UDC 81.33
DOI https://doi.org/10.32782/2522-4077-2025-213-23

CORPORA-BASED ANALYSIS OF SPECIALISED TEXTS FOR TRANSLATION TRAINING: TERMS AND NEOLOGISMS

КОРПУСНИЙ АНАЛІЗ ФАХОВИХ ТЕКСТІВ ДЛЯ НАВЧАННЯ ПЕРЕКЛАДУ: ТЕРМІНИ ТА НЕОЛОГІЗМИ

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

Language corpora are one of the most effective tools of applied linguistics, which are actively used in various fields of human life. The automated selection, compilation and analysis of text corpora of virtually unlimited size open up new perspectives not only for linguistic research, but also for professionals who use such data to solve practical problems. Corpus-based methods have great potential for improving language teaching, including translation, as they allow for the accurate and targeted selection of specialised linguistic materials necessary for mastering the lexical minimum, peculiarities of usage and translation of key language units, as well as for identifying current language trends in a particular field.

Among the tools for working with corpora, Sketch Engine stands out as one of the most powerful, as it does not only analyse existing corpora but also creates your own, including multilingual ones. This makes it possible to quickly and efficiently research professional texts, identify key terminology and common phrases, analyse translation strategies, and create training materials for future translators. The use of the CQL query language allows improving search accuracy and obtaining more relevant linguistic data.

The given article, which is a part of a larger study, discusses such an important function of Sketch Engine for searching, analysing and selecting lexical material as term recognition and extraction using the built-in Sketch Engine Keywords tool. This tool not only allows to identify terms and term combinations in professional texts with high accuracy, but also to compare the frequency of use of such words and combinations in both the studied and the reference corpora, which significantly increases the efficiency of search in general and linguistic analysis of selected units in particular. Another aspect of this study is the methodology of corpus search for neologisms and rarely used words. The latter is a challenge for corpus-based text analysis, as there are no universal search formulas or even principles for finding such vocabulary, which, however, is an important component of professional texts. The study is based on a corpus of English-language legal texts related to the IT sector, including licence agreements and contracts.

Key words: corpus-based research, corpus, professional text, core vocabulary, term, neologism, Sketch Engine, CQL universal query language.

Мовні корпуси – один із найефективніших інструментів прикладної лінгвістики, що активно застосовуються в різних сферах діяльності. Автоматизований підбір, компіляція та аналіз текстових масивів практично необмеженого обсягу відкривають нові перспективи не лише для філологічних досліджень, а й для фахівців, які використовують такі данні для вирішення практичних завдань. Корпусні методи мають значний потенціал для вдосконалення викладання мов, зокрема перекладу, оскільки дозволяють точно та цілеспрямовано відбирати спеціалізовані лінгвістичні матеріали, необхідні для засвоєння лексичного мінімуму, особливостей вживання та перекладу ключових мовних одиниць, а також для виявлення актуальних лінгвістичних тенденцій у конкретній галузі.

Серед інструментів для роботи з корпусами Sketch Engine виділяється як один із найпотужніших, оскільки забезпечує не лише аналіз існуючих корпусів, а й створення власних, у тому числі багатомовних. Це дає змогу швидко та ефективно досліджувати галузеві тексти, виявляти ключову термінологію, типові словосполучення, аналізувати перекладацькі стратегії та складати навчальні матеріали для майбутніх перекладачів. Використання мови запитів CQL дозволяє підвищити точність пошуку та отримувати більш релевантні лінгвістичні дані.

У пропонованій статті, що є продовженням більш масштабного дослідження, розглядаються така важлива функція Sketch Engine для пошуку, аналізу та відбору лексичного матеріалу, як розпізнавання та екстракція термінів за допомогою вбудованого інструмента Sketch Engine Keywords. Цей інструмент не лише дозволяє з високою точністю ідентифікувати терміни та термінологічні сполучення у фахових текстах, а й порівнювати частотність вживання таких слів та сполучень у досліджуваному та референтному корпусах, що значно підвищує ефективність пошуку загалом та лінгвістичного аналізу відібраних одиниць зокрема. Іще одним аспектом даного дослідження є методика корпусного пошуку неологізмів та рідковживаних слів. Останній являє собою певний виклик для корпусного текстового аналізу, адже не існує універсальних пошукових формул або навіть принципів пошуку такої лексики, яка, однак, є важливою складовою фахових текстів. Дослідження виконано на основі створеного корпусу англомовних юридичних текстів, пов'язаних із IT-сферою, зокрема ліцензійних угод і договорів.

Ключові слова: корпусні дослідження, корпус, фаховий текст, термін, неологізм, Sketch Engine, універсальна мова запитів CQL.

The issue at hand. Emphasizing the practical application of linguistic analysis through language corpora and methods to address everyday and scholarly language problems, corpus linguistics has continuously developed a wide range of approaches: including both conventional fields, namely, text analysis, lexicography, language description, corpus-based translation studies, terminology development, and more recent endeavours, such as creating annotated corpora for natural language processing, machine translation systems, and linguistic modelling for artificial intelligence. Notably, the new methods have proven to be quite useful in both language teaching and translation studies, with new corpora-supported insights into real-world language usage patterns and contextual shifts. Through the efficient selection of core vocabulary, most common collocations, and their translation counterparts, corpus-based techniques assist in the learning process of field translation. They enable faster vocabulary acquisition and field translation proficiency and significantly improve the accuracy and contextual precision of language transfer. Therefore, one of the current topics in translation teaching approaches seems to be comprehending the mechanism of field vocabulary selection and usage.

The latest research analysis. Corpus linguistics is the realm of studies which is approached by scholars from diverse perspectives. Researchers have been examining its connections to computational linguistics, language acquisition, and machine translation, using corpus methods to develop both analysis and teaching practices. The uniqueness of the discipline lies in maintaining active research while adapting to new digital tools, thus making itself continuously relevant in the field of scientific investigations. Scholars who have contributed to the filed most are the representatives of computational models developers, applied linguists who perfect teaching methods, and researchers in the sphere of quantitative linguistics. What drives them together is their reliance on the empirical, data-driven framework which develops with the improvement of digital technologies. Thus, the issue of vocabulary extraction and selection for various linguistic purposes has been a target area with the focus on NLP research [1] as well as on neural machine translation [2]. Another significant aspect of corpus-based research, which is of special interest in our studies, is language acquisition through the perspective of corpus analysis. It is here where the DDL approaches comes to the fore [3; 4]. The issue of corpus-based terminology identification and extraction, which is analysed in the given article, is of interest to many reputable linguists [5; 6; 7; 8; 9], their studies revealing boundless possibilities for terms research. Neologisms, presenting a problem for corpus linguists, is also a potentially promising direction of automated linguistic research, with the attempts directed towards new words identification using both quantitative and qualitative approaches to detect the semantic shift [10; 11; 12]. There is also a bunch of home scholars successfully investigating the potential use of corpora to facilitate both linguistic research and language acquisition processes [13; 14; 15]. With new study findings appearing every year, the field is extremely dynamic and diversified, and this trend is only predicted to get stronger in the future.

The article is aimed at examining the potential of corpora and the capability of corpus software to support the teaching of translation. Using the corpus generated in Sketch Engine, the research aims

to demonstrate how to choose the most frequently used lexical elements and collocations with them, as well as constructions that provide specific translation challenges. The latter might develop into an essential tool for translation students to build a lexicon for efficient vocabulary learning and analysis of the nuances of translating specialized professional texts. Special attention in the article is given to the ways of selection and analysis of terms and terminological collocations as well as to the ways of new words extraction. The potential to directly use the developed core vocabulary for teaching field translation to students in the «Translation» as well as «Applied Linguistics» specialities and the opportunities for expanding the core vocabulary's features through additional corpus analysis of parallel bilingual texts are what gives this study its practical value.

The main body of the article. To ensure objective and yielding results of the research we had to create a corpus of specialized professional texts belonging to the same stylistic realm. As the legitimacy and effectiveness of the linguistic data gathered depend on the sequence and precision of the steps involved in creating a core vocabulary utilizing corpus management software, the first stage included the selection of core texts, adding them to a fresh bulk, and examining the lexical content of the results. A total of thirty software product legal support papers, including English-language end-user licensing agreements (EULAs), were chosen. The DOCX and PDF versions of the texts were uploaded to the Sketch Engine corpus management and text analysis program. The Corpus Info page displays the output data, which includes the total amount of words, phrases, documents, and tokens. Details on the quantity of distinct units in the corpus, such as separate words, tags, or lemmas, are provided by the LEXICON SIZES block. Token coverage and document type information are shown in the TEXT TYPES block.

An important dimension of our complex research was the possibility of using the term recognition and extraction function with the built-in Sketch Engine Keywords tool, which allows us to automatically find monocomponent corpus keywords, multicomponent terms in the format of the corpus terminology language, and high-frequency combinations that are specific to this corpus compared to the reference one [7, p. 459–460]. Since the glossary is intended to be used for training future specialists in professional translation, the selection of such vocabulary is an important component of the lexical minimum. So, in the Advanced tab, we selected the Identify Terms and Identify Keywords options and obtained two lists with the corpus keywords and characteristic terminological combinations, respectively. Since the principle of selection and ordering of both lists was the higher frequency of use in the given corpus compared to the reference corpus (English Web 2021 (enTenTen21) by 52,268,286,493 tokens), the resulting lists were useful for compiling a glossary of the most common words and terminological combinations in the texts of IT product licence agreements.

The Advanced tab also enables us to choose to focus on more or less frequent words in the reference corpus: moving the Focus slider to 'rare' will give you the lists of words sorted from most to least unique ones in the corpus, and vice versa. Both sorting methods proved to be useful for selecting the thematic lexical minimum, as more rare words represent the uniqueness of the corpus vocabulary, and more common words and collocations help to effectively work on more general vocabulary that is typical of legal language and the language of agreements and contracts in particular.

As for the relevance and validity of the keywords and term combinations, the results were ambiguous. While the keywords included a large amount of «noise», which is approximately 30%, the term combinations were selected more efficiently, with only about 10% of the expressions being invalid. The outcome can be explained by the complicated algorithm of corpora term identification [8, p. 83–85]. In any case, such lists were definitely useful for compiling a glossary of corpus texts.

The other dimension of the frequency glossary is neologisms, or words that are quite rare. Finding such words is problematic in Sketch Engine, as the program often tries to mark an unfamiliar word as a part of speech based on the context or lemmatises it, equating the lemma with the original word. On the other hand, the vocabulary of legal documents does not tend to expand its composition with neologisms and is quite stable and conservative, while the IT sector may introduce a certain amount

of new vocabulary due to its rapid development. However, we could not find a single effective way to find neologisms. There is the option that they are not represented in the texts, though. The CQL was used to search for an unrecognised lemma (‘[]’), but the system identified literally all the words in the corpus, possibly due to the peculiarities of text marking. Using Wildcard, as an alternative, to find words with productive suffixes or prefixes, such as ‘[word=’.*ware‘]’ or ‘[word=’self*‘]’ or ‘[word=’cyber.*‘]’, did not bring any unexpected results either. Another potentially promising way to find new words in English texts is to search for compound words formed by combining stems or words that are written together or hyphenated. To look for compound words written together, we used the following CQL expression:

• [word=‘^[A-Z][a-z]+[A-Z][a-z]+$’] – search for a continuous word form with two bases (21 units found).

The query was ineffective, as the retrieved units were mostly names of digital products. Our next query was an expression to search for hyphenated compound words:

• [word=‘[A-Za-z]+-[A-Za-z]+’] – search for compound words which were hyphenated (210 items found).

The latter was more successful, as it allowed us to select unconventional word forms that are quite uncharacteristic for the genre of formal licence agreements. To analyse the words of this query, we used the Relative Frequency parameter, since words with a frequency of less than 5-10 per million are considered rare. We identified 127 lexemes that satisfy this parameter. Among them, there were many formations with semi-prefixes («non-Toshiba», «re-performance») and prepositions («version-up», «one-off»), as well as abbreviations, such as «AI-enhanced», «XML-based», and frequently used words with non-traditional spellings («micro-fibre», «human-readable»).

Thus, a comprehensive approach to the problem has yielded some results, but it should be remembered that not all rare words are neologisms, as, for example, infrequent use may indicate that the word is no longer utilised in texts of a particular filed. It should also be borne in mind that, although legal discourse is highly stable, the language of the IT sector is traditionally full of neologisms due to the rapid development of technology, which means that new words quickly lose their novelty and become standardised. So the words we find are rare, but not necessarily new. The last query prompted us to search for all the words with low relative frequency in the corpus using the Wordlist tool, which we have already tried sorting words by their frequency by default. However, it turned out that the words with low relative frequency account for 31.8% (1,640 units) of all words in the corpus (5,195 units excluding «non-words», as marked by Sketch Engine), and therefore considering them is not optimal for finding rare units. It is an established fact that one of the unique ways of word formation in English is conversion, which is extremely productive for vocabulary enrichment. Thus, we tried to find words tagged as both nouns and verbs in the corpus, hoping to find word formations among them. Yet, the standard CQL query «[pos=’n‘&pos=’v‘]» did not yield any results, possibly due to the peculiarities of tagging the corpus with TreeTagger markup, a modified version of Penn Treebank. Then we created two lists of lemmas using Wordlist, a list of all corpus nouns and all verbs, exported them to XLSX documents, merged the lists, and used the Excel formula «=IF(COUNTIF(B:B;A599)>0;"Є у B";"Немає у B")» to compare them. Among the found matches, we selected potentially non-standard ones, checked the context in Concordance using the CQL formulas «[word=’selected word‘&pos=’n‘]», «[word=’selected word‘&pos=’v‘]» and confirmed or rejected their uniqueness. As a result, the found lexemes did not display any characteristics of new words; we identified isolated cases of non-typical use of nouns in the function of verbs («evidence», «document») and verbs in the function of nouns («send»), which, although they cannot be considered new words, clearly characterise the features of the vocabulary typical of this genre. Another notable fact is that some words were identified by Wordlist as nouns or verbs, but were not identified like that by the corresponding CQL formula. For example, in Wordlist, «reference» and «open» occurred in both the noun and verb lists, but the CQL formula «[word=’reference‘&pos=’n‘]» and

«[word=’open‘&pos=’n‘]» did not recognise them as ones. This confirms our assumption about the different approaches to the principles of tagging and functioning of the Wordlist and CQL tools. To summarise, we can say that the issue of effective search for new words using corpus queries remains open and obviously requires new non-standard approaches [10; 11].

Conclusions. Let us outline the potential of the corpus-based core professional vocabulary selection for the research purposes and in translation teaching using Sketch Engine software [16]. The frequency glossary of the most commonly used lexemes in the legal support texts for IT products, which we created using the frequency list of all lemmas in the corpus with the help of Sketch Engine Wordlist option, can serve as a resource for further linguistic analysis of these texts and can be used to develop a translation glossary for practical translation training. Frequency analysis of words by parts of speech will make it easier to choose words for the glossary, break it up into blocks according to parts of speech, and use that information to make more accurate translation of the texts. Since the core vocabulary selected represents distinctive aspects of the professional texts and is frequently linked to translator issues, analysis of multi-component noun phrases and hyphenated word combinations can aid in the creation of distinct blocks of lexical minimum for IT legal support texts. The most common prepositional phrases in the corpus texts are a distinct layer of vocabulary that is crucial for building a bilingual glossary of professional specialised texts, and as the linguistic units can be challenging to learn and to translate, the latter might be helpful for the training purposes. Finding words that are specific and unique to the texts of the specialised professional texts and compiling a list of words and combinations that are typical of the genre in general are two possible tasks of creating a lexical minimum that can also be accomplished by using Sketch Engine Keywords to search for typical lexemes and terminological combinations as well as neologisms and rare words.

The near-term prospect of the study is the corpus research of similar specialized professional texts of other areas with the view to selecting their core vocabulary of frequent terms and terminological combinations as well as prepositional phrases and neologisms. The latter can be obtained not only from the monolingual but also from the parallel or multi-lingual corpora, which considerably expands the possibilities of further linguistic research, and just as importantly, enhances the efficiency of specialised field translation training.

BIBLIOGRAPHY:

1. Bucur Ana-Maria, Dincă Andreea, Chitez Madalina, Rogobete Roxana. Automatic Extraction of the Romanian Academic Word List: Data and Methods. Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing. 2023. Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria. pp. 234–241.

2. 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

3. 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.

4. 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.

5. Culpeper J., Demmen J. Keywords. In: Biber D., Reppen R. (Eds.). The Cambridge Handbook of English Corpus Linguistics. Cambridge University Press. 2015. pp. 90–105. DOI: 10.1017/CBO9781139764377.006

6. Moreno-Ortiz, A. Making Sense of Large Social Media Corpora. An Open Accesss Publication. Palgrave Macmillan. 2024. 192 p. DOI: 10.1007/978-3-031-52719-7

7. 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.

8. 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

9. 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

10. Hengchen, S., Tahmasebi, N., Schlechtweg, D., & Dubossarsky, H. Challenges for Computational Lexical Semantic Change. In N. Tahmasebi, L. Borin, A. Jatowt, Y. Xu, & S. Hengchen (Eds.), Computational Approaches to Semantic Change. Language Science Press. 2021. pp. 341–372. DOI: 10.5281/zenodo.5040322

11. Tahmasebi N., Borin L., Jatowt A., Xu Y., Hengchen S. (Eds.). Computational Approaches to Semantic Change. Language Science Press. 2021. DOI: 10.5281/zenodo.5040302.

12. Afentoulidou V., Christofidou A. It's a Long Way to a Dictionary: Towards a Corpus-Based Dictionary of Neologisms. EURALEX Proceedings. 2021. Vol. 2. pp. 597–606.

13. 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.

14. 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.

15. 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.

16. 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