UDC 81.33
DOI https://doi.org/10.32782/2522-4077-2025-212-4

COMPILATION OF A CORE VOCABULARY FOR SPECIALISED PROFESSIONAL TEXTS USING THE SKETCH ENGINE SOFTWARE FUNCTIONALITIES

СТВОРЕННЯ ЛЕКСИЧНОГО МІНІМУМУМУ ВУЗЬКОСПЕЦІАЛІЗОВАНИХ ФАХОВИХ ТЕКСТІВ З ВИКОРИСТАННЯМ МОЖЛИВОСТЕЙ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ SKETCH ENGINE

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

One of the most powerful tools of applied linguistics is corpora, which are created and used in various fields of human activity. Automation of the process of selecting, compiling and analysing text corpora of virtually unlimited size provides new opportunities not only for researchers in the realm of philology, but also for experts who use the data as the basis for successful completion of practical tasks. Thus, corpus-based research has great potential for improving the effectiveness of language teaching, and translation in particular, as it allows for a more accurate and efficient selection of linguistic material in a particular highly specialised field, which is necessary for the future translators to successfully master the lexical minimum of professional texts, learn the peculiarities of functioning and translation of such commonly used units, and analyse existing and new linguistic trends in a particular field. Sketch Engine, which is one of the most famous and renowned software products for compiling and managing corpora, is the best suited to the tasks that arise when working with professional texts for translation training, as it allows not only analysing the corpora available on the platform, but also creating your own, including multilingual ones, for the purpose of quick and qualitative analysis of industry-specific texts, selection of active vocabulary and significant terminology and typical collocations, analysis of translation peculiarities and difficulties in rendering certain units, creation of glossaries and exercises to develop and improve the translation skills of future translators. A thorough analysis of all the functions of the Sketch Engine corpus manager can significantly increase the efficiency of methodological work with professional texts, and the possibility to create search queries in CQL can improve the accuracy of the linguistic results obtained. The proposed study describes the main capacities and methods of searching, analysing and selecting typical lexical material from professional texts based on the example of a corpus of English-language texts of legal support for IT products, namely the texts of licence agreements and contracts.

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

Одним з найпотужніших інструментів прикладної лінгвістики є корпуси, що створюються та використовуються у різних галузях людської діяльності. Автоматизація процесу підбору, укладання та аналізу текстових масивів практично необмеженого обсягу надає нові можливості не лише дослідникам-філологам, а й фахівцям, для яких такі данні є основою успішного виконання практичних завдань. Так, корпусні дослідження мають великий потенціал для підвищення ефективності навчання мовам, і зокрема перекладу, оскільки дозволяють більш точно та ефективно добирати лінгвістичний матеріал певної вузькоспеціалізованої галузі, необхідний для успішного опанування майбутніми перекладачами лексичного мінімуму фахових текстів, засвоєння особливостей функціонування та перекладу таких найуживаніших одиниць, а також аналізу існуючих та нових лінгвістичних тенденцій певної сфери. Sketch Engine, що є одним з найвідоміших та найкращих програмних продуктів для укладання та роботи з корпусами, якнайкраще відповідає завданням, які постають під час роботи з фаховими текстами для навчання перекладу, оскільки дозволяє не лише аналізувати наявні на платформі корпуси, а і створювати власні, у тому числі і багатомовні, з метою швидкого та якісного аналізу галузевих текстів, відбору активної лексики та значущої термінології та типових колокацій, аналізу перекладацьких особливостей та труднощів передачі певних одиниць, створенню глосаріїв та вправ для відпрацювання та удосконалення навичок перекладу фахових текстів студентів-майбутніх перекладa-

чів. Ретельний аналіз усіх функцій корпусного менеджера Sketch Engine дозволяє суттєво підвищити ефективність методологічної роботи з фаховими текстами, а можливості створення пошукових запитів мовою CQL – підвищити точність отриманих лінгвістичних результатів. Пропоноване дослідження описує основні можливості та методи відшукання, аналізу та відбору типового лексичного матеріалу з фахових текстів на прикладі корпусу англомовних текстів юридичного супроводу IT продуктів, а саме текстів ліцензійних угод та договорів.

Ключові слова: прикладна лінгвістика, корпус, корпусні дослідження, фаховий текст, лексичний мінімум, Sketch Engine, універсальна мова запитів CQL.

Problem under consideration. Corpus linguistics, emphasizing the practical application of linguistic analysis via language corpora and techniques to resolve both routine and academic language issues, has consistently exhibited a diverse array of directions: encompassing both traditional areas (text analysis, lexicography, language description, corpus-based translation studies, terminology development, etc.) and contemporary pursuits (creation of annotated corpora for natural language processing, machine translation systems, and linguistic modelling for artificial intelligence). Notably, the new approaches have demonstrated significant utility in language instruction and translation research, providing corpora-supported insights into genuine language usage, trends, and contextual changes. Corpus-based techniques facilitate field translation learning process by effective selection of the core vocabulary together with most typical collocations as well as of their translation counterparts; they allow for the faster vocabulary acquisition and field translation competence and considerably increase the accuracy and contextual fidelity of language transfer. Thus, understanding the mechanism of field vocabulary selection and utilizing appears to be one of the topical issues in the recent translation teaching methodologies.

The latest research analysis. Corpus linguistics issues have been deeply studied by a great number of linguists, though from slightly different angles, taking into account the vastness of the field. Here we can mention scholars researching general NLP and corpus linguistics questions, like A.-M. Bucur, M. Chitez, A. Dincă and R. Rogobete, who combine methods of corpus and computational linguistics to ensure effective vocabulary extraction [1], as well as B. Byrne, T. Domhan, S. E. Hasler, F. Hieber, K. Tran, S. Trenous, engaged in linguistic integrated models development in neural machine translation [2]. Norbert Schmitt, who is a leading scholar in applied linguistics in the realm of language acquisition, effectively combines vocabulary frequency methods and corpus-based techniques to enhance language teaching process [3]. Another prominent researcher, Stefan Th. Gries, a quantitative corpus linguist who applies statistical methodologies for language analysis, improving the accuracy of vocabulary selection across diverse domains [4]. Abdurrahman Kilimci and Aslı Nur Akkoyunlu make an emphasis on the improvement of the language acquisition process (namely, English as a foreign language) via the use of data-driven learning (DDL) approaches in translation courses. Thus, they conducted the experiment which showed considerable progress of learners' collocational knowledge against the background of general positive perception of the DDL approaches by the students [5]. Amel Lusta, Özcan Demirel and Behbood Mohammadzadeh are the researchers to successfully combine the DDL approach and corpus linguistics for increased performance in both language teaching and learning [6]. Home scholars have also contributed to the corpus studies with a view to applying various corpora techniques into the process of language and translation learning. Among the renowned names are T. Anokhina [7], V. Babych [8], I. Kobyakova [7], N. Lemish [9; 10], S. Matvieieva [8; 10], S. Schvachko [7], A. Zernetska [8; 10] and many more. The field is highly diverse and dynamic, with new research developments emerging year by year, and this tendency is expected to intensify in the future.

The article is aimed at analysing the corpora potential as well as the corpora software functionality to facilitate the translation teaching process. In order to fulfil the goal the research means to show the ways to select the most commonly used lexical items and collocations with them, as well as constructions that pose certain difficulties in translation, using the corpus created in Sketch Engine. The

latter could become an indispensable tool to create a glossary for effective vocabulary acquisition and analysis of the peculiarities of translation of specialised professional texts by translation students. The practical value of this study lies in the possibility to directly use the created minimum for teaching scientific and technical translation with students of the speciality «Translation» (English-Ukrainian), and the prospects for development of the glossary's features through further corpus analysis of parallel bilingual texts of legal support for IT products.

The main body of the article. Compiling a core vocabulary with the help of corpora managing software involves several consecutive stages and the order and accuracy of their implementation determines the validity and efficiency of the linguistic data obtained. The initial stage of the process included selection of the core texts for the future corpus, adding them to a newly-created bulk and analysing the resulting lexical material: 30 documents of legal support for software products were selected for the corpus, including end-user licence agreements (EULAs) in English, i.e. texts of the official business style of the genre of legal contracts. The texts were uploaded to Sketch Engine corpus management and text analysis software in DOCX (23) and PDF (7) formats. On uploading the texts to Sketch Engine and creating the corpus «Legal Documents for IT Products», its output data is available in the Corpus Info tab, where you can see that all texts have been recognised and added to the corpus (despite the fact that the formats differ). In the COUNTS block, we can see the total number of tokens (156,461), words (133,016), sentences (4,101), and documents (30). No less informative is the LEXICON SIZES block, which contains information about the number of unique units in the corpus, for example, the number of unique (i.e., occurring at least once) words, tags, or lemmas – 8,013, 61, and 4,567, respectively. In the TEXT TYPES block, there are two important icons that allow you to see useful statistics on document types (in fact, the share of one document in the corpus) and on the number of tokens in each document (Token coverage). The data is presented both in the form of a table and a pie chart (the latter is unfortunately limited in terms of the number of representations – a maximum of 20 documents).

The second stage included sorting all the lexemes by frequency. To do this, we chose the created corpus «Legal Documents for IT Products», then selected the Wordlist tool, and by default, the words were sorted by frequency. Predictably, the most frequent words on the list were articles, prepositions, conjunctions, and auxiliary verbs. Another problem was the large number of units to analyse (6,031 words and 4,185 lemmas). For this reason, we started changing the Frequency min parameter from 0 to 999. We settled on a Frequency min of 10 (a lower frequency critically increased the number of words, while a higher frequency increased the presence of service words at the top of the list), and got 1,335 (we did not include «nonwords» to reduce noise) lemmas, from which we manually selected 500 units of the most frequent vocabulary, among which nouns, verbs, participles, and adjectives predominated. The DOCF (Document Frequency) parameter (i.e., how many different documents contained a particular word) was also useful in compiling the list, as the higher the DOCF was the reason for selecting words. The next step was to select the vocabulary by parts of speech, including nouns, adjectives, verbs and participles with subsequent sorting them by frequency. To do this, we sorted the words in Wordlist using variable criteria («noun» (2,518 lemmas), «adjective» (863 lemmas), «verb» (810 lemmas). At the same time, we ran queries in CQL in Sketch Engine:

– [tag='N.*'] – to search for nouns (45,672 in total, 2,818 lemmas);

– [tag='JJ.*'] – for adjectives (9,706 in total, 866 lemmas);

– [tag='V.*'] – for verbs (18,847 in total, 812 lemmas);

– [word='.*ed'&tag='VVN.*'] – to search for -ed participles (3,404 in total, 390 lemmas);

– [word='.*ing'& tag='VVG.*'] – to search for participles ending in -ing (2,124 in total, 305 lemmas).

The selection procedure was as follows:

1). entering the search query using CQL;

2). sorting the results by lemma frequency;

3). Exporting them to an XLSX file.

It is noteworthy that the number of lemmas belonging to one and the same part of speech differed when using the Wordlist tools and the CQL queries. This slight discrepancy may be related to the counting methods applied to the corpus and the very principle of operation of the tools.

Whereas the results for the most frequently used nouns, verbs and adjectives could be used immediately for manual selection of the core vocabulary, this was not possible with the -ed and -ing verbal forms due to the overlap in tense and participle forms. In order to process the verbal forms, we had to analyse them in a broader context using CQL formulas:

– [tag='N.*'][word='.*ed'& tag='VVN.*'] – to search for the combination of a noun + an -ed participle, which not only effectively detects 'real' participles, but also selects material for frequency collocations typical of the field (391 collocations);

– [word='.*ed'&tag='VVN.*'][tag='N.*'] – to search for collocations of past participle + a determined noun (143 collocations);

– [word='.*ing'&tag='VVG.*'][tag='N.*'] – to search for collocations of the participle ending in -ing + a noun (285 collocations).

The resulting collocations were sorted by frequency and exported to XLSX files for further manual processing.

Another phenomenon essential for professional English vocabulary, in particular terminology, is noun clusters, and we also searched for them with the help of standard Concordance tools using CQL formulas:

– [tag='N.*'][tag='N.*'] or a more compact regular expression quantifier [tag='N.*']{2} – to search for two-component noun combinations (3,403 units);

– [tag='N.*'][tag='N.*'][tag='N.*'] or [tag='N.*']{3} – to search for three-part noun phrases (1,038 units);

– [tag='N.*'][tag='N.*'][tag='N.*'][tag='N.*'] or [tag='N.*']{4} – to search for four-component noun phrases (325 units);

– [tag='N.*'][tag='N.*'][tag='N.*'][tag='N.*'][tag='N.*'][tag='N.*'] or [tag='N.*']{5} – to search for five-component noun phrases (119 units);

– [tag='N.*'][tag='N.*'][tag='N.*'][tag='N.*'][tag='N.*'][tag='N.*'][tag='N.*'] or [tag='N.*']{6} – to search for six-component noun phrases (61 units);

– [tag='N.*'][tag='N.*'][tag='N.*'][tag='N.*'][tag='N.*'][tag='N.*'][tag='N.*'][tag='N.*'] or [tag='N.*']{7} – to search for seven-component noun phrases (31 units);

– [tag='N.*'][tag='N.*'][tag='N.*'][tag='N.*'][tag='N.*'][tag='N.*'][tag='N.*'][tag='N.*'][tag='N.*'] or [tag='N.*']{8}– to search for eight-component noun phrases (18 units);

– [tag='N.*'][tag='N.*'][tag='N.*'][tag='N.*'][tag='N.*'][tag='N.*'][tag='N.*'][tag='N.*'][tag='N.*'][tag='N.*'] or [tag='N.*']{9}– to search for nine-component noun phrases (12 units).

Further analysis of the data showed that two-component, three-component and four-component combinations were among most frequent compound nominal terms, while the vast majority of five-, six-, seven-, eight- and nine-component clusters mostly contained the names of the manufacturer or licensed product and thus couldn't be considered terms. In addition, due to the peculiarities of the formatting of legal documentation, among the noun clusters with a large number of elements, there were often repetitions of words in the end of the phrase, such as «eco Utility End User License Agreement TOSHIBA eco», and among nine-component clusters, there was only one comprehensive phrase: «TSG Interactive Gaming Europe Limited End User License Agreement», the rest repeated words in a line.

In the process of analysis of noun phrases with two, three and four components, we noticed that proper names are often found among them, so we decided to try to sort them out using the CQL formula [tag='N.*'&word='[A-Z].*']{2,4}. Unfortunately, because of the peculiarities of formatting of licence agreements texts, a large number of phrases though not proper names were also capitalised,

so the final selection of proper names had to be done manually. Hyphenated phrases were worth the attention too, as they are typical of the licence agreements language; therefore, we tried to find all the phrases of the kind in the texts using the CQL formula [word='.*-.*-.*'] (26 units). Among the notable word combinations were, for instance, «non-software-based (systems)», «TEXT-TO-SPEECH (SOFTWARE)», «Not-for-resale (SOFTWARE)», «not-for-profit (entity)», «out-of-court (resolutions)», etc.

Another important aspect of specialized professional vocabulary are prepositional phrases. Particularly noteworthy are the English combinations «verb + preposition» and «verb + adverb», which often build up phrasal verbs. The search was performed using the CQL formulas:

– [tag='V.*'][tag='IN.*'] – to search for verb-prepositional phrases (1,155 units);

– [tag='V.*'][tag='RB.*'] – to search for verb and adverb phrases (351 units);

– [tag='V.*'][tag='RP.*'] – to search for verb plus participle (according to the Sketch Engine mark-up language) (58 units).

As a result of processing the data, it turned out that the most productive formula for finding «verb-preposition» combinations was [tag='V.*'][tag='IN.*'] (for identifying verbs and prepositions). Despite the large number of units obtained (1,155), it was not difficult to select the most frequent combinations for the core vocabulary, since the number of combinations with a frequency of 3 or more was 347 units, and of 2 – only 546. Obviously, there was some information noise, such as verb combinations with pronouns («recommended that») or with conjunctions («replaced if»), but the overall sample was quite accurate. Verb combinations with prepositions like «on», «off», «upon», «out», «around», «through», «up» and «down» were rather illogically marked as particles in the corpus, so including «particles» in the search was also useful for our core vocabulary selection. The search query for verb-adverb combinations was the least productive, as the results included, for example, the particle «not», frequent adverbs, such as «still», «only», «also» and others that do not form stable expressions with verbs. Importantly, verb combinations with prepositions ending in -ly constitute a separate group of collocations typical of the field. Therefore, we have chosen a different CQL formula to separate them:

– [tag='V.*'][word='.*ly' & tag='RB'] – to search for collocations of a verb + an adverb ending in -ly.

The search results revealed that a significant number of collocations were made up of the combination of the verb «to be» and the adverb ending in -ly, but due to grammatical restrictions of the English language, the adverb is often automatically bound to the following word, usually an adjective or participle, so we had to add a restriction and exclude all forms of the verb «to be»:

– [tag='V.*'& lemma!='be'][word='.*ly' & tag='RB'] – to search for collocations of the verb except for «to be» and an adverb ending in -ly (100 units).

The results showed some interesting collocations, for example, «(the law shall) apply exclusively» (23 occurrences) or «(The Licence shall) terminate automatically» (7 occurrences), etc.

Also, the search for verb-preposition combinations revealed a large number of participial forms in combination with a preposition, so we performed a separate search for such phrases using the CQL formula:

– [word='.*ed'&tag='VVN.*'][tag='IN.*'] – to search for participles ending in -ed + an adverb (491 units).

Although the majority of the identified combinations are part of the passive voice verb form, they can be selected to the core vocabulary as an independent group of collocations typical of the given professional texts.

Conclusions and further research prospects. Let us finally summarize the steps taken to compile the corpus-based core vocabulary in the texts of the EULAs. As a result of the queries, we received data from the files uploaded to a separate corpus in Sketch Engine in the form of concordance tables and frequency tables for further verification and analysis. As an additional method, we used Manual Data Validation to check the accuracy, completeness and consistency of the data manually and

spreadsheet-based data analysis (tabular data analysis) [11]. The ultimate goal of our research was to create a frequency glossary and lexical minimum based on the corpus of texts of legal support for IT products, which would be intended for students of the Translation speciality. Given that the obtained vocabulary is potentially multifunctional, we have developed the following sequence of data processing and interpretation: out of the frequency list of all lemmas in the corpus obtained using Wordlist, we created a frequency glossary of the most often used lexemes in the texts of legal support for IT products. The latter can be used to create a translation glossary as well as a reference material for working with the documents. Frequency analysis of words by part of speech will help to select lexemes for the glossary more efficiently, divide it into blocks by part of speech, and use it to compile its translation version [1]. Analysis of multi-component noun phrases and hyphenated word combinations selected by CQL queries will help to create separate blocks of lexical minimum for IT legal support texts, as they represent unique features of these texts and often cause difficulties in translation. The search for the most frequent prepositional phrases in the corpus texts is a separate layer of vocabulary, which, firstly, can be difficult to learn and therefore requires special attention, and secondly, is an important bulk of material for creating a bilingual glossary of such phrases [12, c. 19]. The search for typical lexemes and terminological combinations using Keywords will help to solve two potential tasks of creating a lexical minimum, that is, to create a list of words and combinations typical for the genre of contract language and agreements in general, and to identify words that are specific and unique to the texts of IT product licence agreements.

The immediate prospect of the study is the corpus research of IT legal support texts using other corpus tools, such as Korpusomat, to perform a comparative analysis of their functionality and identify potentially new text processing capabilities, as well as the analysis of bilingual corpora of texts of this genre to expand the possibilities of using them for translation training [13, c. 433-435; 14, c. 807-808; 15, c. 30-32]. Further potential directions for the development of the study include the creation of similar corpora to work with professional texts in specialised fields for research and teaching purposes.

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. Schmitt N. Vocabulary in Language Teaching (2nd ed.). Cambridge University Press. 2020. 304 pages.

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

5. 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.

6. 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.

7. 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.

8. 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.

9. Леміш Н. Є. Англо-український паралельний корпус текстів для студентів спеціальності «Переклад». Актуальні проблеми романо-германської філології та прикладної лінгвістики. 2018. Чернівці. Вип. 1 (15). С. 207–210.

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. 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.

12. Gamallo Otero P. Evaluating Two Different Methods for the Task of Extracting Bilingual Lexicons from Comparable Corpora. Proceedings of the LREC 2008 Workshop on Comparable Corpora. Marrakech, Morocco, 2008. pp. 19–26.

13. Baños R., Borja A. The Application of a Parallel Corpus English-Spanish to the Teaching of Translation (ENTRAD Project). New Trends in Translation and Cultural Identity / Ed. Muñoz-Calvo M., Buesa-Gómez C., Ruiz Moneva M. A. Cambridge Scholars Publishing, 2008. pp. 433–444.

14. Kübler N., Mestivier A., Pecman M. Teaching Specialised Translation Through Corpus Linguistics: Translation Quality Assessment and Methodology Evaluation and Enhancement by Experimental Approach. Meta, 2018. Vol. 63. No. 3. pp. 807–825.

15. Saralegi X., San Vicente I., Gurrutxaga A. Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain. Proceedings of the Workshop on Building and Using Comparable Corpora, 6th International Conference on Language Resources and Evaluation (LREC). 2008. pp. 27–32.