

UDC 81.33

DOI <https://doi.org/10.32782/2522-4077-2025-215-8>

CORPUS TECHNIQUES FOR SPECIALIZED TEXTS TRANSLATION: DECONSTRUCTING COMPLEXITY FOR TRANSLATOR TRAINEES

КОРПУСНІ ТЕХНОЛОГІЇ ДЛЯ ПЕРЕКЛАДУ ФАХОВИХ ТЕКСТІВ: ПОДОЛАННЯ ТРУДНОЩІВ У ПІДГОТОВЦІ ФАХІВЦІВ

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*

Language corpora are a powerful tool not only for philologists and researchers, but also for translators and applied linguists. Their use opens up new, previously unknown opportunities for more effective and in-depth linguistic research and more accurate and faster work while translating and editing texts. However, the productivity of corpus tools directly depends on «corpus literacy», i.e. the range of knowledge about corpora, their types, features, and capabilities, as well as practical «corpus skills», and, last but not least, on mastery of the coding basics for solving linguistic tasks, in particular for constructing queries in Python or with the help of CQL. It follows that the effectiveness of corpus research or its use is dependent not only on philological or translation competence, but also on specific knowledge and skills directly related to the technical side of dealing with corpora. When it comes to training future translators, especially if they do not have the technical skills to work with corpora yet, their use can become an obstacle to acquiring the necessary skills, both translational and applied. To prevent such problems, it is important to realize that corpus skills, similar to translation skills, must be acquired gradually, from simple to more complex, and that the ability to work with a corpus must be improved step by step, just like translation skills. This rather complex interdependent process requires a well-thought-out strategy for presenting the teaching material, which is often not straightforward and needs flexibility in changing and applying different methods, depending on the initial knowledge and skills of the philology students.

The article offers a deeper look at a comprehensive approach to teaching professional translation, given that linguistics students often lack the skills and abilities to work with corpora. A step-by-step guidelines for working with specialized texts using corpora is proposed, which allows not only to teach students the basics of translating highly specialized texts, but also to develop their skills in working with corpora, the correct use of built-in corpus tools, and the implementation of accurate and effective search queries using corpus language. Particular attention is paid to the interconnection between specialized translation training and the translation of subject-specific vocabulary, as well as professionally oriented work with language corpora.

Key words: corpus research, specialized translation training, Sketch Engine, specialized text, subject-specific vocabulary, corpus search query.

Мовні корпуси є потужним інструментом не лише для філологів-науковців, а й для перекладачів та прикладних лінгвістів. Їх використання відкриває нові, до цього незвідані можливості для більш ефективних та глибоких лінгвістичних досліджень та більш точної та швидкої роботи під час перекладу та редагування. Проте, продуктивність корпусних інструментів безпосередньо залежить від «корпусної грамотності», тобто від діапазону знань про корпуси, їх види, особливості та можливості, від практичних «корпусних умінь», й у неостанню чергу – від володіння основами програмування для вирішення лінгвістичних задач, зокрема для побудови запитів мовою Python або за допомогою CQL. З цього випливає, що ефективність корпусних досліджень та їх використання залежить не лише від філологічної чи перекладацької компетентності, а й від специфічних знань та умінь, пов'язаних безпосередньо з технічною стороною роботи з корпусами. Якщо йдеться про підготовку майбутніх перекладачів, особливо якщо вони ще не володіють суто технічним потенціалом роботи з корпусами, то їх використання може обернутися на перешкоду на шляху до здобуття необхідних умінь та навичок, як перекладацького, так і прикладного характеру. Щоб запобігти виникненню таких про-

блем, важливо усвідомлювати, що навички роботи з корпусами, так само як і перекладацькі компетентності, мають набуватися поступово, від простого до більш складного, при цьому вміння працювати з корпусом має удосконалюватися так само покроково, як і навички перекладу. Цей доволі складний взаємозалежний процес вимагає продуманої стратегії подачі навчального матеріалу, яка часто не є прямолінійною і змушує гнучко змінювати та застосовувати різні підходи, залежно від вихідних знань та умінь студентів-філологів.

Стаття пропонує більш глибокий погляд на комплексний підхід до навчання фахового перекладу, за умови недостатнього володіння студентами-лінгвістами навичками та вміннями роботи з корпусами. Пропонується поступальний підхід до роботи з галузевими текстами із залученням корпусів, який дозволяє не лише навчити майбутніх фахівців азів перекладу текстів вузькоспеціалізованої тематики, а й розвинути їх навички роботи з корпусами, правильному застосування вбудованих корпусних інструментів, а також створенню коректних та ефективних пошукових запитів за допомогою корпусної мови. Особливу увагу приділено взаємозв'язку навчання фаховому перекладу та перекладу фахової лексики і професійно орієнтованій роботі з мовними корпусами.

Ключові слова: корпусні дослідження, навчання фаховому перекладу, Sketch Engine, фаховий текст, фахова лексика, корпусний пошуковий запит.

The issue at hand. Teaching specialized translation to university students can be a problem by itself, as it requires not only good mastery of the source and the target languages, but also at least basic competence in the subject field the texts belong to, as well as understanding of the essential principles of specialized text structure, vocabulary and grammar peculiarities. That is why, when it comes to the introduction of corpora techniques into the translator training process, the logical question is immediately raised of whether it is worth further complicating the training in specialized translation, despite the prospects that corpora offer [1]. The answer hugely depends on the methodology applied within the course. There are various premises we can rely on. Suppose the students are familiar with the basic techniques of corpora analysis: in this very case corpus capabilities will immensely enhance the studying process. However, in the majority of cases, the students know very little to nothing of the potential and efficient usage of corpora engines in their professional activities. Here, a thoroughly-considered and calculated method should be applied for the corpus research to become a facilitator of the specialized texts translation teaching process [2].

The latest research analysis. The field of corpus linguistics has been widely explored by many scholars from different perspectives, which reflects the broad scope of this discipline, with particular emphasis on its use in language and translation training. In applied linguistics, Norbert Schmitt has been a leading figure in combining vocabulary frequency analysis with corpus-based methods to improve language teaching, while Stefan Th. Gries has applied advanced statistical techniques to refine vocabulary selection across specialized domains [3; 4]. Important contributions to data-driven learning in translation education have been made by Abdurrahman Kilimci and Asli Nur Akkoyunlu as well as by Ying Lyu and Ziman Han, whose empirical studies showed notable progress in learners' collocational competence, along with positive feedback from students on corpus-based approaches [5; 6]. Likewise, Amel Lusta, Özcan Demirel, and Behbood Mohammadzadeh have effectively used corpus linguistics principles to enhance both teaching and learning outcomes [7]. Ukrainian scholars have also played a significant role in this area, with T. Anokhina, V. Babych, I. Kobyakova, N. Lemish, S. Matvieieva, S. Schvachko, and A. Zernetska [8; 9; 10] examining various ways of applying corpus techniques in language and translation pedagogy. The discipline continues to show strong development and diversity, with ongoing research strengthening both its theoretical basis and practical applications, a trend that is clearly expected to grow further. This expanding body of work highlights the potential of corpus linguistics to transform language teaching and translation training through empirical, data-driven approaches that link theory with classroom practice.

The article aims to show the effective ways of ensuring high-performance collaboration of traditional translation teaching methods of specialized translation and corpora-based techniques, analysing a particular case and meticulously identifying the stages and the strategies applied at each stage to achieve the highest degree of teaching efficiency.

The main body of the article. The very first step in the process of corpora-based translation training is the selection of the specialised field, which is presumably defined by the academic curriculum, and searching for the appropriate corpus within the given corpora engine. The fact that the corpus can be also created by the teacher with the view to more efficient training process makes this phase highly flexible and easily adjusted. We begin by introducing students to the corpus of the license agreements, detailing its overall volume and key statistical features. This initial step, however, goes far beyond mere familiarization. Its deeper value lies in overcoming the psychological barrier that often accompanies specialized texts. It's common for students to feel intimidated by such specific content, so we demonstrate that these documents are not incomprehensively complex but rather structured systems that can be broken down and analysed through the metrics like token counts and unique word frequencies. Also it is worth while demonstrating why the specialized texts seem difficult to translate, highlighting the most typical features of the latter (specific terms, terminological precision, term ambiguity, complex sentence structure). This analytical approach reframes the students' perception, reducing initial stress and building confidence. Moreover, this process fosters a professional mindset by training the students to view translation not merely as an art but as a managed project requiring thorough source data analysis. Ultimately, this isn't an abstract exercise: it's contextualized learning that prepares them for working with real-world documents signed by millions every day.

Building upon the foundational introduction to the corpus, the next phase of the learning process might be strategically designed around a list of the 500 most frequent lexemes, which include nouns, verbs, and adjectives essential to the field. The important part of this work includes training skills of corpus units selection by frequency. The primary objective of this focused approach is to help students get acquainted with the core vocabulary necessary for navigating IT-legal documentation with confidence. To achieve this, the module incorporates targeted practical exercises. For instance, the students begin by translating key individual terms (license, agreement, software, liable, infringement, etc.), but they do so not automatically. Instead, they are required to identify the most accurate Ukrainian equivalent by examining these terms within authentic contexts, utilizing tools like the Concordance feature in Sketch Engine or pre-selected examples provided by the teacher. This method ensures that the students' understanding is based on real usage rather than abstract definitions. Furthermore, to promote active recall and long-term retention, each student then creates detailed term flashcards. These cards go far beyond simple word-pairing; they include the translated term, a clear example sentence extracted directly from the corpus, and crucial translator's notes on usage, nuances, and potential pitfalls. By doing so, this continuous and logically integrated process bridges the gap between recognizing a word and being able to utilize it accurately and appropriately in a professional translation context. The deep value of this stage lies in its immense efficiency, directly applying the Pareto principle: by focusing efforts on learning these core 500 words – which represent roughly 20% of the lexical effort – students gain the ability to understand approximately 80% of a typical text in the domain. This strategic focus accelerates the learning process. The students don't waste time on marginal vocabulary and instead concentrate their attention on the lexical core, which yields rapid and visible results. Furthermore, this approach provides a powerful motivational boost. After just a few sessions, students can already operate this basic set of terms and comprehend significant portions of professional texts, which gives them a sense of progress and achievement [11, p. 5–6]. Ultimately, this set of 500 lexemes is not just a word list; it functions as the essential framework upon which all subsequent knowledge – including collocations, grammatical structures, and more specific terminology – will be systematically built.

An inseparable part of any specialized text is typical terminological collocations, which, on the one hand, are highly representative of the terminological, grammatical and stylistical peculiarities of the domain, and are of particular difficulty for the translator trainees with the need of finding accurate equivalents in the target language. Therefore, our following important step in the development of corpora-based translation skills will be closely linked to the problem. Sketch Engine has several

powerful tools for selection and analysis of terminological collocations and typical word combinations in texts, among them are CQL-language and Sketch Engine Keywords with its in-built function of keywords detection (Identify Keywords). For the stage to be productive, it is essential to familiarize the students with the functional tabs in Sketch Engine programs, show how to select collocations and sort them out by frequency, using the given corpus or any similar corpus in the database. There is also the need to teach the students how to build basic CQL queries, however, at this stage to might be overwhelming for the students, so it is more appropriate to provide them with a set of ready-made prompts specially designed for searching typical terminological collocations, such as verbs plus prepositions combinations as well as phrasal verbs, participles plus prepositions, verbs plus adverbs, etc. The acquisition of the above mentioned skills and knowledge naturally leads to the next critical phase of training: the analysis and translation of collocations. This will form the core of sessions dedicated to working with typical word combinations. The primary objective here is to train students to not only identify but also accurately translate genre-specific multi-word units. The practical application of the information extracted from the corpus is paramount. To this end, a series of practical exercises shall be employed. These might include collocation reconstruction tasks, where students are given sentences with gaps and therefore must fill them using the appropriate collocations. Furthermore, collocation translation exercises will be implemented, providing students with a list of English collocations and challenging them to find and propose established equivalent Ukrainian counterparts, fostering the crucial link between the source and target language phrases and moving from passive recognition to active recall. To enhance a deeper understanding of translation peculiarities beyond simple one-to-one matching, multiple-choice tasks could also be introduced. Here, students must select the most accurate translation out of several options, evaluating subtle differences in meaning, register, and domain precision. This analytical skill is further developed through contextualized translation exercises, where learners translate full sentences containing the target collocations, thus practicing their correct integration into syntactically and stylistically appropriate specialized text. The module then advances to productive activities, challenging students through sentence generation tasks to create their own original sentences in both English and Ukrainian, which demonstrates full productive mastery and ownership of the typical terminological collocations. This creative application is balanced with critical analysis of common translation errors, where students diagnose and correct inaccurate versions, explicitly addressing false friends and calques to build resilience against making typical mistakes. Finally, all acquired skills are synthesized in a comprehensive text-level analysis and translation. Students engage with an authentic or adapted specialized text filled with the studied collocations, requiring them to identify, comprehend, and accurately reproduce the prominent features of the genre in the target language. This concluding activity ensures that the learned collocations are not treated as isolated items but as integral components of a coherent professional text. The strategical shift towards the analysis and translation of collocations signifies a fundamental transition from learning vocabulary in isolation to mastering language as a system of pre-constructed blocks. This approach defies a critical challenge in translator training: the move from “dictionary-based” to “natural” production. A common pitfall for beginners is the tendency to translate words individually and then attempt to combine them, which often results in unnatural and erroneous texts. Working with collocations trains students to perceive and process language in a holistic way, a definitive characteristic of high-level translation proficiency. Furthermore, this method is crucial to prevent calquing, as collocations are where the most subtle and frequent translation errors occur. For instance, the English collocation ‘make a decision’ corresponds to the Ukrainian «ухвалити рішення» and not a literal, word-for-word translation. By analysing authentic examples from linguistic corpora, students develop an evidence-based understanding of correct usage, effectively eliminating mistakes resulting from the attempts to translate the collocations straightforward. Ultimately, repeated exposure to these recurring word combinations in diverse contexts develops a genuine «feel for the language», fostering an intuitive understanding of

what natural or idiomatic expressions are. The latter enhances both the accuracy and fluency of the students' translanguaging production.

The following sessions are dedicated to the critical analysis of multi-component terms and complex noun clusters, one of the defining features of specialized texts. This phase moves beyond simpler collocations to address the challenges posed by more complicated terminological units such as «end user license agreement» or «software license grant». The primary objective is to equip students with a methodological framework for accurate deconstruction and translation of these intricate structures. The process begins with a comprehensive structural analysis, where students learn to parse each term into its constituent components to identify the semantic head, the core element, and its modifying parts. This analytical step is fundamental, as it reveals the logical relationships within the cluster and prevents misinterpretation. Afterwards, the focus shifts to translation strategy. Students engage in critical discussion and practice to determine the most appropriate approach for each term: whether to employ a calque (if it is established and unambiguous), to seek a functional analogy term in the target language, or indulge into a descriptive translation when a direct equivalent is missing. For instance, the acronym EULA (End User License Agreement) necessitates a strategic choice, resulting in the well-adjusted Ukrainian translation «Ліцензійна угода з кінцевим користувачем», which carefully balances accuracy and naturalness in the target language. This comprehensive exercise ensures that students can demonstrate confidence dealing with noun clusters. As we can see, practising with multi-component terms constitutes a critical stage in specialized translation training, focusing on the deconstruction of complex terminological units into their constituent parts. This process is not merely about memorizing terms but is aimed at developing analytical thinking. The student learns to understand the underlying logic of a term rather than relying on rote learning. For instance, deconstructing «end user license agreement» reveals its logical architecture: an agreement concerning a license intended for an end user. This cultivated skill empowers the future translators to independently analyse and accurately decode even unfamiliar terminology in the future. Consequently, this method ensures translation precision, as misidentifying the head noun within a noun cluster frequently leads to significant errors in terms of meaning. Therefore, this practice effectively boosts attention and accuracy, which are paramount for producing reliable and professional translations.

Conclusions. In summary, this article has outlined a clear, step-by-step method for successfully combining traditional methods with modern corpus-based tools for specialized translation training [12]. The given model moves logically from first introducing the corpus to finally analysing complex terms, showing that corpora are much more than just extra resources. They could become a central part of learning that significantly improves and speeds up how students gain real translation skills.

The strength of this combined approach lies in its careful design. It starts by breaking down the fear of specialized texts, showing students they can be analysed like data to build confidence. Next, it uses frequency lists to focus on the most important words, making the process highly efficient. A key step then is moving to the mastery of common word combinations (collocations), which tackles a major problem in translation: shifting from translating single words to working with whole, natural language blocks. This is essential for creating accurate and fluent translations and for avoiding common errors. The final stage trains students to break down and understand complex terms, developing the sharp analytical skills needed for specialized texts and enabling them to solve problems on their own. In the end, this method goes beyond theory to provide practical learning. It provides students not only with the specialized vocabulary translation skills, but, more importantly, with a reliable working method and a professional way of thinking. By moving from terms, to collocations, to complex phrases, the process ensures that students progress from simply understanding words to being able to confidently produce accurate and natural translations of real specialized documents. This thorough preparation is crucial for meeting high demands of today's translation market.

As for the prospects, the future research could apply the exact same phased strategy outlined here, which uses IT-legal documentation as a case study, to other specialized domains, such as medical,

financial, or technical translation; furthermore, a comparative study could then analyse whether the efficiency of the model varies across fields and identify any domain-specific adjustments needed. In terms of the technical aspect, a prospective study could compare learning outcomes and efficiency when using different corpus analysis platforms (e.g., AntConc, LancsBox, Korpusomat) or even comparing pre-built corpora with smaller, teacher-compiled corpora tailored to a very specific sub-field. Apart from the immediate acquisition of skills, a long-term study could track students over a semester or a year to measure the retention of terminology and collocations learned through this method. Also it might be worth investigating how well these analytical skills transfer to translating new types of specialized texts. In a similar practical vein, research could explore placing this approach within a collaborative, project-based framework, for example, by tasking student teams with building a specialized corpus for a particular domain and then using it to create a terminological base.

BIBLIOGRAPHY:

1. Lu J., Xu Y. Practical Research on Corpus-based Translation Teaching Modes of College English. *International Journal of Education and Humanities*. 2023. 8(2). pp. 241–245. <https://doi.org/10.54097/ijeh.v8i2.7822>
2. Pietrzak P., Kornacki M. Focus on the Translation Trainee. *Research in Language*. 2021. 19(2). pp. 107–115. <https://doi.org/10.18778/1731-7533.19.2.01>
3. Schmitt N. *Vocabulary in Language Teaching* (2nd ed.). Cambridge University Press. 2020. 304 p.
4. Gries S. T. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R* (2nd ed.). Cambridge University Press. 2021. 374 p.
5. Akkoyunlu Ashi, Kilimci Abdurrahman. Application of Corpus to Translation Teaching: Practice and Perceptions. *International Online Journal of Education and Teaching*. 2017. Vol. 4. pp. 369–396.
6. Lyu Y., Han Z. Applying data-driven learning in self-translation of academic discourse: A case study of a Chinese medical student. *Frontiers in Psychology*. 2023. 14. 1071123. <https://doi.org/10.3389/fpsyg.2023.1071123>
7. 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](https://doi.org/10.1016/j.heliyon.2023.e22731).
8. 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.
9. 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.
10. 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.
11. Cruz E. C. Teach Smarter, Not Harder: The Pareto Principle in the ELT Classroom. Proceedings of the 16th International ELT Conference: «Ideal vs. Reality». Muscat: Sultan Qaboos University. 2016. pp. 1–9.
12. Yan D., Wang J. Teaching data science to undergraduate translation tragoopinees: Pilot evaluation of a task-based course. *Frontiers in Psychology*. 2022. 13. 939689. <https://doi.org/10.3389/fpsyg.2022.939689>

Дата надходження статті: 03.10.2025

Дата прийняття статті: 10.11.2025

Опубліковано: 30.12.2025