocation networks, while NoSketch Engine provides a good open-source alternative option for institutions with limited budgets. The study closes with recommendations for researchers and educators on how to select the appropriate tool depending on corpus size, research aims, and computational resources.

Keywords: Corpus linguistics; AntConc; Sketch Engine; LancsBox; NoSketch Engine; concordance; linguistic tools; learner corpus; computational linguistics

1. INTRODUCTION

Corpus linguistics has emerged as a methodology at the heart of linguistics in our time. It enables the empirical study of language based on large collections of authentic texts. By applying computational tools to linguistic theory, it offers the researcher the ability to notice patterns of lexical and grammatical usage that would have gone unnoticed in traditional qualitative approaches [1]. In the last three decades, corpus-based methods have radically transformed diverse areas of inquiry, such as lexicography, language teaching, sociolinguistics, computational linguistics, as well as natural language processing [2].

The rising importance of corpus linguistics has also inspired the creation of many software resources aimed at processing, analysing, or visualizing linguistic data. The early software tool available was only capable of concordance searches, having programs such as WordSmith Tools or MonoConc, but with the availability of AntConc developed by Laurence Anthony in 2005, everything changed, allowing users to enjoy a free, cross-platform, user-friendly interface for researchers, lecturers, or even students. Over the years, greater platforms with many capabilities, Sketch Engine, are also available today, providing access to hundreds of corpora with part-of-speech tags, with "word sketches" for the lexicogrammar analysis, developed by Kilgarriff et al. in 2014 [6].

Despite the popularity of these platforms, there is limited systematic comparison between them available in the academic literature. Most existing work has revolved around particular functionalities or educational uses, but few comparisons have taken place from a more general tech or research tool perspective [3]. This has left many researchers unsure about which platform is appropriate to their linguistic needs or constraints on the data size on hand.

The proposed work attempts to bridge the gap existing in the current state of affairs by comparing the strengths, weaknesses, and appropriateness levels of the four leading platforms available in the field of corpus linguistics, namely AntConc, Sketch Engine, LancsBox, and NoSketch Engine, with the help of a comprehensive comparison between the platforms.

The proposed work will help design future corpuses for research in the area of computational linguistics with an alternative or best-suitable platform selected from those described in the work.

2. RELATED WORK

The scope of comparison studies on different corpus linguistics software tools have typically been ad hoc, with the focus of the majority of the studies on the capabilities of each software package instead of overall reviews. The early comparison studies focused on the practicality of WordSmith Tools, MonoConc, and AntConc, with the main focus being on language teaching in the classroom [4]. The AntConc succeeded in showing how the modular nature, key-word analysis, and concordance displays of AntConc could revolutionize the world of corpus analysis even in resource-limited settings [5]. The work, nonetheless, also pointed out the shortcomings of working with AntConc, especially its scalability on large corpora without the availability of other levels of linguistic annotation, such as lemmatized or dependency-parse corpora, shown in Figure 1.

Following these, more integrated approaches emerged with Sketch Engine, characterized by automated part-of-speech tagging, “lexico-grammatical” sketching, with API-access to 500 corpora or more [6].

Simultaneously, recent years have seen the revision of other systems, with LancsBox standing out with its focus on graphical interfaces, visualizing data, especially in the context of co-occurrence networks, and concordance dispersion plots. TheNoSketch Engine provides an open-source solution to replicate the capabilities of Sketch Engine, ideal for hosting corpora on the local machine. Notably, these contributions

highlight the need for cross-comparative analysis across these tools, going beyond merely the usability aspects, applicable analysis, or its feasibility with the scope of linguistics, which is the focus of the current study.

Figure 1. *The work platform interface example of the AntConc tool*

3. MATERIALS AND METHODS

A comparison analysis framework was employed in the study to compare the effectiveness of four popular platforms used in corpus linguistics, namely AntConc, Sketch Engine, LancsBox, and NoSketch Engine, on the basis of the specified evaluation criteria, which was designed to compare the technical capabilities, linguistic capabilities, or both, of the platforms.

3.1. *Selection of platforms*

The four platforms selected are categorized into different platforms available in the field of corpus linguistics. AntConc is a basic, standalone application that prioritizes accessibility and basic functions for working with corpora, while Sketch Engine is a commercial, web-based application boasting broad multilinguality with in-depth grammatical relation extraction capabilities. LancsBoxis designed with the greatest attention paid

to visualize simplicity and educational soundness, while NoSketch Engine is its fully open-source, server-side sibling of Sketch Engine, letting users work with the corpora on their own local computers. The logos of the four selected platforms are presented in the figure below, represented in Figure 2.

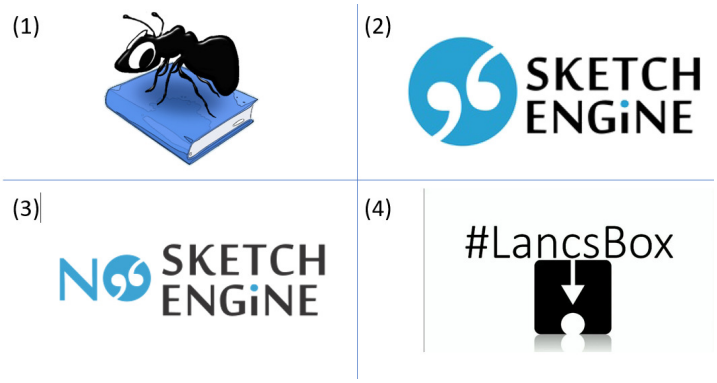

Figure 1. *All the chosen four corpus linguistics platforms – AntConc(1), Sketch Engine (2), NoSketch Engine (3), and the Lancs Box X (4).*

3.2. *Data and experimental setup*

A balanced test corpus of about 100,000 words was developed, comprising learner essays, news, and academic literature to represent different kinds of language data. All the tools worked on the same test corpus for fair comparison on standard tasks designed to check functionality and accuracy on the following steps, illustrated in Figure 3.

For each task, execution speed, result clarity, interface intuitiveness, and feature depth were recorded. In addition, installation requirements, documentation quality, and accessibility (offline vs. online) were qualitatively assessed.

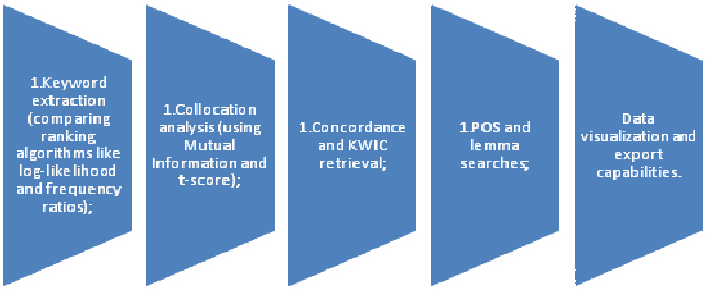

Figure 3. *Tool evaluation tasks*

3.3. *Evaluation metrics*

To maintain objectivity, both quantitative and qualitative metrics were employed. The quantitative aspects involved processing time and the number of supported analytical functions. The qualitative metrics involved the application of a structured rubric, guided by the framework proposed by McEnery & Hardie [8], which covered the aspects of usability, scalability, reproducibility, and educational value. Lastly, the consistency of the platforms was validated by comparing the analysis of collocation conducted on the same statistical metrics across platforms, proposed by Gries [9].

4. RESULTS AND DISCUSSION

The comparative evaluation revealed clear distinctions among the four corpus linguistics platforms in terms of functionality, scalability, and pedagogical applicability.

As for the analytical capabilities of the tools, among all tools, Sketch Engine demonstrated the most comprehensive analytical power. Its Word Sketch function effectively summarized grammatical relations for target lemmas, producing collocational patterns beyond what frequency-based tools could achieve. Its CQL (Corpus Query Language) allowed complex linguistic queries integrating part-of-speech and lemma constraints. However, this sophistication came at the cost of accessibility – its learning curve and subscription model limit adoption in resource-constrained settings.

AntConc, on the other hand, was highly effective in tasks involving the extraction of keywords or creating concordance, although its processing speed made it the best tool for small-scale corpora or teaching purposes only, since AntConc was devoid of the capabilities of lemmatization or syntactic markup. LancsBox was the middle ground solution. It was strong on visualization, particularly with GraphColl co-occurrence networks or KWIC dispersion graphs, allowing the user to readily visualize co-occurrence networks between words. Yet, the processing was poor on bigger datasets with over 10 million words, and the export facilities were poor compared with the other tools available.

TheNoSketch Engine tool was able to replicate mostly the capabilities of its commercial version but required technical skills for installation and maintenance. Although NoSketch Engine was open-source and scalable, suitable for an institutional corpus, updates were not provided regularly or resources integrated as required by the Sketch Engine.

Regarding usability, AntConc and LancsBox are the most accessible to users with less technical competency, while NoSketch Engine is least accessible due to the need to configure the server, with command-line interface familiarity being assumed.

Although Sketch Engine is fully accessible via the web with the latest technology, there is lack of clarity about how the system operates, which could be problematic if the output is reliant upon the corpora, which are already pre-tokenized, and are proprietary resources. Regarding data formats, AntConc's flexibility is its greatest strength, followed by Sketch Engine, due to their adaptability to different platforms, followed by the others From a pedagogical standpoint, LancsBox and AntConc were most suitable for classroom corpus literacy, allowing learners to visualize concordances and collocations quickly. For advanced research, Sketch Engine provided unparalleled analytical flexibility and multilingual coverage, making it a preferred choice for large-scale projects, national corpora, or lexicographic research [10].

As a summary, the findings suggest a complementary relationship among tools rather than a competitive one, as shown in Figure 4:

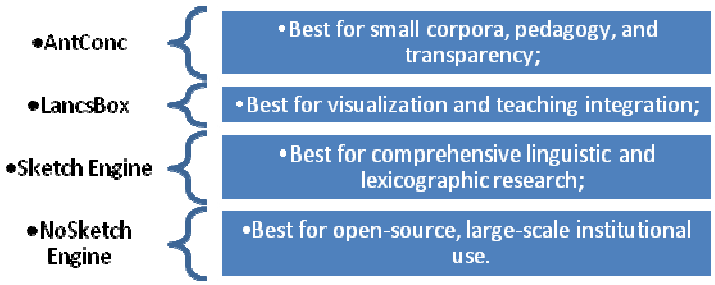

Figure 2. *Complementary relationship among tools compared*

This comparative analysis underscores the need for multi-tool workflows, where researchers combine tools based on their analytical goals rather than relying on a single platform.

5. CONCLUSION

This study made a structured comparison of the four prominent platforms of corpus linguistics, namely AntConc, Sketch Engine, LancsBox, and NoSketch Engine, to determine their strengths, drawbacks, and appropriateness for different tasks. The study confirms that none of the platforms is able to fulfill the different needs of corpus-based studies, but each platform is beneficial in its own way, depending on the purpose of the user, the size of the corpora, or the computing system available.

Sketch Engine is recognized for its strong functionality, complex grammatical links, and scalability, which makes the tool the most appropriate for academic or national corpora, lexicographic work, and other tasks that involve serious academic or professional analysis of language corpora. However, the most suitable tool for educational purposes is still AntConc, with LancsBox being the tool of choice for teaching corpus literacy, while NoSketch Engine is the most appropriate tool for the development of dedicated corpora within academic or

educational institutions without the limitations that are typically applied by commercial providers.

The comparison generally illustrates an underlying methodological lesson: the effectiveness of corpus-based analysis is best achieved by the complementarity of the toolsets, rather than integration with one platform or another. A multi-tool approach, gathering the strengths of AntConc, LancsBox, and Sketch Engine, is recommended for achieving clarity and rigor, respectively, in its pedagogical or scientific contributions.

Future studies may build on the current study by exploring the integration of newer AI-based corpus platforms, automated annotation processes, or the comparison of the current state of multilingual corpora regarding the lack or availability of disclaimer or copyrighting declarations on the corpora's content.

REFERENCES

[1] McEnery, T. & Hardie, A. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

[2] Kutlimuratova, B., Kuriyozov, E. & Tillaeva, M. 2022. Teaching English as a foreign language for primary school children: Literature review. *Foreign Language Teaching and Applied Linguistics*, 161-171.

[3] Anthony, L. 2013. A critical look at software tools in corpus linguistics. *Linguistic Research*, 30/2, 141-161.

[4] Kutlimuratova, B. 2022. Literature review of Gabrielatos' corpora and language teaching: Just a fling or wedding bells? *Proc. Talented Youth Conf*, 47-50.

[5] Anthony, L. 2014. AntConc: A learner and classroom-friendly tool for corpus research. In J. Thomas & A. Boulton (Eds.), *Researching Corpora in English Language Teaching*. London: Routledge.

[6] Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. 2014. The sketch engine: Ten years on. *Lexicography*, 1/1, 7-36.

[7] Sharipov, M., Kuriyozov, E., Yuldashev, O. & Sobirov, O. 2024. Uzbek verb detection: Rule-based detection of verbs in Uzbek texts. In *Proc. 2024 Joint Int. Conf. Comput. Linguistics, Lang. Resour. Eval. (LREC-COLING 2024)* (pp. 17343-17347).

[8] McEnery, T. & Hardie, A. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.
[9] Matlatipov, S. G., Rajabov, J., Kuriyozov, E. & Aripov, M. 2024. UzABSA: Aspect-based sentiment analysis for the Uzbek language. In *Proc. 3rd Annu. Meeting Special Interest Group Under-resourced Lang. @ LREC-COLING 2024* (pp. 394-403).
[10] Brezina, V., McEnery, T. & Wattam, S. 2015. Collocations in context: A new perspective on collocation networks. *Int. J. Corpus Linguistics*, 20/2, 139-173.

BARNO KUTLIMURATOVA
PHD STUDENT,
URGENCH STATE UNIVERSITY,
UZBEKISTAN.