UDC 81.33
DOI https://doi.org/10.32782/2522-4077-2025-214.1-16

CORPUS-BASED APPROACH TO SPECIALIZED TRANSLATION TRAINING: SKETCH ENGINE TOOLS AND CQL QUERIES

КОРПУСНИЙ ПІДХІД ДО ВИКЛАДАННЯ ФАХОВОГО ПЕРЕКЛАДУ: ІНСТРУМЕНТИ SKETCH ENGINE ТА CQL-ЗАПИТИ

Tarnavska M. M.,
orcid.org/0000-0002-5476-911X
Candidate of Philological Sciences, Associate Professor,
Associate Professor at the Chair of Translation, Applied and General Linguistics,
Volodymyr Vynnychenko Central Ukrainian State University

Modern applied linguistics cannot be imagined without language corpora. They are both its powerful and efficient research tool and a virtually unlimited database of linguistic data for various fields and needs of sciences that use the language arsenal in one way or another. Automation of text corpus generation and research creates new opportunities not only for philology, but also for specialists who use such data for practical purposes. Corpus-based methods play an important role in improving language teaching, particularly translation, as they allow for the accurate and systematic selection of specialised language materials necessary for mastering vocabulary, peculiarities of use and translation of key linguistic units, as well as for identifying current language trends in a particular field. Despite the huge selection of platforms and software for managing corpora and text analysis, Sketch Engine stands out due to its power, as well as the size and versatility of its text collections: it allows you to not only analyse existing corpora, but also create your own, including multilingual ones, research vocabulary, phrases, terminology, translation equivalents, and generate learning materials using CQL queries and built-in linguistic functions.

Sketch Engine is particularly effective in teaching professional texts and their translation, as it allows you to effectively research specialised texts, identify key terminology and typical phrases, analyse translation approaches and prepare teaching materials for translation students, while using the CQL query language to make the process of working with specialised texts more focused and flexible.

The article explores a comprehensive approach to using Sketch Engine in teaching professional texts translation, focusing on the practical aspects of working with corpora. The author offers systematic methods that cover both basic skills of working with corpus data (creation of specialised corpora, frequency analysis, terminology research) and advanced techniques of analysing professional vocabulary and specific grammatical structures. Particular emphasis is placed on the practical application of these tools in the educational process, which allows translation students to effectively master the key aspects of specialised translation.

Key words: corpus-based research, teaching specialised texts translation, Sketch Engine, professional text, professional vocabulary, CQL universal query language.

Сучасну прикладну лінгвістику неможливо уявити без мовних корпусів. Вони одночасно є як її потужним та ефективним дослідницьким інструментом, так і практично необмеженою базою лінгвістичних даних для різних галузей та потреб наук, що так чи інакше використовують мовний арсенал. Автоматизація створення та дослідження текстових масивів створює нові можливості не тільки для філології, а й для фахівців, які застосовують такі дані у практичних цілях. Не останню роль корпусні методи відіграють у вдосконаленні навчання мов, а саме перекладу, оскільки дають змогу точно та системно підбирати спеціалізовані мовні матеріали, необхідні для опанування лексики, особливостей вживання та перекладу ключових лінгвістичних одиниць, а також для визначення сучасних мовних тенденцій у конкретній сфері. Незважаючи на величезний вибір платформ та програмного забезпечення для керування корпусами текстів та для текстового аналізу, Sketch Engine займає особливе місце завдяки своїй потужності, а також обсягів та універсальності колекцій текстів: він дозволяє не тільки аналізувати наявні корпуси, а й створювати власні, у тому числі багатомовні, досліджувати лексику, словосполучення, термінологію, перекладацькі відповідники та генерувати навчальні матеріали за допомогою CQL-запитів і вбудованих лінгвістичних функцій. Особливо високу ефективність Sketch Engine демонструє у навчанні роботі з фаховими текстами та їх перекладу, оскільки дає можливість ефективно досліджувати спеціалізовані тексти, виявляти ключову термінологію, характерні словосполучення, аналізувати перекладацькі підходи та готувати навчальні матеріали для студентів-перекладачів; при цьому використання мови запитів CQL дозволяє зробити процес роботи зі спеціалізованими текстами більш спрямованим та гнучким.

У статті досліджується комплексний підхід до використання Sketch Engine у навчанні фахового перекладу, зосереджуючись на практичних аспектах роботи з корпусами. Пропонуються систематизовані методики, що охоплюють як базові навички роботи з корпусними даними (створення спеціалізованих корпусів, частотний аналіз, термінологічні дослідження), так і більш складні прийоми аналізу фахової лексики та специфічних граматичних структур. Особливий акцент робиться на практичному застосуванні цих інструментів у навчальному процесі, що дозволяє студентам-перекладачам ефективно опанувати ключові аспекти фахового перекладу.

Ключові слова: корпусні дослідження, навчання фахового перекладу, Sketch Engine, фаховий текст, фахова лексика, універсальна мова запитів CQL.

The issue at hand. Corpus linguistics has evolved significantly by developing diverse methodologies that apply linguistic analysis to solve both everyday and academic language-related problems. This field encompasses traditional areas such as textual analysis, lexicography, linguistic description, corpus-assisted translation studies, and terminology development [1], alongside with emerging applications like building annotated and tagged corpora for NLP [2], enhancing machine translation systems [3], and advancing AI-driven linguistic modelling [4; 5]. These innovative approaches have demonstrated particular value in language education and translation studies, offering data-driven insights into authentic language use and contextual variations. Within this framework, corpus analysis and related technologies play a pivotal role in translation training, particularly in professional translation teaching. By facilitating the systematic identification of key terminology, frequent collocations, and their translation equivalents, corpus-based methods enhance the acquisition of professional vocabulary and improve specialised translation skills. Such techniques not only accelerate learning but also improve the accuracy and contextual appropriateness of translated content. The process of corpus-based translation training might turn out to be more complex, as it involves both research and analysis of the corpus-based data as well as more basic skills of comprehensive utilization of diverse and multifaceted corpus software functionalities.

The latest research analysis. The field of corpus linguistics has been extensively investigated by numerous scholars from various perspectives, reflecting the broadness of this discipline, with particular attention to its applications in language and translation training. In the domain of applied linguistics, Norbert Schmitt has pioneered the integration of vocabulary frequency analysis with corpus-based techniques to enhance language instruction, whereas Stefan Th. Gries has employed sophisticated statistical approaches to refine vocabulary selection processes across various specialized domains [6]. Significant contributions to data-driven learning approaches in translation education have been made by Abdurrahman Kilimci and Aslı Nur Akkoyunlu, whose empirical research demonstrated substantial improvements in learners' collocational competence together with positive student reception of corpus-based methodologies [7]. Similarly, Amel Lusta, Özcan Demirel and Behbood Mohammadzadeh have successfully implemented corpus linguistics principles to optimize both language teaching and learning outcomes [8]. The Ukrainian academic community has also made substantial contributions to this field, with scholars including T. Anokhina, V. Babych, I. Kobyakova, N. Lemish, S. Matvieieva, S. Schvachko, and A. Zernetska [9; 10; 11] investigating various applications of corpus techniques in language and translation teaching. The discipline continues to demonstrate remarkable dynamics and diversity, with ongoing research developments enhancing the theoretical foundations and practical implementations of corpus-based approaches in educational contexts, and this trend clearly shows signs of intensifying in future scholarly work. This growing body of research highlights the potential of corpus linguistics in revolutionizing language teaching and translation training methodologies through empirical, data-driven approaches that bridge theoretical insights with classroom applications.

The article is aimed at researching how corpus linguistics methodologies can enhance professional translation teaching through two complementary approaches: corpus-based research techniques and hands-on training in corpus compilation and analysis. Focusing on the application of Sketch Engine software, the research demonstrates practical methods for (1) developing essential competencies in building customized corpora for domain-specific translation needs, and (2) identifying high-frequency lexical patterns and specific translation constructions. The article presents a framework for training future translators in critical skills, including text selection criteria for specialized corpora, formulation of CQL queries for terminological extraction, analysis of terminological collocations, and identification of frequent words in professional discourse. Particular emphasis is placed on developing students' ability to construct and utilize specialized corpora, starting from initial text compilation to advanced query design. The study highlights how these dual competencies (practical corpus-building skills and analytical corpus research) enable translation students to create targeted lexical resources for specific domains, while simultaneously developing the meta-skills needed for professional terminological research. For «Translation» and «Applied Linguistics» educational programs, this approach offers both immediate practical benefits through ready-to-use core vocabulary sets as well as long-term methodological value by teaching students to expand and adapt these resources through corpora text analysis. The given model bridges the gap between theoretical corpus linguistics and applied translation practice, equipping students with research-backed strategies to face specialized translation challenges as well as fostering independence in professional terminology management. And this is what gives this study its practical value.

The main body of the article. As we have repeatedly pointed out, the current stage of applied linguistics development facilitates active implementation of corpus technologies in the training of future translators. This is particularly relevant for working with highly specialized texts, where traditional methods often prove insufficient. Sketch Engine, as one of the most functional corpus analysis tools, helps address this issue by enabling the creation of specialized text databases and utilizing the powerful capabilities of CQL queries to identify and analyse key linguistic phenomena [12, c. 7–16]. The practical use of corpus-based methods in teaching and learning encompasses multiple interconnected stages. These may include: corpus selection and corpus design, when the students choose or compile a specialized text corpus relevant to the learning objectives; data querying and analysis with the corpus tools application to extract linguistic patterns, collocations and terminology; hands-on training at which the students are engaged in corpus searches to analyse authentic language use; interpretation and discussion, the stage that could be viewed as an important consolidation phase to refer the findings to the translation or applied linguistics theory; practical implementation which aims to apply the insights to various related translation tasks, text production, or error analysis.

The above-mentioned stages naturally represent the cycle of corpus building and using at the same time enhancing the students professional knowledge and skills acquisition. Here we can be more specific and dwell on practical implementation of the ideas. First, this involves developing skills in working with the Sketch Engine interface, where students learn to create their own corpora, upload texts in various formats (DOCX, PDF), and analyse their structure using the Corpus Info tool. Thus, when working with the «Legal Documents for IT Products» corpus, special attention should be paid to analysing parameters such as the total number of tokens, words, and documents, which helps understand the scale and specificity of the texts.

Second, a crucial aspect involves mastering CQL (Corpus Query Language), which enables advanced terminology search and analysis. The CQL language can pose some problems when working with specialised corpora in a professional texts translation class, as CQL queries are extremely important for correct and efficient work with this type of texts. There are two kinds of difficulties students face here. The first one is the query language itself, because, despite the fact that Sketch Engine offers a short guide to creating queries, as well as the so-called CQL Builder, a tool for prompting that makes writing queries much easier, developing the skills to build the prompts might take time and guidance for a gradual transition from elementary queries to more complex ones. The second problem is that even if students have the skills to write CQL prompts, they do not always fully understand their potential and what kind of result they will get in the process of querying, how they should analyse and interpret the data obtained. Therefore, the primary task at the beginning of working with corpora of professional texts is to show students the basic ways of creating queries and how they can use them to analyse professional vocabulary as well as particular structural features of a professional text. Thus, a number of typical CQL queries can be used to analyse legal and IT texts. For instance, the query [tag="N.*"] identifies all nouns in the corpus, essential for building a terminology database. Similarly, to analyse passive constructions typical of legal texts, the query [tag="V.*"] [tag="VVN"] detects phrases like «is granted» or «are prohibited». For translators, queries targeting multi-word terms are particularly valuable, e.g., [tag="N.*"]{2} is utilized to extract two-component noun clusters, such as «license agreement» or «intellectual property». The following query [word="[A-Z][a-z]+"] is used to identify capitalised terms, which may include proper or company names («Microsoft», «Adobe», «End User Licence Agreement», etc.), and [word=".*-.*"] to find hyphenated combinations («non-exclusive», «user-generated», «up-to-date», etc.). To detect collocations, we can use queries, such as [tag="V.*"] [tag="IN"] for verbs with a preposition («grant under», «comply with», «terminate upon», etc.); [tag="VVN"] [tag="N.*"] for constructions such as «licensed software» or «registered trademark»; and [tag="N.*"]{2,3} for noun clusters of two or three components («license agreement», «intellectual property rights», etc.). To search for grammatical constructions, it is convenient to use [tag="V.*"] [tag="VVN"] for passive voice («is granted», «are prohibited», etc.); [lemma="shall" | lemma="must"] [tag="V.*"] for modal verbs in texts of agreements («shall comply», «must notify», etc.); and [word="if"] [tag="PRP"] [tag="V.*"] for conditional constructions («if the user violates», «if the software is modified», etc.).

In order to develop skills of applying CQL queries in the classroom, several practical tasks can be offered as well. First, students receive a ready-made CQL query, for example, to search for noun clusters, then enter it into Sketch Engine, analyse the results, paying attention to which terms are most common, and then suggest ideas for the Ukrainian equivalents. For instance, the query [tag="N.*"]{2} will help find most common 2-component noun clusters such as «data protection», «user agreement», «copyright notice», which are translated as «захист даних», «угода з користувачем», «повідомлення про авторські права». Students can also search the corpus for a sentence with a particular term using the appropriate CQL query and translate it into Ukrainian, taking into account the context. For example, the query [lemma="infringe"] allows detecting all forms of the word «infringe». As a result, there might be found a typical sentence like «You shall not infringe our intellectual property rights.», which can be translated as «Ви не повинні порушувати наші права на інтелектуальну власність» or «Ви не маєте права порушувати наші права на інтелектуальну власність», etc. Editing is another effective way to practise using CQL language. Students could be given a translation excerpt of an agreement with deliberate mistakes, and they have to use a CQL query to find the correct version in the corpus and correct the error. For example, in the sentence «The licence shall be cancelled» the wording is incorrect, because in legal texts, the construction „shall be terminated» is more commonly used. Using the query [lemma="licence"] [lemma="shall"] [lemma="be"] [tag="VVN"], students find the correct form – «The licence shall be terminated.». Additionally, students can independently compose CQL queries to search for specific constructions, for example, cases where the licence is restricted. Working with real documents might also be useful: you can download EULAs from Microsoft, Adobe or other companies and compare the terminology using CQL, which helps to better understand the peculiarities of legal and IT texts.

Another important aspect is the practical application of the obtained data in the educational process. An effective approach is organizing discussions to analyse specific types of documents, for example

EULAs (End-User License Agreements), where students learn to identify characteristic language features: standard formulations, passive constructions, specific terminology. Group project work, when each team analyses a certain aspect of the text (noun clusters, phrasal verbs, etc.) with subsequent presentation of results, promotes better acquisition of the material. Particular attention should be paid to creating thematic glossaries based on corpus data, as this activity extends far beyond lists of terms; on the contrary, it represents a comprehensive process of developing professional linguistic awareness. Tools like Sketch Engine enable frequency analysis of terms and their typical collocations, revealing not just key vocabulary but also its actual functioning within professional discourse. When the students examine terms like «license», «agreement» or «liability» in their natural contexts, they begin seeing not isolated words but complete communicative patterns. For instance, analysing collocations such as «grant a license», «terminate automatically», etc., reveals both lexical and grammatical features of legal style. This corpus-driven immersion helps students develop an intuitive understanding of professional terminology that differs fundamentally from mechanical dictionary memorization. The approach proves particularly valuable when combined with authentic translation practice. After working with corpora, students translate real text excerpts and then compare their solutions with professional translations. This comparative process provides more than just vocabulary knowledge – it helps develop true translator thinking: the ability to analyse contexts, anticipate interpretations, and justify lexical choices.

Conclusions. To sum up, we have to once again emphasise that corpus linguistic offers powerful tools for researching and analysing various linguistic data as well as a virtually unlimited repository of language data. The automation of corpus creation and analysis has led to the immense popularity of corpus methods for practical applications in language teaching and translation. They are particularly valuable in professional translation training, as they allow for the precise selection of language materials, identification of lexical patterns typical of the field, and analysis of ongoing language trends in specific domains. Sketch Engine is distinguished due to its versatility, vastness of language data, and advanced features like Corpus Query Language (CQL). It enables users to create custom corpora, analyse vocabulary and collocations, and therefore, effectively use the materials in language and translation training. The corpus manager's benefits in teaching professional translation are in its ability to facilitate targeted research on specialized texts, identify both specific and commonly used terminology, and offer tailored learning resources. The methodology of corpus-based specialised translation teaching can be divided into stages, each focusing on the particular tool or part of the corpus engine with gradual increase of the tasks complexity as new knowledge and skills are acquired. At the initial stage students learn to extract linguistic patterns, collocations, and terminology through CQL queries. With more practice students engage in various kinds of corpus search to analyse authentic language use. An important part of the educational process is to link translation theory and corpora practical applications, which would foster deeper understanding of terminological and structural processes typical of specialised texts of the given field. No less important is the application of the corpus research and analysis methods to the translation tasks, text production and linguistic errors analysis.

The corpus-based approach to teaching specialised texts translation offers several advantages among which is enhanced terminological accuracy, as the students, being exposed to authentic language use combined with linguistic understanding of the specialised text production processes, improve their ability to identify and thus translate the terms appropriately. Analysing collocations and grammatical structures helps students build up contextual understanding of the specialised vocabulary. Eventually, learning to create and query corpora, students develop skills for independent terminology research and, therefore, gain higher level of translation autonomy.

The prospects of the study include corpus analysis application to specialized texts in various fields, like medicine, finance, or engineering, as this type of linguistic search is highly suitable for specialised language and translation teaching. Terminology is a remarkably dynamic phenomenon, which means that corpus-based methods might be particularly adjusted to identification and analysis of emerging terms in professional discourse. Lastly, and of no lesser significance is the infinite possibilities to employ parallel corpora to study translation equivalents across languages for both training and professional purposes.

BIBLIOGRAPHY:

1. Peñas A., Verdejo F., Gonzalo J. Corpus-Based Terminology Extraction Applied to Information Access. UCREL Technical Papers, 13. Presented at the Corpus Linguistics 2001 conference, Lancaster University, United Kingdom. pp. 458–465.

2. Cabré Castellví M.T., Estopà Bagot R., Vivaldi Palatresi J. Automatic Term Detection: A Review of Current Systems. *Terminology.* 2001. Vol. 7(2). pp. 53–88. DOI: 10.1075/term.7.2.07cab

3. Hewavitharana S., Vogel S. Enhancing a Statistical Machine Translation System by Using an Automatically Extracted Parallel Corpus from Comparable Sources. Proceedings of the LREC 2008 Workshop on Building and Using Comparable Corpora. Marrakech, Morocco, 2008. pp. 7–10.

4. Domhan T., Hasler E., Tran K., Trenous S., Byrne B., Hieber F. The Devil Is in the Details: On the Pitfalls of Vocabulary Selection in Neural Machine Translation. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2022). 2022. Association for Computational Linguistics. pp. 1840–1851. https://doi.org/10.18653/v1/2022.naacl-main.136

5. Van Eck N.J., Waltman L., Noyons E.C.M., Buter R.K. Automatic Term Identification for Bibliometric Mapping. *Scientometrics.* 2010. Vol. 82(3). pp. 581–596. DOI: 10.1007/s11192-010-0173-0

6. Gries S. T. Analyzing Linguistic Data: A Practical Introduction to Statistics Using R (2nd ed.). Cambridge University Press. 2021. 374 pages.

7. Akkoyunlu Aslı, Kilimci Abdurrahman. Application of Corpus to Translation Teaching: Practice and Perceptions. *International Online Journal of Education and Teaching.* 2017. Vol. 4. pp. 369–396.

8. Lusta A., Demirel Ö., Mohammadzadeh B. Language Corpus and Data Driven Learning (DDL) in Language Classrooms: A Systematic Review. *Heliyon.* 2023. Vol. 9. e22731. 10.1016/j.heliyon.2023.e22731.

9. Anokhina T., Kobyakova I., Schvachko S. Innovative Methodology for Teaching European Studies Using a Corpus Approach. *Philological Treatises.* 2023. Vol. 15. No. 2. pp. 7–16.

10. Matvieieva S. A., Lemish N. Ye., Zernetska A. A., Babych V. I., Torgovets M. S. English-Ukrainian Parallel Corpus: Prerequisites for Building and Practical Use in Translation Studies. *Studies about Languages.* 2022. Vol. 1. pp. 61–74.

11. Lemish N. Ye., Aleksieieva O. M., Denysova S. P., Matvieieva S. A., Zernetska A. A. Linguistic Corpora Technology as a Didactic Tool in Training Future Translators. *Information Technologies and Learning Tools*. 2020. Vol. 79. No. 5. pp. 242–259.

12. Kilgarriff A., Baisa V., Bušta J., Jakubíček M., Kovář V., Michelfeit J., Rychlý P., Suchomel V. The Sketch Engine: Ten Years On. *Lexicography*. 2014. Vol. 1(1). pp. 7–36. DOI: 10.1007/s40607-014-0009-9.

Дата надходження статті: 16.07.2025
Дата прийняття статті: 21.08.2025
Опубліковано: 16.10.2025