European Master in Lexicography

Universidade de Santiago de Compostela

Faculty of Philology

“Methodology for the Corpus-based English-German-Ukrainian Dictionary of Collocations”

Master’s Thesis

Student: Mariia Polova

Supervisor: Dr. Marcos Garcia González

July 2022

Acknowledgments

I would like to express my gratitude to my supervisor Dr. Marcos Garcia González for his unwavering guidance and willingness to share his extensive professional expertise. Although I was not particularly well-versed from the computational and technical perspective before embarking on this Master's project, the support from Dr. Garcia González helped me broaden my knowledge and succeed in applying the computationally demanding methodology onto my sample dictionary data.

This work would also not have been possible without the constant guidance from the EMLex coordinator from the University of Santiago de Compostela, Prof. Dr. María José Domínguez Vázquez who ensured that we were on track during these two years.

I would also like to thank all the EMlex professors, who shared not only their knowledge during lectures and seminars but also their passion for lexicography.

Last but not least, I would like to thank the EMLex consortium and organization committee for making this degree extraordinary with the two mobility semesters at Friedrich-Alexander-Universität Erlangen-Nürnberg and Roma Tre University, and numerous lexicographic conferences and events.

Abstract

This Master's thesis recounts the vision of the multilingual collocations dictionary project for the English, German, and Ukrainian languages (*"Corpus-based English-German-Ukrainian Dictionary of Collocations"* or *EDU-Col*) and elaborates on the methodology for compiling the dictionary and its key dictionary structures. The dictionary will cater to the needs of language learners, translators, text producers (journalists, copywriters), and native speakers. Tapping into the latest developments in NLP and the capabilities of corpora, the methodology for creating the proposed dictionary relies on the automatic extraction of dictionary information types, namely collocation candidates, example sentences, and translation equivalents for collocations. The automatic extraction is followed by manual validation in order to maintain the quality of the obtained lexicographic data.

Key words: *collocations; multilingual dictionary project; automatic generation of dictionary data*

Abstract

Diese Masterarbeit befasst sich mit der Konzeption des mehrsprachigen Kollokationswörterbuchs für die englische, deutsche und ukrainische Sprache ("Corpus-based English-German-Ukrainian Dictionary of Collocations" oder EDU-Col) und erläutert die Methodik für die Erstellung des Wörterbuchs und seine wichtigsten Wörterbuchstrukturen. Das Wörterbuch ist auf die Bedürfnisse von Sprachlernern, Übersetzern, Redakteuren (Journalisten, Werbetextern) und Muttersprachler ausgerichtet. Die Methodik zur Erstellung des vorgeschlagenen Wörterbuchs basiert auf der automatischen Extraktion von Wörterbuchinformationen, nämlich Kollokationskandidaten, Beispielsätzen und Übersetzungsäquivalenten für Kollokationen. Auf die automatische Extraktion folgt eine manuelle Überprüfung, um die Qualität der erhaltenen lexikografischen Daten zu gewährleisten.

Schlüsselwörter: *Kollokationen; mehrsprachiges Wörterbuchprojekt; automatische Generierung von Wörterbuchdatei*

Table of Contents

Introduction

As a learner of foreign languages, namely English and German, I often faced challenges with finding a proper word combination to sound natural while expressing a thought or idea in writing or speech. The same occasionally happened when I was working on translation projects that involved English and my native language of Ukrainian. Even at advanced levels of language proficiency, it is not uncommon for learners to encounter difficulties with collocations (see Altenberg & Granger, 2001; Gilquin, 2007), which are often pointed out by the editors, who are usually native speakers, as the first thing that draws their attention while proofreading translations.

1.1 Motivation

As far as the English, German, and Ukrainian languages are concerned, there is a clear gap in collocations dictionaries pertaining to bilingual and multilingual lexicography, especially for such language pairs as Ukrainian and English and Ukrainian and German, nor is there a separate English-German collocations dictionary known to us to date. Besides, the demand even for monolingual Ukrainian collocation dictionaries can be seen among the Ukrainian readership, with the questions if such dictionaries exist for the Ukrainian language appearing on Facebook groups for translators and language enthusiasts quite often. For instance, in an inquiry on the Facebook group Ukrainian translation[1], a user[2] asked if there exist Ukrainian resources similar to The English Collocations Dictionary online[3]. Among the pieces of advice from users that can be found in the replies to the post, there is a recommendation to look for collocations of a particular word using corpora. Such an approach can undoubtedly be an efficient solution but only if a user is familiar with corpora and corpus tools. In other cases, when the users are laymen, the most efficient resource would unquestionably be a dictionary of collocations. As can be seen from the account of the Ukrainian monolingual collocations dictionary (see Section 2.3), no comprehensive resource, which is accessible to the wider public, is available. Furthermore, among the few existing collocations dictionaries, which represent the Ukrainian monolingual lexicography, there are no corpus-based dictionaries.

[1] Facebook group for translators “Ukrpereklad”. Retrieved February 20, 2022, from https://www.facebook.com/groups/ukrpereklad/

[2] Users discussion in the facebook group for translators. Retrieved February 20, 2022, from https://www.facebook.com/groups/ukrpereklad/permalink/3687342137968194/

[3] https://ozdic.com/

In an attempt to fill these gaps, a multilingual project that combines all three languages and the methodology based on natural language processing (NLP) for compiling a dictionary with collocations as its subject matter would be a challenging yet promising endeavor.

1.2 Aim

In line with the outlined vision of the multilingual collocations dictionary project, this Master's thesis undertaken within the frames of the EMLex Master's thesis component aims at elaborating on the methodology for compiling the *"Corpus-based English-German-Ukrainian Dictionary of Collocations"* (hereinafter, the *EDU-Col*, the naming referring to the first letters of languages, namely English-Deutsch-Ukrainian) and its features from a lexicographic point of view, and presenting the results of the first attempts to gather collocation candidates and other key dictionary information types using the automatic approaches.

1.3 Objectives

In order to fulfill the overarching aim set for the master thesis, the following objectives are defined:

1) to provide an account and identify peculiar features of collocations dictionaries in English, German, and Ukrainian lexicographic traditions;
2) to create a dictionary plan elaborating on general characteristics of EDU-Col, including dictionary typology, target user, etc;
3) to investigate the approaches to the definition and treatment of collocations in linguistics and lexicography;
4) to establish a methodology, which will be used for the dictionary compilation;
5) to create corpora and obtain sample results, documenting the steps and procedures of dictionary information types generation, including collocation candidates, examples, and translation equivalents;
6) to conceptualize the structure of the dictionary, i.e., macro-, micro-, medio-, and access structure, as well as outer texts;
7) to present model dictionary articles[4] for different types of collocations.

[4] Here, we follow the terminology as suggested by Gouws (2014), i.e., where the term dictionary article is preferred, as opposed to dictionary entry, which is treated as an individual part of dictionary articles and other texts in dictionaries.

1.4 Methodology

The methods applied in this Master's thesis project include a combination of methods pertaining to lexicography (Schierholz, 2015), i.e., methods used in practical lexicography and metalexicographic methods related to the phase of preparation, acquisition of the material and the data, treatment, and evaluation of the material and the data. Most importantly, outlining a dictionary conception and identification of the dictionary type and the dictionary functions, was performed, resulting in the metalexicographic description of the dictionary articles. Linguistics and corpus linguistics methods were used to obtain corpora, specifically the method of compiling and subsequently natural language processing methods for corpora preparation, including POS-tagging, lemmatization, syntactic parsing. In addition, concrete theory-related methods (Garcia et al., 2019a, 2019c; Orenha-Ottaiano et al., 2021) were employed for the automatic generation of dictionary information types.

1.5 Thesis organization

The thesis contains seven chapters and is organized as follows. The introductory chapter presents the aim and objectives of the work, as well as elaborates on the motivation for undertaking the project. Chapter 2 gives an account of existing collocation dictionaries across English, German, and Ukrainian. Chapter 3 describes the general characteristics of EDU-Col including the type of the dictionary, its potential users, and its functions. Then, the approaches to the treatment of collocations are delineated in Chapter 4, and the view on collocations adopted for EDU-Col is presented. Chapter 5 presents the methodology adopted for extracting collocation candidates, example sentences, and translation candidates of collocations that will be eventually used in the dictionary articles of EDU-Col. Some thoughts on the dictionary structure of EDU-Col, among which macrostructure and lemma selection, and microstructure are explored in Chapter 6. Finally, Chapter 7 contains the results and conclusions of our study, as well as the ideas for further work.

2 Collocation dictionaries in lexicography

Before dwelling on the methodological approach selected for the dictionary compilation as well as the characteristics of the dictionary structures, let us observe how EDU-Col will fit into the wider lexicographic practice context. The next sections will provide a brief account of the existing collocations dictionaries across lexicographic traditions of English, German and Ukrainian and review the multilingual collocations projects available as online resources.

2.1 Anglophone lexicography

The English language undoubtedly boasts the largest number of collocations dictionaries, with big names such as Oxford, Longman, and Macmillan among their publishers offering both print and online versions. We will focus mainly on online versions of the above-mentioned dictionaries since they are more accessible nowadays to the users than the print ones.

In particular, the *Oxford Collocations Dictionary*'s online version, available for purchase at the Oxford learners dictionaries website[5], gives an insight into 250,000 word combinations and 9,000 nouns, verbs, and adjective collocations. The dictionary is based on the three billion Oxford English Corpus, and its dictionary articles are illustrated with 75,000 examples and various usage notes. Similarly, *Longman Collocations Dictionary,* available through subscription on the publishing house's website, offers over 70,000 collocations and a substantial number of examples. The *Macmillan Collocations Dictionary*[6] focuses on English learners' productive needs, providing them with collocations grouped in semantic sets and illustrated with examples in context, whereas a special emphasis is placed on collocations frequent in academic and professional writing. The dictionary is based on a corpus of almost 2 billion words, featuring over 4,500 keywords.

Another resource worth mentioning is the Online Collocation Dictionary[7], which is reportedly an older version of the Oxford Collocations dictionary available freely, and according to the website, provides users with over 150,000 collocations based on 100 million word British National Corpus. Collocations in the dictionary articles are grouped according to meaning and parts of speech, additionally containing prepositional collocations and common phrases, whereas the examples are provided occasionally. The same contents can be accessed through ozdic.com.

[5] About the Oxford Collocations Dictionary. Retrieved April 15, 2022, from https://www.oxfordlearnersdictionaries.com/definition/collocations/

[6] About the Macmillan Collocations Dictionary. Retrieved April 15, 2022, from https://www.macmillandictionary.com/collocations/about.html

[7] https://www.freecollocation.com/

As far as the print dictionaries are concerned, the two works can be named, namely *LTP Dictionary of Selected Collocations* (Hill & Lewis, 1997, 1998) and *the BBI Combinatory Dictionary of English* by Benson et al., (1997).

Apart from conventional dictionaries, users, especially those who are working on writing tasks, can utilize a variety of tools available on the web that generate collocations automatically from corpora, such as ProWritingAid[8], which offers an extensive collocations list created by analyzing books available on the web. Another resource, Flax[9], provides its users with an artificial intelligence-driven toolkit that allows searching across BNC, the British Academic Written English (BAWE) corpus, and the Wikipedia crowd-sourced corpus, viewing collocations on Flax's website, as well as saving and re-using them in game-based activities for playing and learning. A similar writing aid tool, Inspirassion's[10], functionality allows users to check which words can be combined together, organized by parts of speech.

2.2 German lexicography

Monolingual collocations dictionaries of German are not as widespread as the English ones, with two print dictionaries available on the market, one of which additionally offers an online version. In particular, *Wörterbuch der Kollokationen im Deutschen* (Quasthoff, 2011), which according to the preface is aimed at German native speakers is a corpus-based dictionary with over 3200 of the most common bases (nouns, verbs, adjectives). The articles in the dictionary are organized taking into account both parts of speech that collocates represent and the semantic criteria. Thus, noun base collocations are first followed by typical verbs which are used with the noun in the role of a subject in the nominative case, then typical verbs that go together with the noun in the dative or accusative cases. Verb collocates are followed by adjective collocates which are then organized according to the semantic criteria.

Another resource, available both in print and online versions, "*Feste Wortverbindungen des Deutschen. Kollokationenwörterbuch für den Alltag"* (Buhofer et al., 2014) is primarily aimed at language learners, hence, it covers collocations for a basic vocabulary list of 2000. In total, over 95,000 word combinations are presented and illustrated with over 30,000 example sentences. In addition, to account for the learners' needs, the articles are provided

[8] https://prowritingaid.com/Free-Online-Collocations-Dictionary.aspx
[9] http://flax.nzdl.org/greenstone3/flax?a=fp&sa=collAbout&c=collocations
[10] https://inspirassion.com/en/

with stylistic markers, which signal to a reader that the collocation, for instance, implies irony or is derogatory.

There have also been created bilingual dictionaries with German as one of the languages, such as "*Kollex. Deutsch-ungarisches Kollokationslexikon. Korpusbasiertes Wörterbuch der Kollokationen. Deutsch als Fremdsprache"* (Hollós, 2014); however, due to the extensiveness of the subject, we will not go into detail, thus limiting the account only to the languages our project is concerned with.

2.3 Ukrainian lexicography

The Ukrainian lexicography is not so rich in monolingual collocations dictionaries with few available resources, which are not known among the wider public. In particular, two online resources were revealed, which cover only the prepositional collocations and noun-verb and verb-noun collocations, and several print dictionaries, which are not accessible for purchase due to the low number of copies printed initially.

Online Ukrainian collocation dictionaries are represented by *the dictionary of Ukrainian prepositional collocations*[11], which was created as a result of comparing frequency lists of lemmata and word forms of collocations of subcorpus of legislative texts of the Ukrainian language corpus MOVA.info[12], with the subcorpus accounting for 2.7 million tokens. The dictionary includes more than 1415 collocations for 29 Ukrainian prepositions. The subcorpus was also used to create *the dictionary of Ukrainian predicative collocations*[13]*,* focusing on collocations with verbs as a base, which includes 440 collocations. These two resources are very important attempts to create collocation dictionaries of the Ukrainian language, but they cover only a fraction of all the collocations types and do not aim to cover other domains, as well as provide general language collocations.

As far as the print dictionaries are concerned, the *Dictionary of the compatibility of the words of the Ukrainian language* (Sakhno, 1999) can be mentioned. It contains the words of the general lexicon of the Ukrainian language, where collocations is only a small part, with the author including also idioms and proverbs. The resource is not available to the wider public –

[11] http://www.mova.info/Page.aspx?l1=66

[12] http://www.mova.info/carticle.aspx?l1=210&DID=5347

[13] http://www.mova.info/Page.aspx?l1=208

only a few copies are available in state scientific libraries. Other print resources that describe collocations are *the Dictionary of the epithets of the Ukrainian language* by Bybyk et al., (1998) and *the New Dictionary of the epithets of the Ukrainian language* by Yermolenko et al., (2012)*,* which focus on attributes that describe nouns, particularly characteristics of a person or thing mentioned, thus corresponding to the adjective-noun collocations type. However, the focus of these dictionaries is not on the common frequency-based attributes but rather on occasional attributes bearing stylistic expressiveness, which are used by novelists and poets in their works.

2021 saw a publication of the print *Ukrainian-English Collocation Dictionary* (Shevchuk, 2021), compiled by Yuri Shevchuk, the lecturer of Ukrainian language at Columbia University. The dictionary is primarily aimed at Ukrainian language learners including a comprehensive introduction to the Ukrainian language and grammar and covers idioms and proverbs in addition to collocations. The dictionary information types come not from corpora, as the work of the result of a survey and analysis of existing general language dictionaries. It should be noted that none of the above-mentioned print dictionaries were created relying on corpora.

2.4 Multilingual collocations dictionaries projects

Multilingual collocations dictionary projects are not generally widespread; however, a few attempts have been made to create multilingual platforms where users can find information about collocations in different languages. One of the most recent projects is PLATCOL, an “online Platform for Multilingual Collocations Dictionaries” (Orenha-Ottaiano et al., 2021), which covers collocations with verbal, adjectival, nominal, and adverbial bases. The languages covered are Portuguese, Spanish, English, French, and Chinese. The dictionary data is extracted with the help of automatic methods from corpora in combination with post-editing performed by lexicographers. The platform can be used multidirectionally, accounting for monolingual, bilingual, or multilingual layouts of the dictionar.

Part of the methodology and design of PLATCOL is based on another earlier online corpus-based Portuguese-English dictionary of collocations (Orenha-Ottaiano, 2017), which was eventually transformed into PLATCOL. The dictionary was bi-directional allowing users to view it either as a monolingual (Portuguese or English) or as a bilingual (English-Portuguese and Portuguese-English).

Another project that should be mentioned is the investigation of a methodology for the automatic construction of a multilingual dictionary of collocations from large corpora using distributional semantics (Garcia et al., 2019a) and a proposal for a multilingual online dictionary of collocations of English, Portuguese, and Spanish. The dictionary which was planned to be released as the tool aims to feature the following types of collocations: verb-object, adjective-noun, and nominal compounds, serving both as a monolingual and multilingual resource, with the system showing equivalents in other languages, ranked by a translation confidence value, for each collocation. The collocation candidates in Garcia et al., (2019a) are extracted using dependency parsing and statistical association measures, whereas the equivalents in target languages are obtained using compositional methods based on cross-lingual models of distributional semantics. Extraction of collocations is possible due to the capabilities of corpora, whereas the cross-lingual embeddings are mapped relying on unsupervised approaches. The methodology described in Garcia et al. (2019a) is used in the process of creating the above-mentioned multilingual collocations platform PLATCOL and will be also applied in the collocations dictionary project for English, German, and Ukrainian languages.

3. General characteristics of the dictionary

3.1 Dictionary typology

In line with the phenomenological typology, i.e., which answers the question “What does the dictionary look like?”, the structural and content-related characteristics of dictionaries should be considered as the classification criteria. Hausmann’s typology (1989, p. 977, see Figure 1) distinguishes between general dictionaries and specialized dictionaries, with monolingual and bilingual dictionaries belonging to both groups. Accordingly, *EDU-Col* falls within the second specialized type of dictionaries since it describes the syntagmatic properties of languages. As far as the lexical units and dictionary information types are concerned, it should be noted that collocations presented in *EDU-Col* will cover noun, adjective, and verb bases.

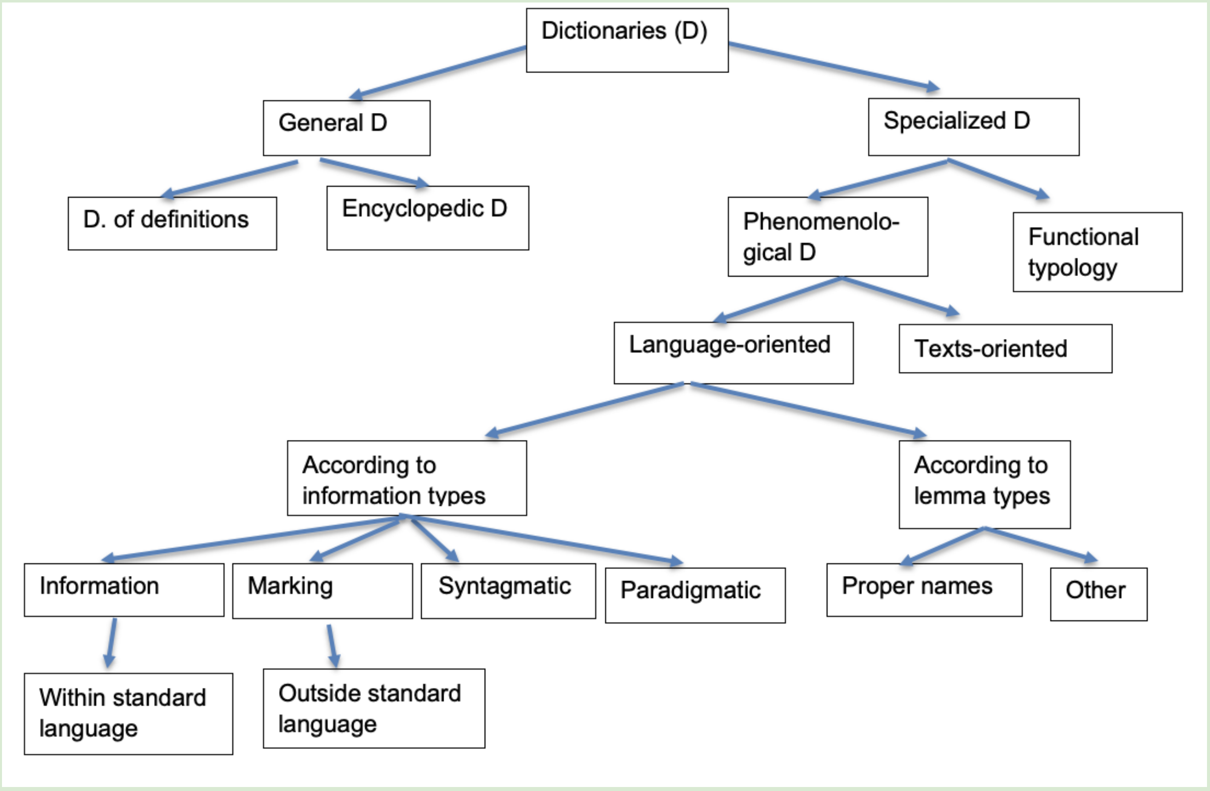


Figure 1. HSK Typology of Dictionaries (Hausmann, 1989, p. 977)[14]

Another aspect that should be taken into account in this line of thinking is the diasystematic criteria, i.e., what subsystems of the language the lexical units and information types represented in the dictionary are affiliated with. Thus, on the time axis, *EDU-Col* covers the synchronic stage of language development. In turn, the spatial characteristics cover only the standard language excluding dialects. For the English corpus, the varieties common in the UK and USA are covered. However, the texts in the German corpus are restricted to the variety used in Germany, not accounting for other German-speaking countries. Finally, the Ukrainian corpus is constituted by the texts representing the variety of Ukrainian which is primarily spoken in Ukraine.

The subject matter of the dictionary is the syntagmatic relations of lemmata, and the dictionary itself can be characterized as polyselective (in line with Wiegand et al., 2020, p. 206) because three classes of linguistic units, namely nouns, adjectives, and verbs are considered for collocation bases.

[14] Note: Translation from German is mine

Moreover, *EDU-Col* will cover not only collocations characteristic of general language but also field-specific ones, such as law, economics, and environment for which specialized corpora will be used to obtain the collocations. In terms of the medium, the dictionary will be compiled for online use striving for a user-friendly interface and a straightforward consultation procedure.

The positioning of *EDU-Col* within the functional typology, which answers the questions "Who is the addressee of the dictionary?" and "What is the purpose of the dictionary?" will be elaborated on in detail in the next sections.

3.2 Target users

Anderson & Fuertes Olivera (2009, p. 214) argue that the lexicographic process should start with delineating the target user group of the dictionary and the functions or situations in which it can be used. This step of dictionary-making is of particular importance since it prevents confusion in potential users, which might appear unless the step had been taken. Thus, the following types of users can be the potential addressees of *EDU-Col*:

- language learners
- translators
- text producers (journalists, copywriters)
- experts in particular fields who are learning a foreign language
- native speakers.

Collocations are often discussed in relation to language fluency. In the words of Laufer (2011, p. 29), "Native speakers of a language operate with a large number of collocations which contribute to idiomaticity and fluency of their expression while foreign learners do not seem to perceive collocations as chunks and often produce them by combining separate words that do not go together in a given language". Thus, acknowledging the challenges language learners face with collocations, *EDU-Col* aims to account for the needs of the learners of English, German, and Ukrainian who will be one of the target user groups of the planned dictionary.

Multiple languages in the dictionary will allow using it not only as a monolingual but also as a bilingual dictionary for a specific language pair chosen by the user or even set the settings to show translation equivalents in all available languages. Two languages can accommodate users with assistance in situations where a monolingual dictionary is of limited use. As

Adamska-Sałaciak (2010, p. 133) notes, while monolingual dictionaries are suitable primarily for advanced (or at least upper-intermediate) students, bilingual dictionaries can take care of learners at all levels – from beginners to advanced students. Hence, *EDU-Col* will cater to the needs of a much wider target audience, including beginners who are often overlooked by the publishers of monolingual dictionaries.

Laufer (2011, p. 31) notes that learners can comprehend the majority of collocations in receptive situations if at least one of the individual words that form a collocation is familiar to them. However, difficulties arise in production language use, such as translation or writing. Bahns and Eldaw's study (1993, as cited in Laufer, 2011, p. 31) revealed that in translation tasks, students make significantly more errors related to collocations than to single lexical items, with the number of errors being almost double. Therefore, another target user group of our dictionary project is students of translation and interpreting, as well as practicing translators.

Finally, native speakers can in turn benefit from using the dictionary to learn more about the intricacies of their own language, especially related to the collocations used in specific domains.

3.3 Dictionary functions

Having identified the potential target users of *EDU-Col*, the situations in which it can be used will be dwelled upon in this section. Dictionary usage situations are closely related to the notion of a lexicographic function, which in the words of Tarp (1998, as cited in Tarp, 2002, p. 610) is "the endeavour and ability of the dictionary to cover the complex of needs that arise in the user in a particular user situation." The scholar subdivides the functions of a dictionary into communication-orientated and knowledge-orientated.

The essential communication-orientated functions, according to Tarp (2002, p. 611), are:

(1) to assist the production and reception of texts in the native language;
(2) to assist the production and reception of texts in a foreign language;
(3) to assist the translation of texts between the native language and a foreign language.

In turn, the most important knowledge-orientated functions are:

(1) to give general or special cultural and encyclopaedic information;
(2) to give information about the language.

According to the Engelbeg and Lemnitzer's (2009, p. 20) typology of dictionaries according to possible uses, the dictionaries are viewed as either reference works ("Nachschlagerwerk") or reading books ("Lesesbuch"). The former presupposes selective dictionary consultations related to such reasons as language competence problems in text production situations, comprehension problems in text reception situations, equivalence problems in translation, as well as research interests (e.g. finding in which lexemes a particular suffix appears). In turn, dictionary as a reading book can be explained by the words of Wiegand (2000, p. 736):

> "Jemand, der in einem Wörterbuch liest, hat z.B. Interesse an der oder einer Sprache, am Bau der Sprache, an ihrer Geschichte, an der Geschichte bestimmter Wörter etc. Er lässt sich bei der Lektüre vom Wörterbuchtext führen, um zu entdecken, studiert im Wörterbuch und sucht Belehrung. Es kann sogar sein, dass er den besonderen Reiz lexikographischer Texte genießt, z.B. das nuancenreiche Beieinander der Wörter im Paradigma und ihre Variationen in Kollokationen".
>
> ["Those who read a dictionary take an interest, for instance, in a particular language, its structure, history, or history of certain words, etc. They are guided in their reading by the dictionary text, study and look for instructions. They may even enjoy the special appeal of lexicographic texts, e.g., the nuanced juxtaposition of words in the paradigm, and their variations in collocations"].[15]

Considering the typical functions and usage situations, EDU-*Col* can be characterized as a reference work in Engelbeg and Lemnitzer's terms (2009), whereas the functions type is the communication-orientated following Tarp's (2002) definition, and particularly, the dictionary is primarily expected to:

(1) to assist the production of texts in foreign languages;

(2) to assist the translation of texts between the native language and a foreign language;

(3) to assist the users in the production of texts in the native language where the users might have doubts about the combinability of words in the native language, especially in the specialized domains, such as legal or economics domain.

[15] Note: the translation from German is mine.

4 Collocations: definitions and approaches

4.1 Approaches to collocations

There is no agreement as to the definition of collocations, with varying approaches as far as collocations treatment is concerned available in the literature. Herbst (1996) singles out at least three major lines of thinking related to the interpretation of collocations. In particular, those are:

(i) a text-oriented approach
(ii) a statistically oriented approach, and
(iii) a significance-oriented approach.

Further, we will examine the views on collocations treatment by some of the most prominent scholars who influenced the shaping of the term 'collocation' following these three approaches outlined by Herbst (1996) and other approaches.

4.1.1 Collocations and language teaching

The first among the scholars who investigated the issue of collocations in language learning were Palmer and Hornby. However, their interpretation of a collocation, which dates back to 1920-1930s, diverges from the current understanding of the phenomenon, since they used the term rather broadly, with the combinations that would now be classified as idioms also belonging to collocations, according to their view. According to Cowie (1999, p. 56 as cited in Poulsen, 2022, p. 29), Palmer objected to the term 'idiom', instead using 'collocation', which was established and theoretically defined first by Firth (1951) some 20 years later. However, their collocation project (Palmer, 1933a, 1933b) was the first large-scale analysis of phraseology taking into account the needs of the foreign learners[16] (ibid).

Their pedagogical approach can be explicated with the following words of Palmer (Cowie, 1999, p. 52., as cited in Poulsen, 2022, p. 30):

[16] Palmer was appointed as a linguistic adviser to the Ministry of Education in Japan in 1922, where he examined the English teaching process in schools settings and later became a director of a research institute, with an aim to "reform" the methods of English teaching, research and experiment in linguistics, and the training of teachers" (Cowie, 1999, p. 5 as cited in Poulsen, 2022, p. 28).

"It will tend to confirm his [the language teacher's] impression that it is not so much the words of English nor the grammar of English that makes English difficult, but that that vague and undefined obstacle to progress in the learning of English consists for the most part in the existence of so many odd comings-together-of words".

The fact that Palmer and Hornby at their time acknowledged word combinations as a challenge learners face while learning a foreign language undoubtedly makes their research important. Their ideas can be confirmed by more recent studies, which involved experiments with native speakers and learners of English. For instance, Herbst (2014, p. 389) gives an account of a series of tests with German student learners of English from two German universities in Augsburg and Erlangen, and English students, native speakers, from the three universities from the UK. The tests included such tasks as filling in the gaps, completing sentences, and translating typical collocations into English. The results revealed that the English native speakers performed significantly better than German students of English, with doublefold or even threefold percentages for certain sentences in the completion tasks. Herbst (2014, p. 391) provides an illustration of the completion exercise, in which the sentence such as "*A number of objections were but ...,*" were completed correctly as *raise objections* only by 30% of the German students. As far as translation tests are concerned, for instance, the German collocation *schwacher Tee* was translated as *weak tea* by 27% of the German students and as *light tea* by 33%.

4.1.2 Firth

Collocation as a linguistic term is believed to be used by Firth in 1951, although the phenomenon had been known since Samuel Johnson's time, and collocation had been attempted to be defined already in the first edition of the *OED.* In addition, it was used for reports and specialized dictionaries starting from Palmer and further (Barnbrook et al., 2003, p. 35). As Barnbrook et al., (ibid) note, "The real significance of his approach is that makes it possible to consider collocation not just as an observable effect of language use, but as an important element of the causes of language patterns". In particular, in a paper "Modes of meaning" (Firth, 1951) underlines the importance of collocations: "At this point in my argument, still confining our references to the language of limericks, I propose to bring forward as a technical term, meaning by 'collocation', and to apply the test of 'collocability'" (Firth, 1951, p. 123, as cited in Barnbrook et al., 2003, p. 37). Firth provides examples of the word *ass,* which according to his view collocates with *you silly* preceding it and with other ways addressing or naming a person, as in "*He is an ass. You silly ass! Don't be an ass!"*, and points out the limited possibilities of collocation with preceding adjectives *silly*, *obstinate*, *stupid*, *awful*, occasionally *egregious* (ibid).

Herbst (1996, p. 380) argues that there is a certain vagueness in "collocability" as defined by Firth. What seems to be clear though is that Firth viewed collocation as "a co-occurrence relation between individual lexical items and not a relation between classes of items" and "the use of the term collocation is not restricted to combinations of two words" (Herbst, 1996, ibid). Hence, the sentences such as "*You silly ass"* or *Don't be such an ass"* are given as examples of collocations in his works.

4.1.3 Text-oriented approach (Halliday and Hasan)

According to the text-oriented approach to collocations as represented by Halliday and Hasan (1976), all words that contribute to the cohesion of the text are regarded as collocations. Here the notion of "collocational cohesion" is relevant, which following the scholars' view, arises due to the typical co-occurrence of lexical items in similar environments or as precisely put forward in their definition (Halliday and Hasan, 1976, p. 287 as cited in Herbst, 1996, p. 381) of a collocation which they view as:

> "A cover term for the cohesion that results from the co-occurrence of lexical items that are in some way or other typically associated with one another, because they tend to occur in similar environments".

As it is noted in Herbst (1996, p. 381), the use of the term collocation by Halliday and Hasan (1976) does not presuppose:

(a) that collocations have to be immediately adjacent to each other
(b) that the elements of a collocation have to enter into an immediate dependency relation.

It is a rather broad interpretation of collocations, under which in a text about a certain topic, words that are usually used to describe that topic would be treated as collocations.

4.1.4 Statistically oriented approach (Sinclair)

Another approach to the interpretation of collocations, which relies purely on frequency is the statistically oriented approach. It is linked to the Cobuild project and John Sinclar, according to whom, a collocation can be broadly defined as "the co-occurrence of two items in a text within a specified environment" (Jones & Sinclair, 1974, p. 19, as cited in Herbst, 1996, p. 381).

In Sinclair's view (1966, p. 415 as cited in Herbst, 1996, p. 382), a collocation has the following structure: a *node,* which refers to "an item whose collocations we are studying", and a *span*, which can be defined as "the number of lexical items on each side of a node that we consider relevant to that node". The final element is the *collocates*, i.e., "items set by the environment set by the span". Sinclair and Jones were the first to analyze the texts computationally, with the results revealing that in 95% of nodes, a span of collocates includes four words in the left and right directions (Jones & Sinclair 1974, p. 21, as cited in Nesselhauf 2004, p. 8). In addition, the scholar differentiates between casual and significant collocation, which are rare and frequent collocations accordingly.

Herbst (1996, p. 382) argues that frequency of co-occurrence alone is not a sufficient criterion for a word combination to be significant. He draws an example of analyzing a small corpus, where one can arrive at the conclusion that the only salient and frequent collocates of a word such as *house* are the determiners (*the* and *this)* and the verb *sell*, which is "neither particularly surprising nor particularly interesting", as pointed by Herbst (199, ibid). Thus, to achieve significant results, corpora should be sufficiently large.

4.1.5 Significance-oriented approach (Hausmann)

The representative of the significance-oriented approach to collocations treatment is Hausmann (1984) who phreases his interpretation of a collocation as follows:

"Wörter mit begrenzter Kombinierbarkeit verbinden sich entsprechend differenzierter semantischer Regeln und einer gewissen zusätzlichen Üblichkeit mit Wörtern, zu denen sie in Affinität stehen. Affinität sei definiert als die Neigung zweier Wörter, kombiniert aufzutreten" (Hausmann 1984, p. 398).

[Words with limited combinability are combined according to distinct semantic rules and a certain additional commonality with words to which they have affinity. Affinity can be defined as the tendency of two words to occur in a combination][17].

Collocations in Hausman's theory fall within a group of non-fixed combinations, which on the contrary to fixed ones, namely idioms and compounds, can be subdivided into the following types:

[17] Note: translation from German is mine.

(a) co-creations, i.e., free combinations, where the constituents join esch other based on the speaker's creativity.
(b) collocations, i.e., not creatively combined but put together out of some convention;
(c) counter-creations, i.e., non-typical combinations used in fiction and advertisements to create a special effect (Hausmann 1984, as cited in Nesselhauf, 2004, p. 16).

In order to describe the structure of collocations, Hausmann (1984) used the terms 'base' and 'collocator' (Basis and Kollokator in German), known in literature as the base-collocator principle. According to this principle, the constituent parts of a collocation are in a dependency relation, where the base is considered to be "static" and semantically autonomous, thus not changeable, whereas a collocator is assigned to the base and is semantically relational. The two elements form a collocation according to the following 5 structural types: noun + noun, noun + adjective, noun + verb, verb + adverb, adjective + adverb (Hausmann 1989, p. 1010).

4.1.6 Mel'čuk

Similarly to Hausmann, Mel'čuk, who is known for his "'Meaning-Text-Theory" and a model of "lexical functions" (LFs), puts a large emphasis on the collocations' compositionality. In his words (Mel'čuk, 2012, p. 39):

> "[A] collocation is binary – it consists of two major elements: a base, lexical expression chosen freely by the speaker [..], and a collocate, lexical expression chosen as a function of the base to express a given meaning bearing on the base".

According to this view, compositionality refers not only to syntactical relations – it also refers to semantics since the meaning of a collocation can be divided into two parts. The first part corresponds to the base and the second one relates to the collocate. Mel'čuk (2012, p. 39) stresses that the "meaning of the base is always the semantic pivot of the collocation" (2012, p. 39). It should be noted that while the base has a rather stable meaning, the collocate outside the collocation does not usually retain its meaning as it used to have in a combination Mel'čuk's terms its "context-imposed signified". The scholar (Mel'čuk, 2012, ibid) provides such examples as the collocation *sit for an exam* means "undergo an exam", where the verb *sit* has the meaning of "undergo", but in a general English dictionary, the verb *sit* does not necessarily have to indicate this meaning. This also leads to the discrepancy in the combinability of words across different languages; therefore, not infrequently collocations cannot be rendered into another language word-for-word. For instance, in the English collocation *make a decision,* the collocate is the verb *make*, in British English also, *take (a decision)*, whereas in German the collocation is *eine Entcheidung treffen,* where the collocate is the verb *treffen,* which corresponds to the English *"meet".*

In addition, Mel'čuk, (2012, p. 40) distinguishes between two main types of collocations, namely *standard and non-standard collocations,* with an example of the former being ("John despairs": "*John is in despair". ~ "John is desperate". ~ "despair seized* 〈≈ *overcame*〉 *John"*, and the latter: "*leap year"*, **r** = "having 366 days"; "*black coffee"*, **r** = "with no dairy product added". The main criteria which helps to see the difference between the two types lies in the number of bases the semantic relation **r** of a collocation "Base–**r**–Collocate" is applicable. In the first case, the number is large, whereas in the second case, the number is scarce.

4.2 Collocations as viewed in *EDU-Col*

Herbst (1996, p. 383) argues that the usefulness of Halliday and Hasan's approach to collocation interpretation is limited since the relations that in their approach refer to cohesion are usually interpreted through lexical fields or paradigmatic relations (e.g., synonymy). Moreover, Hasan (1984, as cited in Herbst 1996, p. 381) modifies what was covered by collocation in the 1976 model by the so-called lexical chains.

Considering the above, Herbst (1996, p. 384) notes the two approaches, namely the one that relies on the frequency and the "one where significance is seen in such terms as the unpredictability of the combination on semantic grounds" can be considered. What is of relevance for *EDU-Col* is the syntactic relations between the two lexical units, that is the collocations' elements (Hausmann 1984, 1989; Mel'čuk, 2012). These elements do not have the same status within a collocation, with the base being freely selected, whereas the second element, namely the collocate is predetermined by the former (Mel'čuk, 2012).

Garcia et al., (2019c, p. 2) notes that this distinction between the weight of the elements of collocations is particularly important for the automatic generation of translations of extracted collocations. In particular, most of the approaches to automatic extraction of translations of multiword expressions often consider the semantic load of each multiword expressions' element as equal, which results in word-for-word translations. However, there exists a large number of word combinations, elements of which cannot be translated into the target language word-for-word, some of which being the combinations as described by the outlined approach.

In addition, adopting the concept of the base, which is according to Hausmann's view (1984, 1989) is presumably known to the users, whereas the collocates need to be learned or revised (which will also affect how we can organize the collocations in the dictionary articles; it will be elaborated in more detail in the next chapters), the collocates for the three base types, namely nouns, adjectives, and verbs, will be the elements users are searching for.

5 Methodology used in the compilation of *EDU-Col*

The methodology for compiling *EDU-Col* follows the approach presented by Orenha-Ottaiano et al. (2021), namely automatic extraction of linguistic data as described in Garcia et al. (2019a, 2019c) and supplemented by manual validation of the extracted dictionary information types. The choice of the methodology can be justified by the intention to use the capabilities of corpora and the latest developments in NLP to make the process of the dictionary compilation time-efficient while maintaining the quality of the lexicographic data by manual review.

The procedure starts with the compilation of corpora for each language, which will be represented in the dictionary. Each corpus is then tokenized, POS-tagged, and annotated with dependency relations. Collocations are extracted automatically utilizing dependencies computed for the lists of part of speech bases, which are the most frequent nouns, verbs, and adjectives from each corpus. As a result of this procedure, collocation candidates with noun, verb and adjective bases are extracted, with the output data containing the most important statistical measures (Evert et al., 2017; Garcia et al., 2019b as cited in Orenha-Ottaiano et al., 2021). Translations of collocations candidates for language 1 in language 2 and language 3 are generated using distributional models (Garcia et al., 2019c). A final step includes the extraction of a set of examples (Kilgarriff et al., 2008) for the previously generated collocations candidates. Manual validation is performed on the stages of extraction of part of speech collocation bases and collocation candidates. The steps will be elaborated on in detail in the next chapters.

5.1 Dictionary basis

The dictionary basis of *EDU-Col* is constituted by primary sources, i.e., natural language collected in corpora for gaining insights into the syntactic combinability of words in the languages covered by *EDU-Col*, English, German, and Ukrainian, and subsequently generating the required dictionary information types, namely collocation candidates, examples, and collocations translations. For the purpose of this work, the corpora should be either compiled from scratch or utilizing the publicly available corpora as well as repositories of free texts in the corresponding languages.

In this case, the use of corpora collection provided by the linguistics analysis tool *Sketch Engine* (Kilgarriff et al., 2004, 2014) is not considered since it does not use dependency parsing required in line with the methodology chosen for the project. In addition, we aim to cover not only general language collocations but also domain-specific ones, such as law, economics, environment, etc., whereas the corpora present in *Sketch Engine* do not offer such a variety of domain-specific resources for all three languages by default.

For the corpora design, two essential concepts were taken into consideration, namely representativeness and balance. According to Rundell et al (2013, p. 1339-1341), representativeness refers to the register variation of texts. The scholars provide a checklist of corpora design criteria among which such aspects as language (regional and dialectal variation), timespan (synchronic vs. diachronic), mode (spoken vs. written texts), medium (for written texts, e.g., a novel, a magazine, and for spoken conversations, lectures; the web is also regarded as a new channel), domain (also ‘topic’ or subject-matter of a text). As far as the balance criteria, the size of a corpus and its main subcorpora categories play an important role. Thus, in line with the methodology presented in the previous section and the type of the dictionary, corpora for each language represented in *EDU-Col* should be compiled following certain criteria, in particular:

- type of corpora: non-parallel[18] monolingual corpora in English, Gerrman, and Ukrainian;

[18] Non-parallel corpora are easier to get than parallel corpora. Moreover, the approach by Garcia et al. (2019c), which will be used for obtaining translations, allows generating automatic translations of candidate collocations based on a non-parallel (monolingual) corpus.

- medium: primarily written language;
- scope: general language corpora for general language collocations and domain-specific corpora for domain-specific collocations;
- corpus preparation: syntactically annotated;
- language stage of the texts to be included in the corpora: synchronic language texts.

As far as the size[19] of corpora is concerned, in order to extract meaningful results, each general language corpus was planned to amount to over 1 billion tokens, whereas domain-specific corpora 50 million tokens. For the general language collocations to be used in *the Dictionary,* the corpora will comprise the following data visualized in the table below; the token size count is provided accordingly.

Corpora language	**Sources of data**	**Token size**
Ukrainian	UberText corpora comprise a volume of texts from Ukrainian periodicals, news, Wikipedia pages, and fiction. In order to comply with the license restrictions of some publications, which were used in the corpora and which do not allow publication of texts in their original form, the texts' sentences were shuffled.	**0,65 bln**
	Ukrainian corpus[20] based on the free online library of Ukrainian language literature Chtyvo[21]	**< 0,6 bln**
German	deWaC (Baroni et al., 2009): this is generated from the web, with the criteria, which was used to restrict the texts being the .de domain. To compile this corpus, lists of words from SudDeutsche Zeitung corpus were used, as well as the lists of basic vocabulary.	**1,7 bln**
English	ukWaC (Baroni et al., 2009): similarly to deWaC, ukWaC was generated worm web, and the domain used for restriction was .uk. Medium-frequency words from the BNC were used as seeds.	**2 bln**

[19] Back in the 1970s Halliday suggested that a corpus of at least 20 mln words is necessary for collocational analysis (Halliday, 1966, p. 159 as cited in Herbst, 1994, p.382).

[20] http://korpus.org.ua/

[21] https://chtyvo.org.ua/

Table 1. Sources of data used in the general corpora for English, German, and Ukrainian language

Corpora will allow illustrating the dictionary with examples collected from authentic texts using software-based generation processes. Here, we refer to the evidence-based examples or citations, unlike the "pedagogic" examples made up by lexicographers, in order to show the collocations in context.

Since the annotation process of large text files requires significant computational resources and power, within the frames of this Master thesis project, we are limiting the extraction of dictionary data to domain-specific collocations, as elaborated on in detail in the next section.

5.2 *EDU-Col:* domain-specific collocations

For the experiment, we compiled a significantly smaller 50 million token corpora for each language representing the legal domain. In particular, the approach was to select the most important legal documents available for each language and balance it according to certain criteria. Below is the overview of the legal domain corpora. The English legal domain corpus compiled for the purposes of this Master thesis project comprises legal documents that represent both the UK and US legal systems, as well as a number of treaties adopted by the EU institutions to account for the US and UK English varieties and different legal systems (see Table 2)..

Corpora language	**Sources of data**	**Token size**
English	The United States Code[22] (Chapters 1-20) UK Law Public General Acts[23] (2008-2022) EU Treaties[24]	**< 50 mln**
Ukrainian	Codes, laws and resolutions (Verkhovna Rada) Decrees and orders (President of Ukraine) Resolutions and orders (Cabinet of Ministers) Coursebooks	**< 50 mln**

[22] https://uscode.house.gov/
[23] https://www.legislation.gov.uk/ukpga
[24] https://eur-lex.europa.eu/collection/eu-law/treaties/treaties-force.html

German	Contents of the https://www.gesetze-im-internet.de/ web site published by the Directorate-General for Communications Networks, Content and Technology (2019)[25]	**< 50 mln**

Table 2. Sources of data used in the legal domain corpora for English, German, and Ukrainian language

The Ukrainian law corpus was also compiled specifically for this task and comprises data from the official web portal of the legislation of Ukraine[26]. In particular, the codes, laws, and resolutions issued by the Parliament of Ukraine (Verkhovna Rada), decrees and orders by the President of Ukraine, as well as resolutions and orders by the Ukrainian government (Cabinet of Ministers) constitute the major part of the corpus. To achieve the 50 mln token count the corpus was enlarged with the coursebooks on legal topics which pertain to the legislature of Ukraine[27].

For the German language, a legal corpus based on the current federal law published in their current version by the Federal Ministry of Justice and the Federal Office of Justice on their web-based portal was already available for free use.

5.2.1. Corpora preparation

Before utilizing corpora for generating dictionary data, it should be properly prepared, including preparing the metadata and preparing the text (Kilgarriff & Kosem, 2011, p. 32). While metadata allows a lexicographer to find out from what kind of text a particular instance had been extracted (i.e. such as its date of publication, author, mode, and domain), and also limit the searches to particular text types, preparing the text is even more important. Particularly, such processes as tokenization, i.e., distinguishing tokens (usually the words,

[25]https://elrc-share.eu/repository/browse/german-legal-monolingual-corpus-from-the-contensts-of-the-httpswwwgesetze-im-internetde-web-site/c2ec783cac7c11e9a7e100155d026706e433e10f16324131bb3dd1a295dec7c4/

[26] https://zakon.rada.gov.ua/

[27] Initially we used a Corpus of Laws and Legal texts available on the web https://lang.org.ua/en/corpora/#anchor7 to compute the collocation candidates for the Ukrainian part; however, it turned out it seems to contain a part of documents written in Russian, which resulted in poor quality of output data. Therefore, a decision was taken to compile a corpus from scratch.

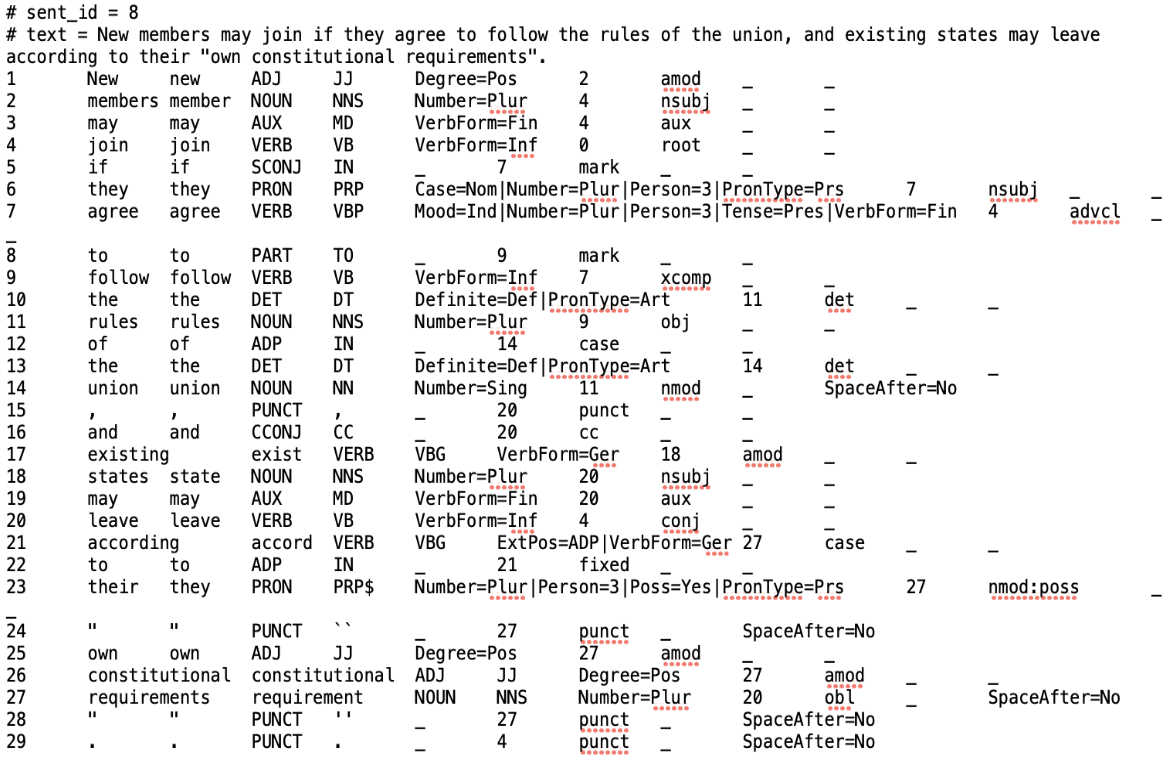
which the user typically searches for), lemmatization, that is finding the base form of the word, part-of-speech tagging, and parsing used to annotate the syntactic structure of sentences will allow retrieving automatically collocation candidates and examples required for our dictionary project.

Among the existing natural language processing tools (e.g., Natural Language Processing Toolkit (Bird et al., 2009), SpaCy (Honnibal & Montani, 2017), etc) we opted for UDPipe (Straka et al. 2016, Straka. & Straková, 2017), which is an open-source tool that automatically generates sentence segmentation, tokenization, POS tagging, lemmatization, and dependency trees, using for training data Universal Dependencies treebanks, i.e., a framework for annotation of grammar (parts of speech, morphological features, and syntactic dependencies) for different languages. What is worth noting is that the use of dependency parsing allows for identifying long-distance dependencies, which cannot be spotted in immediate surroundings of a word (Garcia et al., 2019a, p. 751). As noted by Seretan (2013, p. 16), this is especially important for extracting collocations in cases of syntactic variations, for instance, in interrogative sentences (e.g., "Which *challenges* do online media *face* in terms of press freedom?") or when one of the collocation components occurs in a different clause. For instance, in the sentence part "various global *challenges* that we inevitably have to *face"*, the verb *face* belongs to the dependent clause, and the noun *challenges*, which is an object of the verb *face* is contained in the main clause.

UDPipe provides pre-trained language models for a variety of languages, including a model for the Ukrainian language, which is one of the motivating factors for the choice of UDPipe (for instance, SpaCy does not offer a pre-trained Ukrainian model, and its multi-language model undoubtedly cannot be expected to provide accurate enough results). Moreover, as Straka et al. (2016) note, automatic natural language processing of large texts often poses challenges since the texts are usually first processed by basic processing steps, whereas UDPipe is simple to use as it consists of one binary and one model and does not require any other external data for performing these tasks.

In addition, it allows users to train their own models. Since the latest pre-trained models available for download are Universal Dependencies 2.5, based on Universal Dependencies 2.5 treebanks, we trained new language models using the most recent at the time of work on the project 2.9 Universal Dependencies treebanks (Zeman et al., 2021). Thus, each corpus compiled for the three languages for the purpose of our project was processed by UDPipe, with the resulting CoNLL-U files (an excerpt of the processed English corpora can be seen in Figure 2).

In CoNLL-U format[28], the annotation of each token is presented in a separate line that contains ten fields with various information. In particular, the first column shows a word index. New sentences start from index 1. The index is followed by word form or punctuation symbol and lemma of word form in the 2nd and 3d column correspondingly. The 4th and 5th columns show the universal part-of-speech tag (UPOS) and language-specific part-of-speech tag (XPOS); for instance, NOUN and NNS (noun plural) for the word form *members* (see Figure 2, index 2). From the 6th column onward, the following information types are displayed: a list of morphological features (underscore if not available, column 6), HEAD (head of the current word, which can be the index or zero, column 7), universal dependency relation (deprel) to the HEAD (column 8). The HEAD and dependency relation columns define the basic dependencies and are used to encode a dependency tree over words. Finally, in column 9, there may be displyed an enhanced dependency representation with additional dependency relations (however, this element is optional), whereas column 10 may show any other annotation (not supported in our case in Figure 2).



```
# sent_id = 8
# text = New members may join if they agree to follow the rules of the union, and existing states may leave
according to their "own constitutional requirements".
1       New     new     ADJ     JJ      Degree=Pos      2       amod    _       _
2       members member  NOUN    NNS     Number=Plur     4       nsubj   _       _
3       may     may     AUX     MD      VerbForm=Fin    4       aux     _       _
4       join    join    VERB    VB      VerbForm=Inf    0       root    _       _
5       if      if      SCONJ   IN      _       7       mark    _       _
6       they    they    PRON    PRP     Case=Nom|Number=Plur|Person=3|PronType=Prs    7       nsubj   _       _
7       agree   agree   VERB    VBP     Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin   4       advcl   _
_
8       to      to      PART    TO      _       9       mark    _       _
9       follow  follow  VERB    VB      VerbForm=Inf    7       xcomp   _       _
10      the     the     DET     DT      Definite=Def|PronType=Art       11      det     _       _
11      rules   rules   NOUN    NNS     Number=Plur     9       obj     _       _
12      of      of      ADP     IN      _       14      case    _       _
13      the     the     DET     DT      Definite=Def|PronType=Art       14      det     _       _
14      union   union   NOUN    NN      Number=Sing     11      nmod    _       SpaceAfter=No
15      ,       ,       PUNCT   ,       _       20      punct   _       _
16      and     and     CCONJ   CC      _       20      cc      _       _
17      existing        exist   VERB    VBG     VerbForm=Ger    18      amod    _       _
18      states  state   NOUN    NNS     Number=Plur     20      nsubj   _       _
19      may     may     AUX     MD      VerbForm=Fin    20      aux     _       _
20      leave   leave   VERB    VB      VerbForm=Inf    4       conj    _       _
21      according       accord  VERB    VBG     ExtPos=ADP|VerbForm=Ger 27      case    _       _
22      to      to      ADP     IN      _       21      fixed   _       _
23      their   they    PRON    PRP$    Number=Plur|Person=3|Poss=Yes|PronType=Prs    27      nmod:poss       _
_
24      "       "       PUNCT   ``      _       27      punct   _       SpaceAfter=No
25      own     own     ADJ     JJ      Degree=Pos      27      amod    _       _
26      constitutional  constitutional  ADJ     JJ      Degree=Pos      27      amod    _       _
27      requirements    requirement     NOUN    NNS     Number=Plur     20      obl     _       SpaceAfter=No
28      "       "       PUNCT   ''      _       27      punct   _       SpaceAfter=No
29      .       .       PUNCT   .       _       4       punct   _       SpaceAfter=No
```

[28] Format of the Universal Dependencies. Retrieved July 10, 2022, from https://universaldependencies.org/format.html

Figure 2. An excerpt of the resulting CoNLL-U file after processing the English corpus with UDPipe

5.2.2 Extraction of collocation candidates

Collocation candidates for English, German, and Ukrainian for EDU-Col are obtained relying on the approach by (Garcia et al., 2019a), according to which dependency parsing and statistical association measures are utilized for automatic generation of collocations. The processes involved in this approach and the steps we have taken to obtain the collocation candidates will be described further.

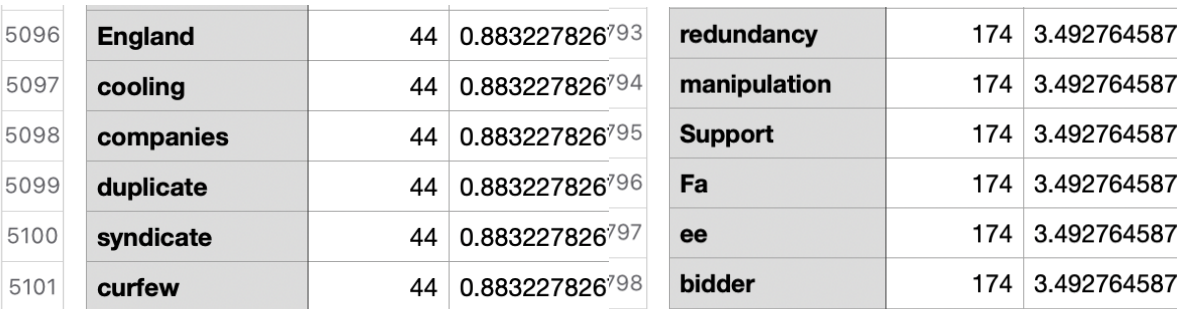
Firstly, to obtain collocation candidates, the lists of part of speech bases were generated from the corpora for each language., i.e., the most frequent nouns, verbs, and adjectives were retrieved, which were subsequently used to extract dependency relations. Manual validation was then conducted for the extracted part of speech bases, as a result of which the output containing typographical errors, non-lexical items, and other noise was eliminated in order to make further results more quality (Orenha-Ottaiano et al., 2021).

The number of extracted bases is indicated further (see Table 3) for each part of speech and the resulting numbers after the manual validation for the English, Ukrainian, and German languages respectively.

Corpus (Legal Domain)	Nouns		Adjectives		Verbs	
	Before	After	Before	After	Before	After
English	10 124	6148	4 143	2571	3 498	1761
Ukrainian	13580	6662	6940	4873	3379	1974
German	23947	13256	6251	4615	3179	1815

Table 3. The number of extracted bases before and after manual validation

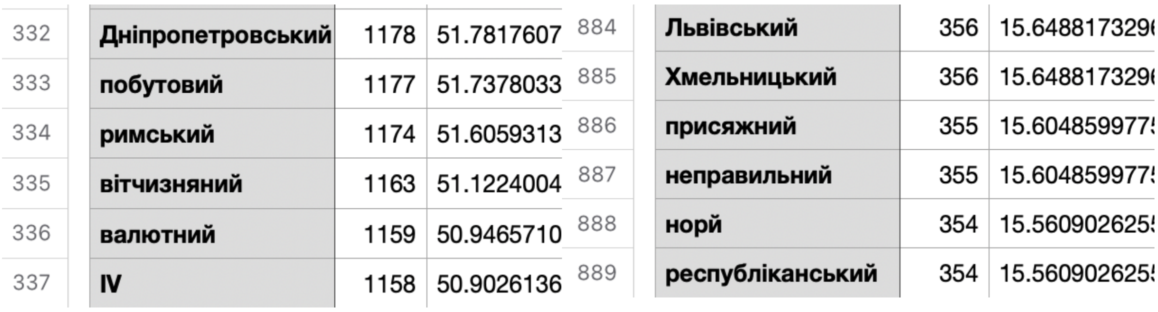
The noise, which was eliminated in the process of validation, included mainly, typographical errors, misspelled words, incorrectly processed parts of speech, and proper names. Figure 3 (a,b,c) shows the excerpts from the output files for English noun bases, Ukrainian adjective bases, and German verb bases. The bases files contain a lemma in the first column and the frequency of the lemma in the corpus in the second column, whereas the third column is constituted by the frequency of the lemma per million. As can be seen in Figure 3a, typographical elements (*Fa, ee*) and the proper name (*England*) had to be eliminated from this short span of the English noun base file. Such elements as *IV*, proper names *(Дніпропетровський, Львівський, Хмельницький*), and combination of letters (*норй*) were deleted from the Ukrainian excerpt of adjective bases (Figure 3b). Finally, only the word *verpacken* was preserved from the German verb bases excerpt (Figure 3c) since other elements are not verbs in the infinitive form.



5096	**England**	44	0.883227826
5097	**cooling**	44	0.883227826
5098	**companies**	44	0.883227826
5099	**duplicate**	44	0.883227826
5100	**syndicate**	44	0.883227826
5101	**curfew**	44	0.883227826

793	**redundancy**	174	3.492764587
794	**manipulation**	174	3.492764587
795	**Support**	174	3.492764587
796	**Fa**	174	3.492764587
797	**ee**	174	3.492764587
798	**bidder**	174	3.492764587

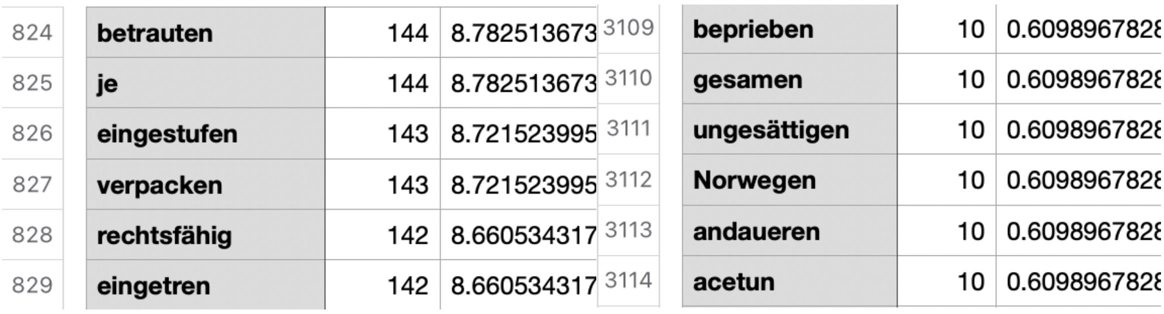
(a) English noun bases



332	**Дніпропетровський**	1178	51.7817607
333	**побутовий**	1177	51.7378033
334	**римський**	1174	51.6059313
335	**вітчизняний**	1163	51.1224004
336	**валютний**	1159	50.9465710
337	**IV**	1158	50.9026136

884	**Львівський**	356	15.648817329
885	**Хмельницький**	356	15.648817329
886	**присяжний**	355	15.604859977
887	**неправильний**	355	15.604859977
888	**норй**	354	15.560902625
889	**республіканський**	354	15.560902625

(b) Ukrainian adjective bases



824	**betrauten**	144	8.782513673
825	**je**	144	8.782513673
826	**eingestufen**	143	8.721523995
827	**verpacken**	143	8.721523995
828	**rechtsfähig**	142	8.660534317
829	**eingetren**	142	8.660534317

3109	**beprieben**	10	0.609896782
3110	**gesamen**	10	0.609896782
3111	**ungesättigen**	10	0.609896782
3112	**Norwegen**	10	0.609896782
3113	**andaueren**	10	0.609896782
3114	**acetun**	10	0.609896782

(c) German verb bases

Figure 3. Types of noise eliminated in the process of manual validation

In the next stage, these parts of speech bases were used to extract dependency relations necessary for obtaining collocation candidates. Collocation candidates are identified as the pairs of lemmata that belong to the dependency relations. The generation of collocation candidates is based on lemmata, with the inflected forms of a word to be covered by a single dictionary article. As Garcia et al. (2009a, p. 751) point out, the number of inflected forms of a verb in Spanish can be very extensive. The same applies to Ukrainian and to a lesser extent German languages since they are also inflectional languages.

Thus, the second step was extracting dependency relations for the part of speech bases we obtained in the first step. In particular, in the words of Garcia et al., (2019a, p. 750), dependency parsing helps to identify the syntactic relations between two lexical units by establishing binary relations between words in a sentence. Thus, the focus during the extraction of dependency relations for *EDU-Col* was on the types of relations for the noun, verb, and adjective bases presented further in Table 4.

Base	Collocation types	Syntactic relation[29]	Example
Noun	verb-noun	*obj*	*reach (an) agreement*
	noun-verb	*nsubj*	*law prohibit(s)*
	adjective-noun	*amod*	*basic rule*
	verb-preposition-noun	*obl*	*burst into tears*
	noun-prep-noun or noun-noun	*nmod and compound*	*ceasefire agreement*
Verb	verb-adjective	*xcomp*	*deem appropriate*
	verb-adverb	*advmod*	*think deeply*
Adjective	adjective-adverb	*advmod*	*highly successful*

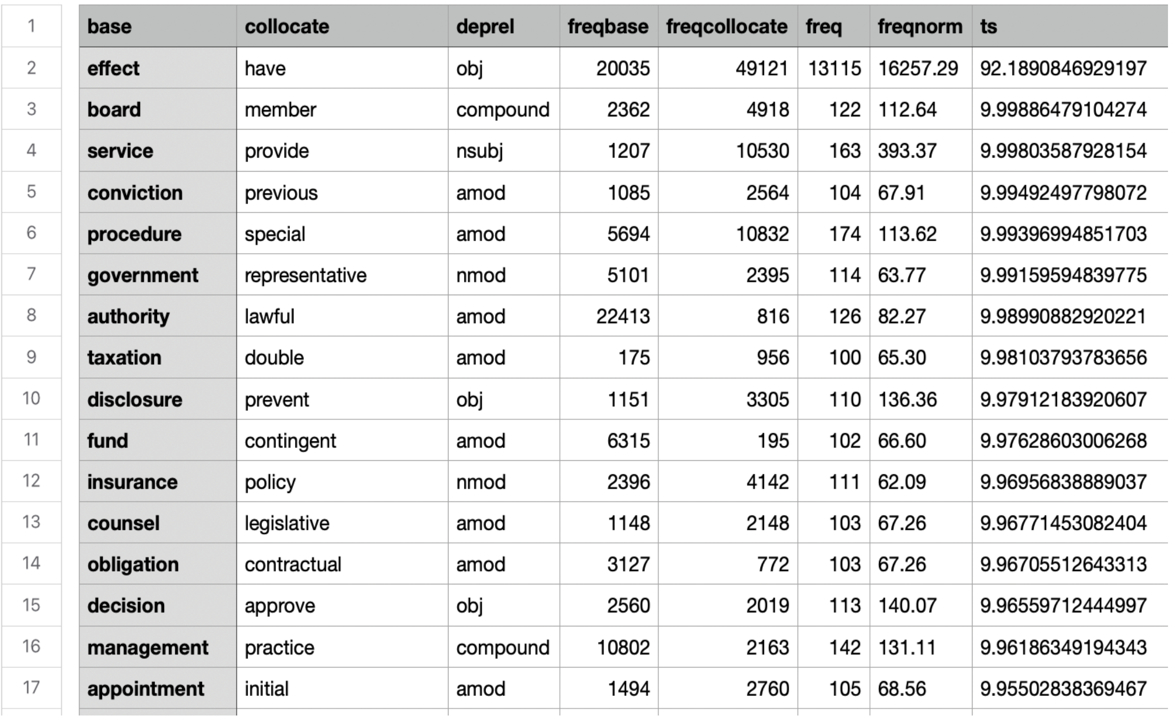
Table 4. Types of extracted dependency relations for noun, verb, and adjective bases

Then, different statistical association measures, namely t-score (ts), MI (mi), Dice (di), ΔP(dp), log-likelihood (ll), simple-ll (sl), z-score (zs), are applied in the process of extraction to spot the collocability of two syntactically related words by ranking the frequent co-occurrence between word pairs using numerical values (Garcia et al., 2019b). In addition,

[29] https://universaldependencies.org/u/dep/index.html

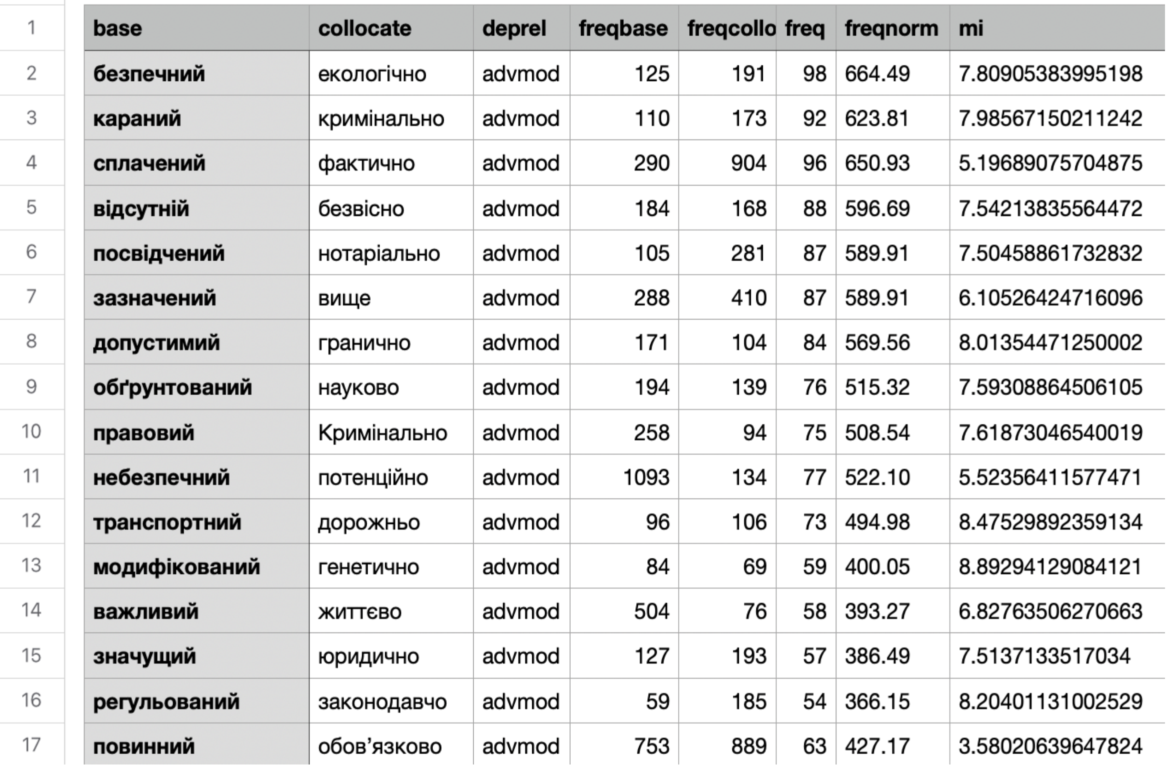
the ranking is supplemented with frequency data to select the top-*n* combinations (Krenn & Evert, 2001 as cited in Garcia et al., 2019a, p. 751).

As it was mentioned in the previous sections, the approach adopted for the dictionary compilation presupposes the use of non-parallel corpora. Hence, the collocation candidates are extracted separately for each language from the respective corpora. Further, the illustration from the output files containing collocation candidates of the legal domain is provided for the three languages, including frequencies and some of the statistical data. As a result, we obtained three large datasets of collocation candidates for English, Ukrainian, and German (Figures 4, 5, and 6).



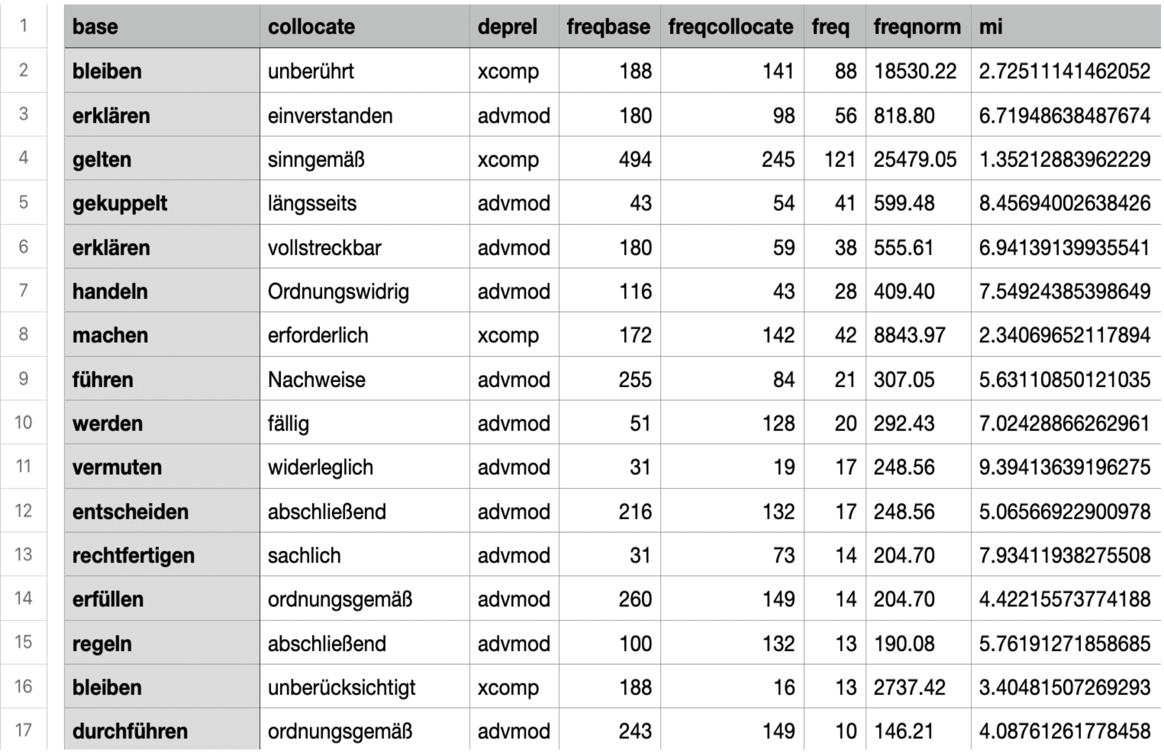
1	base	collocate	deprel	freqbase	freqcollocate	freq	freqnorm	ts
2	**effect**	have	obj	20035	49121	13115	16257.29	92.1890846929197
3	**board**	member	compound	2362	4918	122	112.64	9.99886479104274
4	**service**	provide	nsubj	1207	10530	163	393.37	9.99803587928154
5	**conviction**	previous	amod	1085	2564	104	67.91	9.99492497798072
6	**procedure**	special	amod	5694	10832	174	113.62	9.99396994851703
7	**government**	representative	nmod	5101	2395	114	63.77	9.99159594839775
8	**authority**	lawful	amod	22413	816	126	82.27	9.98990882920221
9	**taxation**	double	amod	175	956	100	65.30	9.98103793783656
10	**disclosure**	prevent	obj	1151	3305	110	136.36	9.97912183920607
11	**fund**	contingent	amod	6315	195	102	66.60	9.97628603006268
12	**insurance**	policy	nmod	2396	4142	111	62.09	9.96956838889037
13	**counsel**	legislative	amod	1148	2148	103	67.26	9.96771453082404
14	**obligation**	contractual	amod	3127	772	103	67.26	9.96705512643313
15	**decision**	approve	obj	2560	2019	113	140.07	9.96559712444997
16	**management**	practice	compound	10802	2163	142	131.11	9.96186349194343
17	**appointment**	initial	amod	1494	2760	105	68.56	9.95502838369467

Figure 4. Extracted English collocation candidates for noun as a base sorted by t-score



1	base	collocate	deprel	freqbase	freqcollo	freq	freqnorm	mi
2	**безпечний**	екологічно	advmod	125	191	98	664.49	7.80905383995198
3	**караний**	кримінально	advmod	110	173	92	623.81	7.98567150211242
4	**сплачений**	фактично	advmod	290	904	96	650.93	5.19689075704875
5	**відсутній**	безвісно	advmod	184	168	88	596.69	7.54213835564472
6	**посвідчений**	нотаріально	advmod	105	281	87	589.91	7.50458861732832
7	**зазначений**	вище	advmod	288	410	87	589.91	6.10526424716096
8	**допустимий**	гранично	advmod	171	104	84	569.56	8.01354471250002
9	**обґрунтований**	науково	advmod	194	139	76	515.32	7.59308864506105
10	**правовий**	Кримінально	advmod	258	94	75	508.54	7.61873046540019
11	**небезпечний**	потенційно	advmod	1093	134	77	522.10	5.52356411577471
12	**транспортний**	дорожньо	advmod	96	106	73	494.98	8.47529892359134
13	**модифікований**	генетично	advmod	84	69	59	400.05	8.89294129084121
14	**важливий**	життєво	advmod	504	76	58	393.27	6.82763506270663
15	**значущий**	юридично	advmod	127	193	57	386.49	7.5137133517034
16	**регульований**	законодавчо	advmod	59	185	54	366.15	8.20401131002529
17	**повинний**	обов’язково	advmod	753	889	63	427.17	3.58020639647824

Figure 5. Extracted Ukrainian collocation candidates for adjective as a base sorted by t-score



1	base	collocate	deprel	freqbase	freqcollocate	freq	freqnorm	mi
2	**bleiben**	unberührt	xcomp	188	141	88	18530.22	2.72511141462052
3	**erklären**	einverstanden	advmod	180	98	56	818.80	6.71948638487674
4	**gelten**	sinngemäß	xcomp	494	245	121	25479.05	1.35212883962229
5	**gekuppelt**	längsseits	advmod	43	54	41	599.48	8.45694002638426
6	**erklären**	vollstreckbar	advmod	180	59	38	555.61	6.94139139935541
7	**handeln**	Ordnungswidrig	advmod	116	43	28	409.40	7.54924385398649
8	**machen**	erforderlich	xcomp	172	142	42	8843.97	2.34069652117894
9	**führen**	Nachweise	advmod	255	84	21	307.05	5.63110850121035
10	**werden**	fällig	advmod	51	128	20	292.43	7.02428866262961
11	**vermuten**	widerleglich	advmod	31	19	17	248.56	9.39413639196275
12	**entscheiden**	abschließend	advmod	216	132	17	248.56	5.06566922900978
13	**rechtfertigen**	sachlich	advmod	31	73	14	204.70	7.93411938275508
14	**erfüllen**	ordnungsgemäß	advmod	260	149	14	204.70	4.42215573774188
15	**regeln**	abschließend	advmod	100	132	13	190.08	5.76191271858685
16	**bleiben**	unberücksichtigt	xcomp	188	16	13	2737.42	3.40481507269293
17	**durchführen**	ordnungsgemäß	advmod	243	149	10	146.21	4.08761261778458

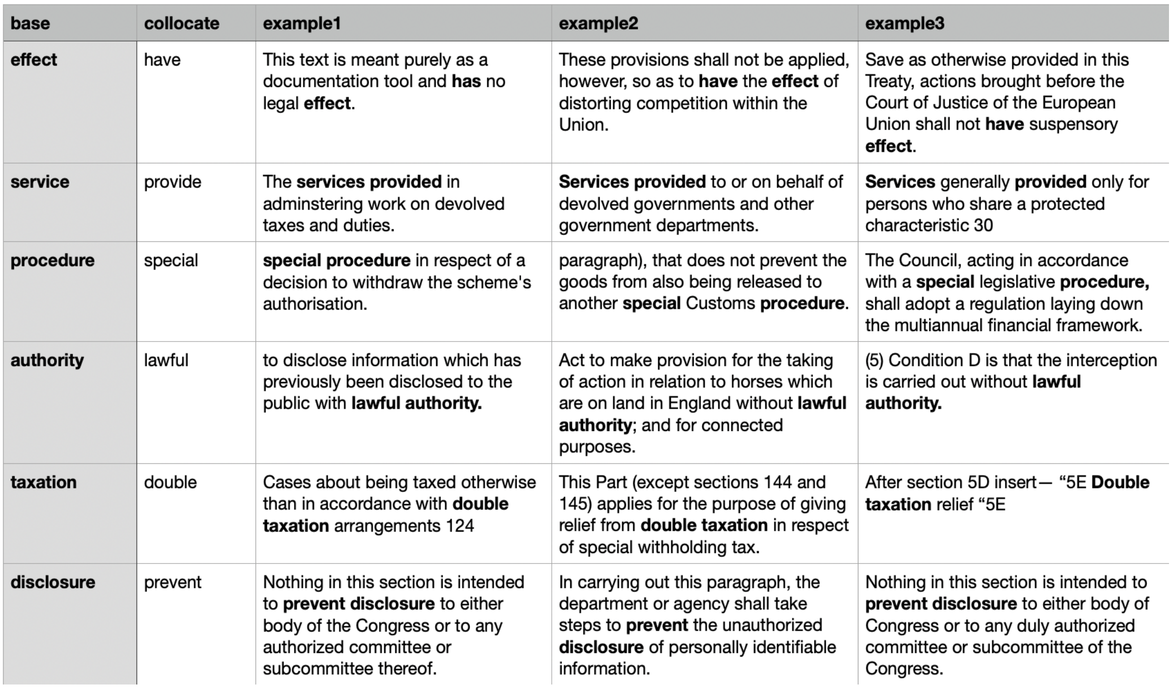
Figure 6. Extracted German collocation candidates for verb as a base sorted by t-score

In line with Orenha-Ottaiano et al. (2021), the automatically generated collocation candidates should be reviewed in the article writing process by lexicographers, thus not all collocation candidates will appear in the final dictionary articles.

5.2.3 Extraction of examples

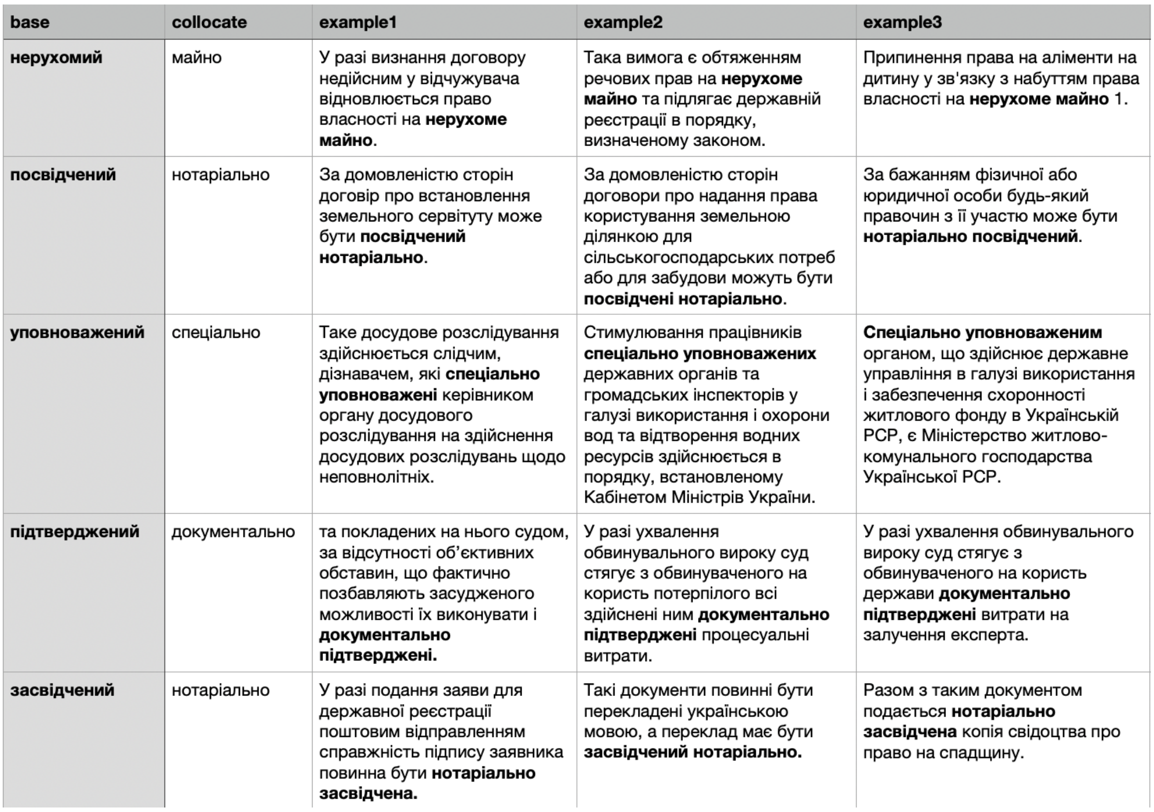
Examples are an indispensable part of dictionary articles, with the examples that come from corpora contrary to the ones created by lexicographers being especially important since they show the lexical unit in its natural context. Frankenberg-Garcia (2012, 2014, as cited in a Kosem et al., 2019, p. 119) study revealed that several examples taken from corpora can sometimes be even more beneficial to users than the definition. Selecting examples manually from corpora would be challenging for lexicographers due to the large size of corpora and the need to follow certain criteria of a good dictionary example. Hence, there has been extensive research into language technologies that look for good dictionary examples, such as the GDEX tool by Kilgarriff et al. (2008, as cited ibid)

Thus, for *EDU-Col*, for each collocation candidate, usage examples were extracted automatically following a set of GDEX-inspired heuristics (Kilgarriff et al., 2008, as cited in Orenha-Ottaiano et al., 2021, p. 6). In particular, at first, example sentences with assigned GDEX values for each base were generated from the corpus. The results were then used together with a list of collocations to get 10 GDEX example sentences for each collocation.



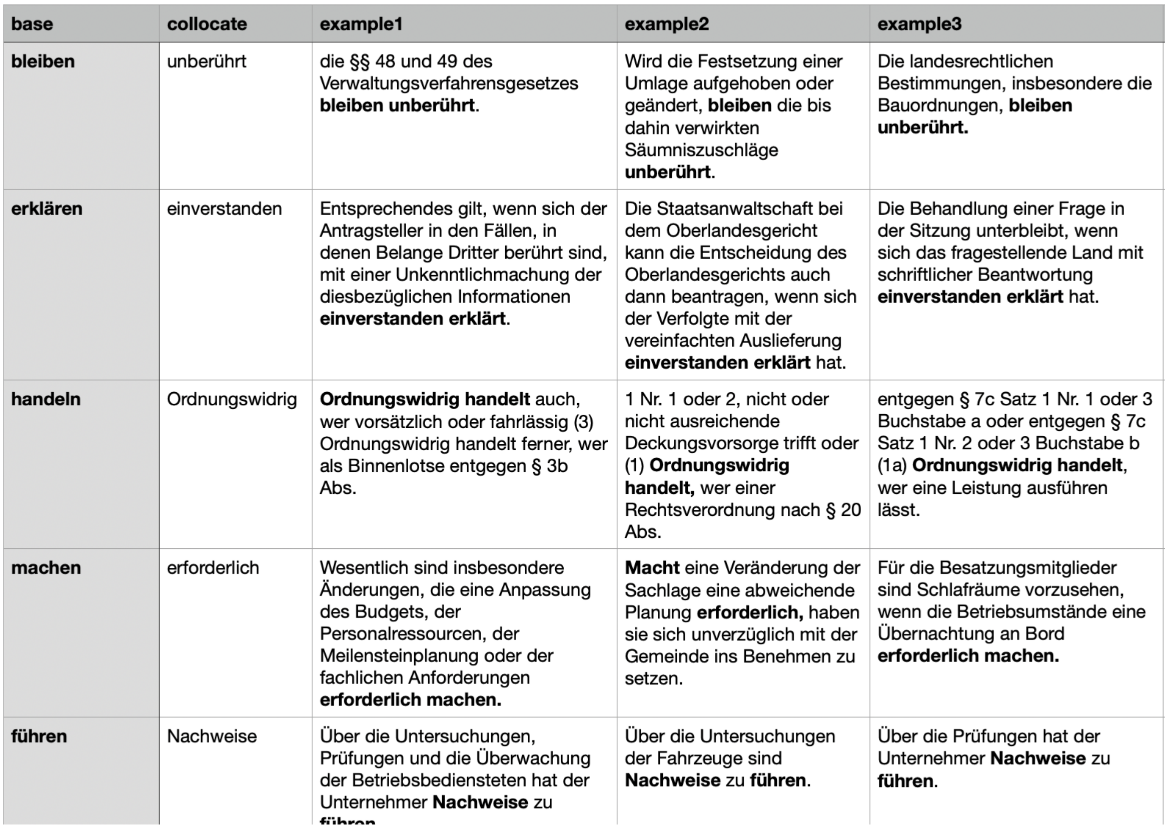
base	collocate	example1	example2	example3
effect	have	This text is meant purely as a documentation tool and **has** no legal **effect**.	These provisions shall not be applied, however, so as to **have** the **effect** of distorting competition within the Union.	Save as otherwise provided in this Treaty, actions brought before the Court of Justice of the European Union shall not **have** suspensory **effect**.
service	provide	The **services provided** in adminstering work on devolved taxes and duties.	**Services provided** to or on behalf of devolved governments and other government departments.	**Services** generally **provided** only for persons who share a protected characteristic 30
procedure	special	**special procedure** in respect of a decision to withdraw the scheme's authorisation.	paragraph), that does not prevent the goods from also being released to another **special** Customs **procedure**.	The Council, acting in accordance with a **special** legislative **procedure,** shall adopt a regulation laying down the multiannual financial framework.
authority	lawful	to disclose information which has previously been disclosed to the public with **lawful authority.**	Act to make provision for the taking of action in relation to horses which are on land in England without **lawful authority**; and for connected purposes.	(5) Condition D is that the interception is carried out without **lawful authority.**
taxation	double	Cases about being taxed otherwise than in accordance with **double taxation** arrangements 124	This Part (except sections 144 and 145) applies for the purpose of giving relief from **double taxation** in respect of special withholding tax.	After section 5D insert— "5E **Double taxation** relief "5E
disclosure	prevent	Nothing in this section is intended to **prevent disclosure** to either body of the Congress or to any authorized committee or subcommittee thereof.	In carrying out this paragraph, the department or agency shall take steps to **prevent** the unauthorized **disclosure** of personally identifiable information.	Nothing in this section is intended to **prevent disclosure** to either body of Congress or to any duly authorized committee or subcommittee of the Congress.

Figure 7. Automatically extracted examples for English collocation candidates (the collocation is marked in bold)



base	collocate	example1	example2	example3
нерухомий	майно	У разі визнання договору недійсним у відчужувача відновлюється право власності на **нерухоме майно**.	Така вимога є обтяженням речових прав на **нерухоме майно** та підлягає державній реєстрації в порядку, визначеному законом.	Припинення права на аліменти на дитину у зв'язку з набуттям права власності на **нерухоме майно** 1.
посвідчений	нотаріально	За домовленістю сторін договір про встановлення земельного сервітуту може бути **посвідчений нотаріально**.	За домовленістю сторін договори про надання права користування земельною ділянкою для сільськогосподарських потреб або для забудови можуть бути **посвідчені нотаріально**.	За бажанням фізичної або юридичної особи будь-який правочин з її участю може бути **нотаріально посвідчений**.
уповноважений	спеціально	Таке досудове розслідування здійснюється слідчим, дізнавачем, які **спеціально уповноважені** керівником органу досудового розслідування на здійснення досудових розслідувань щодо неповнолітніх.	Стимулювання працівників **спеціально уповноважених** державних органів та громадських інспекторів у галузі використання і охорони вод та відтворення водних ресурсів здійснюється в порядку, встановленому Кабінетом Міністрів України.	**Спеціально уповноваженим** органом, що здійснює державне управління в галузі використання і забезпечення схоронності житлового фонду в Українській РСР, є Міністерство житлово-комунального господарства Української РСР.
підтверджений	документально	та покладених на нього судом, за відсутності об'єктивних обставин, що фактично позбавляють засудженого можливості їх виконувати і **документально підтверджені.**	У разі ухвалення обвинувального вироку суд стягує з обвинуваченого на користь потерпілого всі здійснені ним **документально підтверджені** процесуальні витрати.	У разі ухвалення обвинувального вироку суд стягує з обвинуваченого на користь держави **документально підтверджені** витрати на залучення експерта.
засвідчений	нотаріально	У разі подання заяви для державної реєстрації поштовим відправленням справжність підпису заявника повинна бути **нотаріально засвідчена.**	Такі документи повинні бути перекладені українською мовою, а переклад має бути **засвідчений нотаріально.**	Разом з таким документом подається **нотаріально засвідчена** копія свідоцтва про право на спадщину.

Figure 8. Automatically extracted examples for Ukrainian collocation candidates (the collocation is marked in bold)



base	collocate	example1	example2	example3
bleiben	unberührt	die §§ 48 und 49 des Verwaltungsverfahrensgesetzes **bleiben unberührt**.	Wird die Festsetzung einer Umlage aufgehoben oder geändert, **bleiben** die bis dahin verwirkten Säumniszuschläge **unberührt**.	Die landesrechtlichen Bestimmungen, insbesondere die Bauordnungen, **bleiben unberührt.**
erklären	einverstanden	Entsprechendes gilt, wenn sich der Antragsteller in den Fällen, in denen Belange Dritter berührt sind, mit einer Unkenntlichmachung der diesbezüglichen Informationen **einverstanden erklärt**.	Die Staatsanwaltschaft bei dem Oberlandesgericht kann die Entscheidung des Oberlandesgerichts auch dann beantragen, wenn sich der Verfolgte mit der vereinfachten Auslieferung **einverstanden erklärt** hat.	Die Behandlung einer Frage in der Sitzung unterbleibt, wenn sich das fragestellende Land mit schriftlicher Beantwortung **einverstanden erklärt** hat.
handeln	Ordnungswidrig	**Ordnungswidrig handelt** auch, wer vorsätzlich oder fahrlässig (3) Ordnungswidrig handelt ferner, wer als Binnenlotse entgegen § 3b Abs.	1 Nr. 1 oder 2, nicht oder nicht ausreichende Deckungsvorsorge trifft oder (1) **Ordnungswidrig handelt,** wer einer Rechtsverordnung nach § 20 Abs.	entgegen § 7c Satz 1 Nr. 1 oder 3 Buchstabe a oder entgegen § 7c Satz 1 Nr. 2 oder 3 Buchstabe b (1a) **Ordnungswidrig handelt**, wer eine Leistung ausführen lässt.
machen	erforderlich	Wesentlich sind insbesondere Änderungen, die eine Anpassung des Budgets, der Personalressourcen, der Meilensteinplanung oder der fachlichen Anforderungen **erforderlich machen.**	**Macht** eine Veränderung der Sachlage eine abweichende Planung **erforderlich,** haben sie sich unverzüglich mit der Gemeinde ins Benehmen zu setzen.	Für die Besatzungsmitglieder sind Schlafräume vorzusehen, wenn die Betriebsumstände eine Übernachtung an Bord **erforderlich machen.**
führen	Nachweise	Über die Untersuchungen, Prüfungen und die Überwachung der Betriebsbediensteten hat der Unternehmer **Nachweise** zu **führen**	Über die Untersuchungen der Fahrzeuge sind **Nachweise** zu **führen**.	Über die Prüfungen hat der Unternehmer **Nachweise** zu **führen**.

Figure 9. Automatically extracted examples for German collocation candidates (the collocation is marked in bold)

Following the proposed heuristics by Kosem et al., (2019, as cited in Orenha-Ottaiano et al., 2021, p. 7), the software used for extracting examples rejects sentences with few tokens, as well as those containing more than 30 tokens. In addition, sentences with proper nouns, words containing more than 12 characters, and characters not conforming to the main alphabets are also left out. The illustration of the extracted examples for candidate collocations with the noun, adverb, and adjective bases can be seen above further for the three languages in Figures 7, 8, 9.

5.2.4 Generation of translation equivalents

Acquiring translation candidates for *EDU-Col* is performed following Garcia et al. (2019c), i.e, the method employing cross-lingual models of distributional semantics, also known as word embeddings, and an unsupervised approach. Distributional semantics is associated with the famous words by Firth (1957, p. 11), “you shall know a word by the company it keeps,” and is based on the Distributional Hypothesis, according to which, as Boleda (2020, p. 2) summarizes, “similarity in meaning results in similarity of linguistic distribution”. The scholar (Boleda, 2020, ibid) further gives an example of the semantically related words *post-doc* and *student* that are used in similar contexts (*a poor _ , the _ struggled through the deadline*). In its most common form, distributional semantics represents word meaning with semantic representations in the form of vectors, i.e, "lists of numbers that determine points in a multi-dimensional space" Boleda (2020, p. 2) .

Cross-lingual models of distributional semantics can be obtained using two different methods, namely training bilingual models using parallel data or using monolingual models and then mapping them into a shared vector space. As Garcia et al., (2019a, p. 752) points out, the second approach allows to obtain high-quality cross-lingual word embeddings without the need to rely on parallel data. Although the quality of word embeddings is lower than using parallel corpora, comparable corpora are much easier to find; therefore, we opted for this approach, mapping the monolingual models in a fully unsupervised way.

Garcia et al., (2019a, p. 752) further elaborates that monolingual word embeddings represent words as *n*-dimensional vectors, with words that occur in similar contexts having similar vectors, whereas cross-lingual word embeddings represent words of different languages in the same vector space, which enables the computation of distributional similarities between those languages (Rapp, 1999; Ruder et al., 2019 as cited in Garcia et al., 2019a, p. 752). An illustration of monolingual word embeddings as compared to cross-lingual word embeddings in a shared vector space can be seen in Figure 10 (a, b).

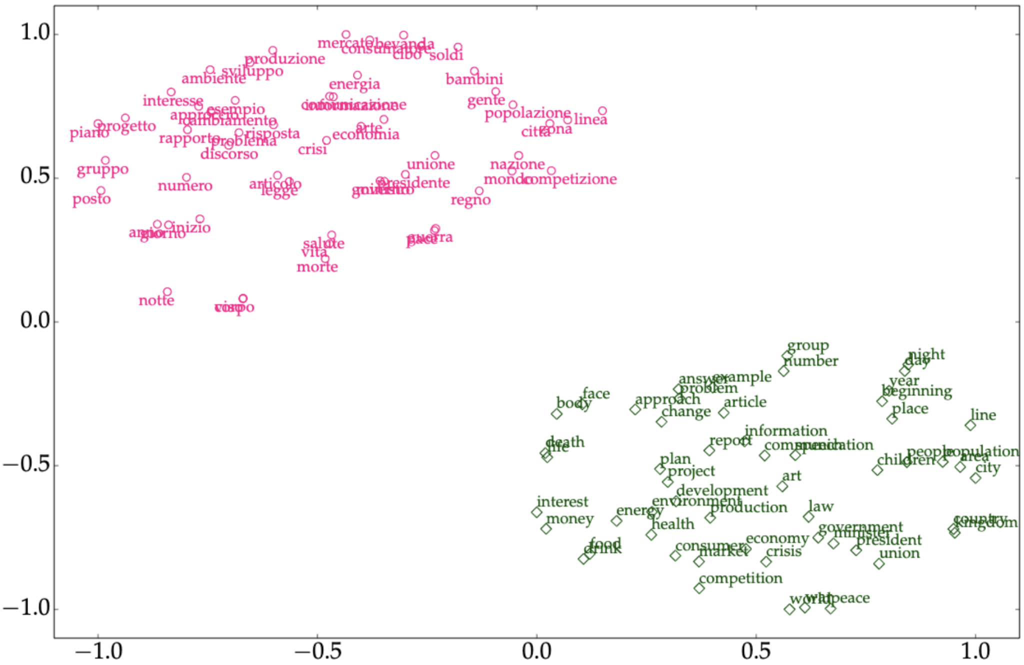


Figure 10 (a). Unaligned monolingual word embeddings (in Ruder et al., 2019, p. 570).

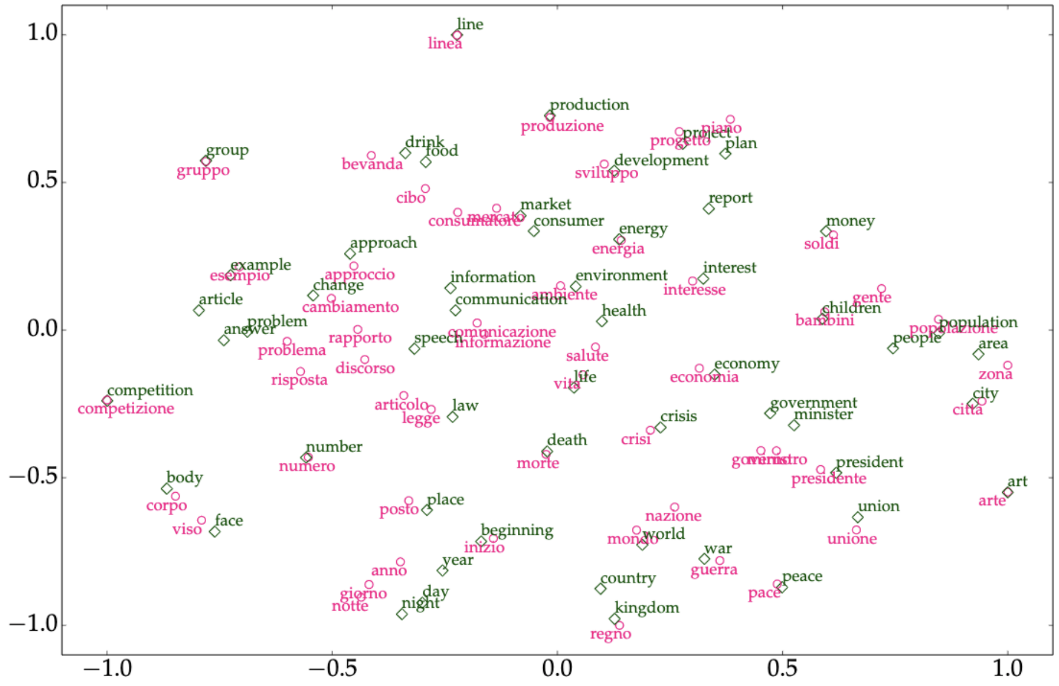


Figure 10 (b). Word embeddings projected into a joint cross-lingual embedding space (in Ruder et al., 2019, p. 570).

Following Garcia et al. (2019c, p. 8), to obtain cross-lingual word embeddings, at first, corpora need to be converted into lemma PoS-tag corpora. The lemma PoS-tag corpora are then utilized to learn monolingual models using word2vec (the skip-gram algorithm, 300 dimensions, a window of 5 tokens, and a frequency threshold of 5). The final step involves mapping the monolingual models into a shared vector space (Artetxe et al., 2018, as cited in Orenha-Ottaiano et al. 2021, p. 8). For our project, we used the models mapped with the fully unsupervised approach., i.e., without any bilingual dictionaries or lists of translated words to generate the files with candidate translation in target languages from the input lists in source languages for each pair of three languages. The lists of dependencies in target languages were used in this procedure to reduce the number of infrequent and unattested combinations.

Garcia et al., (2019c) propose the "weighted approach" to the translation of collocations that accounts for non-congruent collocations, unlike those strategies that consider the semantic load of each multiword expression component as similar and obtain only word-for-word translations. Non-congruent collocations are the collocations "where the meaning of one of their components is not a direct translation in the target language" (as cited in Garcia et al., 2019c, p. 2), for instance, *take a photo* in English vs. *ein Foto machen* in German (where the German *machen "make"* is not the word-for-word translation of *take*).

The internal procedures of the "weighted approach" can be summarized as follows (Garcia et al., 2019c, p. 6): both base and collocate candidates in the target languages are searched for using cross-lingual word embeddings. For the bases, the source vector is used to obtain the n most similar words, which have the same PoS-tag, whereas candidate collocates are searched for using the weighted compositional vector, which allows finding collocate candidates whose meaning is closer to one of the whole combinations than to the collocate alone. To acquire the translated collocation candidates, the software puts together the base and collocate candidates. For each of the collocation candidates, translation confidence (in terms of cosine similarity) is computed using the weighted average vector. The confidence value helps to rank the target collocations regarding source ones in terms of quality.

The approach to finding multilingual equivalents takes into account semantic properties of collocations, namely the fact that bases tend to have a stable meaning, whereas collocate's meaning is rather unstable and may change depending on the combination it is used in. The candidates should have some degree of collocability or have sufficient frequency rate before theu can be accepted as resulting collocations (Garcia et al., 2019a, p. 751).

It should be noted that at this stage, manual validation is also required to check the output translations to ensure the best quality (Orenha-Ottaiano et al., 2021). As can be seen from the Table 5, which contains the Ukrainian translation equivalents of the English collocations, the translation candidate with the highest translation confidence score is not always the appropriate equivalent (the correct equivalents are marked in bold). Automatic translation using the presented method might at times produce errors, with Garcia et al., (2019c, p. 11) giving an account of the most typical error types. For instance, the first collocation in Table 5 *have [an] effect* did not get an acceptable equivalent, with the first two candidates translated using an antonym "do not have" into Ukrainian. Such an error might appear due to the distributional method, which represents antonyms with very similar vectors.

Collocation	**Candidate equivalents**
effect, have,obj	чинник,немати,0.928050; наслідок,немати,0.927425; наслідок,зважати,0.924841; чинник,зважати,0.919831; вплив,зважати,0.914635; наслідок,враховувати,0.906281; наслідок,незважаючи,0.905400;вплив,незважаючи,0.898415; чинник,незважаючи,0.893812; наслідок,бути,0.892243;

board, member, compound	голова,член,0.902166;ради,представник,0.872777; **ради,член,0.872088;**Комітет,представник,0.866286; голова,представник,0.861981;голова,членкиня,0.859971; член,член,0.858062;член,представник,0.857976; комісія,представник,0.857005;керівнико,член,0.855447;
service, provide, nsubj	**послуга,надавати,0.906074;**забезпечення,надавати,0.904858;послуга,забезпечити,0.901774;агентування,забезпечити,0.900914;послуга,надаватися,0.899568;забезпечення,забезпечити,0.898403;забезпечення,надаватися,0.893204;послуга,забезпечувати,0.891758;забезпечення,надати,0.890530;агентування,надавати,0.890468;
conviction, previous, amod	кривосвідчення,попередній,0.912077;звинувачення,попередній,0.900204;обвинувачений,попередній,0.895970;кривосвідчення,наступний,0.883818;звинувачення,останній,0.883554; звинувачення,наступний,0.883187;обвинувачення,попередній,0.877276;злочин,попередній,0.874781**;вирок,попередній, 0.872515;**вирка,попередній,0.870599;

Table 5. Collocations and their automatically generated candidate equivalents

Another drawback of the method is that it cannot extract collocation equivalents of different syntactic structures, thus, for instance, an adjective-noun collocation will get only adjective-noun collocations as candidate equivalents, although other structures might be acceptable, for instance.

Moreover, although the "weighted approach" is supposed to perform well for non-congruent collocations, other simpler cases of congruent collocations should be accounted for too. In Garcia et al., (2019c, p. 7), the proposed "weighted method" was compared to a number of other methods using the same cross-lingual models to select the translation candidates. Among the methods there were the following compositional ones (Garcia et al., 2019c, p. 7):

(a) "baseline", which "creates a collocation selecting the most similar equivalents of both the base and the collocate in the target language";

(b) “addition”, which “generates at most 100 collocation candidates from the top 10 bases and collocates in the target language, and ranks them by cosine similarity of the source and target compositional vectors $v = b + c$”;

(c) “multiplication”, the “same as “addition” but obtaining the collocation vectors by multiplication instead of addition ($v = b \cdot c$)”.

Out of the above methods, the “addition method” yielded the best results; therefore, we employed it as a second working method to extract collocation candidates. Although in Garcia et al., (2019c) the “weighted approach” produced the best results as compared to other compositional methods, on our data, the “addition method” was slightly better, especially for congruent collocations, i.e., probably due to the fact that it obtains the translation of each word individually, thus being more powerful for such type of collocations. The results of a small test are elaborated on in more detail in section 7.1 Results. Thus, for our dictionary project, we will resort to both approaches.

6. Dictionary structure and presentation of model articles

This section will reflect on some of the aspects of lexicographic structures relevant for our dictionary project, with the regard to the dictionary type, the intended target users, and specific functions of the dictionary. Some of these structures emerge only in online dictionaries, whereas others have been discussed by metalexicographers solely in relation to printed dictionaries; however, in an adapted form they are also used in online dictionaries.

6.1 Macrostructure

Wiegand & Gouws (2013, p. 75) give the following definition of a printed dictionary’s macrostructure: “The *macrostructure* of a printed dictionary is that textual structure that presents the ordering of all those elements of the data memories that contribute to the dictionary type specific macrostructural coverage”. It should be noted that in online dictionaries, on the contrary, the organization of the central word list or a head word list according to the alphabetical ordering or subject matter is evidently no longer the deciding factor for the form of the dictionary in which it will be published. However, it is still worth considering what elements will eventually enter the head word list and defining certain criteria for delimiting and organizing those elements. With this in mind, this section will present the criteria for building the candidate head word list for our dictionary project.

6.1.1 The head word list

Since *EDU-Col* is planned to cover collocations in three languages, there is a need to create separate candidate head word lists for English, German, and Ukrainian. In addition, we aim to have additional candidate head word lists for the domain-specific collocations, such as law, economics, environment etc.

6.1.1.1 Collocations types

As it was touched upon in the previous sections, the types of collocations to be represented in the headword lists for English, German, and Ukrainian include for noun bases:

- verb-noun
- noun-verb
- adjective-noun
- verb-preposition-noun
- noun-preposition-noun
- noun-noun.

For verb bases, the collocation types are in turn the following:

- verb-adjective
- verb-adverb

Finally, adjective collocation bases will be represented by the adjective-adverb collocation type.

6.1.1.2 Organization of headwords

Organizing collocations into headwords in a dictionary has certain implications for users and the dictionary consultation procedure accordingly. Hollós (2008, p. 125) points out that in large German learner's dictionaries by Langenscheidt, de Gruyter, Pons or Duden, most collocations are listed under the collocate, i.e. under the element, which is usually unknown to users. On the contrary, Hausmann, (1984) advocates for the opposite treatment of collocations' elements, arguing that the regular monolingual dictionaries are not sufficient in the production situations as the users look first for the element they know, which according to this view is a base and not a collocate (Hausmann, 1984, as cited in Hollós, 2008, p. 125).

Considering the functions of our dictionary project, which aims to assist users primarily in production usage situations, we resorted to the approach that advocates for the base of the collocation to be regarded as a headword. Hence the collocations in *EDU-Col* will be listed under the part which is supposedly known to the users, thus contributing to the higher probability that the users will find what they are looking for. In addition, another solution is to provide users with a functionality that allows them to search not only for bases but also for specific collocates.

6.1.1.3 Treatment of certain types of collocation bases

6.1.1.3.1 Homonymy vs. polysemy

Atkins & Rundell (2008, p. 281-282) provide an account of the most common practices pertaining to the differentiation between homonymy and polysemy, which can be summarised as follows:

- Homographs (i.e., words that differ in sounding but have the same form) and capitalized forms in most dictionaries are treated as separate dictionary articles, e.g., the English *bow* (baʊ) and *bow* (boʊ), and for *may* and *May* would get separate articles;
- Differences based on word class are handled differently depending on the dictionary, with the most common solutions being to create a separate dictionary article with subsections for each word class and, conversely, separate articles for each word class of a word;
- In general, homonymy is becoming less common in many types of dictionaries.

Trying to answer the question whether homonymy is still relevant for lexicography, Atkins & Rundell (2008, p. 281) emphasize:

> “The answer, as always, depends on the intended uses (and target users) of the dictionary. In historical dictionaries, homonymous words always appear as separate entries: describing words’ origins and development is central to the function of dictionaries of this type. But the value of homonymy to a synchronic account of meaning is far less clear”.

On the other hand, the connections between the word forms or absence of them might confuse the users, as Atkins & Rundell (2008, p. 282) recapitulate, “[ ] a rigorous application of homonymy could well cause look-up problems, too.” In particular, such criteria as distinct etymological origins, which are used to identify if a word is a homonym or a polysemous lexical unit, are in most cases not known to the users unless they have a background in linguistics; hence, it might impede their dictionary consultation procedure. For instance, in *COBUILD*'s policy, it is stated, "Because access to an item is through its orthographic form, and because etymological homonymy depends on knowledge that is not available to the dictionary user before he or she locates the word in the dictionary, it was decided to ignore homonymy completely" (Atkins & Rundell, 2008, ibid). Therefore, homonymy should not be the defining criteria for the macrostructural decisions related to the headword list.

For *EDU-Col* we adopted an approach that favors the word class, thus a dictionary article will include all the different senses of a headword, in our case the collocation base, irrespective of the nature of the differences between senses. Instead, these differences will be dealt with on the microstructural level.

6.1.1.3.2 Variant forms

Following Atkins & Rundell (2008, p. 180), another macrostructural decision pertaining to the headword list is deciding on the lexical form of words that will be included in it. In particular, the variation in the spelling of words is of relevance here. For instance, such variant forms and spellings as *aluminium,* (British English) and *aluminum* (American English), *ageing* vs. *aging*, *harbour* vs. *harbor*, or *analogue* vs. *analog,* etc. should be taken into account. In EDU-Col, separate dictionary articles can be developed for the variant forms, accounting for the differences in collocations they might have in British and American English.

Since our dictionary project includes English as one of the languages, and the corpora we used for our experiment includes texts representing both the US and UK varieties of English, it is vital to consider such an issue as variant forms and spelling of words for the headword list building and subsequently microstructure of the dictionary articles. On the contrary, the variant forms are not characteristic of the standard German and Ukrainian languages, thus such consideration in our case is relevant only for English.

6.2 Microstructure

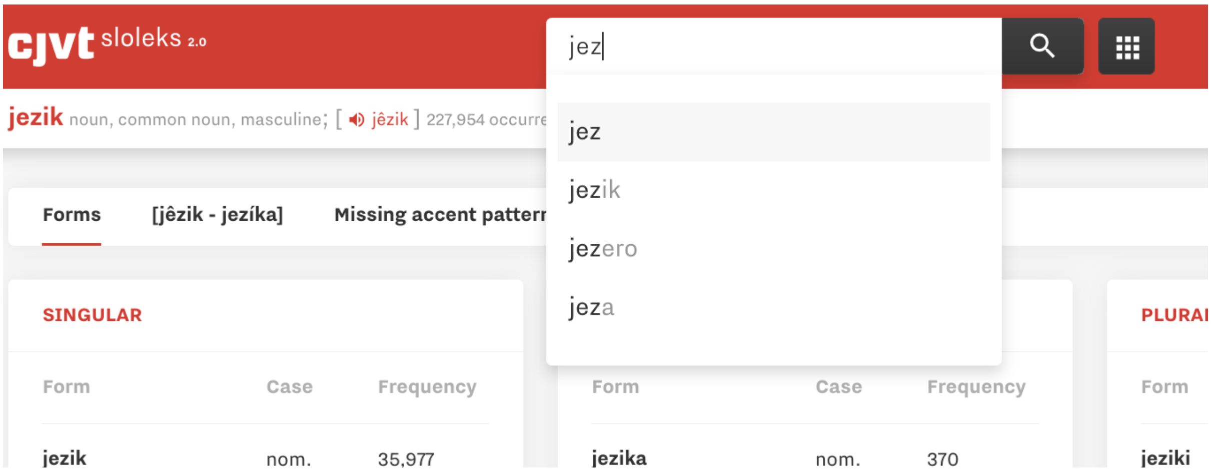
Following Hausmann et al., (1989, as cited in Gouws & Prinsloo, 2005 p. 119-126), the microstructure of a dictionary article is constituted by the "*comment on form* and the *comment on semantics",* where the former is "the search field accommodating those data types that reflect on the form of the lemma sign, i.e. the morphological, phonetic and orthographic form", whereas the latter refers to those data types that "describe the semantic and pragmatic features of the lexical item represented by the lemma". Comment on semantics is typically represented by a variety of data types, depending on the type of dictionary, the dictionary target user, and the usage situations, such as the meaning paraphrase presented by means of lexicographic definitions, examples, translation equivalents in bilingual dictionaries, pragmatics labels, etc.

Further, we will elaborate on a number of considerations relevant to the microstructure of *EDU-Col*'s articles, considering the dictionary type and its potential users.

6.2.1 Comment on form

6.2.1.1 Orthographic information

The headword represented by a lemma sign belongs to the comment on form as it indicates the spelling. As noted by Gouws & Prinsloo (2005 p. 119), "[users] often need orthographic guidance and their dictionary consultation procedure only goes as far as finding the lemma and retrieving the necessary spelling information from the lemma sign". Although our dictionary project is not a general language dictionary, with the potential users being interested primarily in finding collocations, a possible decision is to include additional spelling information for different spelling variants of lemmata. The solution for the issue with variant forms for the English language, which was discussed in section 6.2 Microstructure, could be to give a variant form as, for instance, in **favor** (UK **favour**) in the dictionary article. In turn, the issue with searching for either of the forms can be solved by the suggestion functionality, which will give users variants of words once they start typing in the search field as is, for instance, implemented in the Collocations Dictionary of Modern Slovene (Figure 11) by Kosem et al., (2019).



jezik noun, common noun, masculine; [jêzik] 227,954 occurre

SINGULAR

Form	Case	Frequency
jezik	nom.	35,977

Form	Case	Frequency
jezika	nom.	370

PLURAI

Form
jeziki

Figure 11. Search functionality in the Collocations Dictionary of Modern Slovene (Kosem et al., 2019)

6.2.1.2 Morphological information

In *EDU-Col*, data on the morphology of the lemma and grammatical features is not an obligatory element of microstructure considering the type of the dictionary; therefore, for instance, for a lemma representing a noun, the comment on form will not include morphological data such as the diminutive forms or for adjective the comparative and superlative degrees. However, such morphological information as the part of speech is relevant for our project since it will not only provide the users with information but will also serve as a guiding element on finding the right dictionary article, for instance, for lemmata such as *plan* (verb) and *plan* (noun). In addition, considering the nature of the German and Ukrainian languages, morphological information on grammatical gender will be provided for noun bases, indicating if a noun is masculine, feminine or neutral, hence assisting users in production situations.

6.2.1.3 Phonetic information

Similarly to morphological information, information regarding the pronunciation of words is not an obligatory element of microstructure in *EDU-Col.* However, by adding the guidance on pronunciation at least for the headword, we can accommodate the language learners who are one of the potential target user groups. Among various ways of the representation of the sound form in lexicography, such as transcription, International Phonetic Alphabet (IPA), and recorded audio files that enable the user to listen to the pronunciation or a partial transcription where the only the main stress is indicated (Gouws & Prinsloo, 2005, p. 120), we will resort to the audio guidance and IPA for German and Ukrainian headwords, also accounting for the UK and US varieties in the case of English. The phonetic information is planned to be generated using automatic tools, followed by a manual review performed by lexicographers. Due to the scope of research, phonetic information generation is not covered in this thesis.

6.2.1.4 Collocation patterns and collocations

Collocations are an obligatory element in the microstructure of the dictionary articles in *EDU-Col,* and the collocation patterns will be provided to help the users seamlessly navigate throughout the article. In order not to confuse the users, the patterns will represent only the part of the speech combination, whereas the dependency relations will not be marked.

6.2.2 Comment on semantics

6.2.2.1 Meaning

In general language dictionaries, the information on meaning is one of the most important data types, which is represented by meaning paraphrase or a lexicographic definition. Here, the ordering of the senses should be considered, which is normally motivated by the type of dictionary. As Gouws & Prinsloo (2005, p. 120) point out, "A dictionary based on historical principles will typically order the senses from the oldest to the youngest. In general, in synchronic dictionaries, one usually finds the ordering determined by the usage frequency of the senses. The sense with the highest usage frequency will be given as the first sense."

In *EDU-Col*, the senses of the polysemous lemma, which is a collocate base in our case, will be ordered according to its frequency in the corpus, and numbers will be used as polysemy markers to visually enhance navigating the information in the dictionary article.

Discussing the comment on semantics as a constituent part of the dictionary article's microstructure, Gouws & Prinsloo (2005, p. 127) also mention such information type as *context and cotext entries* (entries in their terms refer not to the dictionary entry, which they call a dictionary article, but to a part of the dictionary article), which are particularly important for dictionaries with the text production function. The *context entry* refers to the data on the typical pragmatic environment of a lemma, and is often indicated by means of glosses (e.g., (of a person) to indicate that a certain adjective can be used to describe humans). Such an information type can be also used in *EDU-Col* where it is relevant.

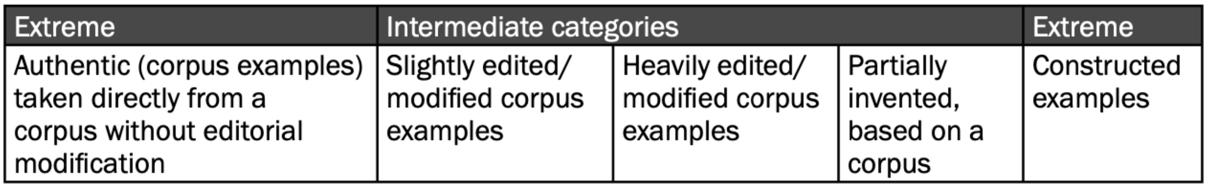
Another solution is using lexicographic labels, which we will discuss further. The *cotext,* in turn pertains to the typical syntactic environment, i.e., examples, collocations etc.

6.2.2.2 Examples

Examples are a very important microstructural element in a dictionary article, with lexicography scholars (Prinsloo & Gouws, 2000, p. 144-145) naming the following functions of examples, which should help users to:

- “disambiguate senses;
- distinguish one meaning from another, clarify an abstract definition;
- supplement the information in a definition;
- show or indicate the selectional range;
- place the word in context;
- place the word in context;
- specify the semantic range;
- indicate the collocational behaviour, including typical collocations,
- illustrate the grammatical patterns;
- specify the word order;
- give pragmatic uses;
- note stylistic features, indicate appropriate registers, reflect the word history;
- be accurate, especially those quoting measurements, technical data, etc., and
- stimulate the users to capture the features or characteristics of the word in question and use the examples as a model to create examples of their own”.

On the scale of authentic examples taken from the corpus versus ‘made-up’ examples of usage, lexicographers single out the following categories that are available to the lexicographer:



Extreme	Intermediate categories			Extreme
Authentic (corpus examples) taken directly from a corpus without editorial modification	Slightly edited/ modified corpus examples	Heavily edited/ modified corpus examples	Partially invented, based on a corpus	Constructed examples

Figure 12. Authentic versus constructed examples (Gouws & Prinsloo, 2005, p. 35)

It should be noted that constructed examples have become much less frequently used, with such arguments in favor of authentic examples as the fact that they are grammatically correct, situationally appropriate, give accurate collocations, and represent real language, among other arguments (Prinsloo & Gouws, 2000, p.146-147).

In *EDU-Col,* automatically extracted examples from the corpora will accompany each collocation in order to place it in context and cotext, indicate the collocational behavior, illustrate the grammatical patterns (active, passive voice, use of prepositions, etc), and specify the word order. Where it is relevant, this slight modification will be employed to make them more readable.

6.2.2.3 Translation equivalents

Translation equivalents of the collocations is another obligatory microstructural item in our dictionary project. In *EDU-Col*, users will be able to choose to display equivalents in one or all the available target languages, depending on the source language they had set as a default.

Extracted as a result of the automatic procedure described in section *5.2.2 Extraction of collocation candidates*, candidate equivalents will be subsequently reviewed by lexicographers to ensure their validity. It should be noted that we will restrict the equivalents to the same syntactic patterns collocations represent. Thus, for instance, for the English adjective + noun collocation, there will be an equivalent with the same syntactic pattern adjective + noun in other languages, although other ways of expressing the meaning of a collocation can be found syntactically. In cases when as a result of the automatic generation procedure certain candidate equivalents are not meaningful, it will be the task of lexicographers to add the missing equivalents.

6.2.2.4 Lexicographic labels

Another data type that contributes to the comment on semantics of a dictionary article and is relevant for our dictionary project is lexicographic labels. Lexicographers single out at least three major classes of labels, i.e., subject field labels, stylistic labels, and chronolectic labels, where the first type of labels indicates a specialized field; stylistic labels in turn signify nonconformity to the standard variety of a language, *e.g., formal, colloquial, figurative, slang, etc.* Finally, chronolectic labels indicate the time of use of a word or one of its senses, for instance, an archaic form or neologism (Gouws & Prinsloo, 2005, p. 129-131).

Considering that collocations that belong both to the general lexicon and specific domains will be represented in *EDU-Col*, we aim to use subject field labels. In particular, they will be employed for macro- and microstructural items in the dictionary. Thus, the label will mark the lemma sign for the collocations extracted from specialized corpora such as, for instance, legal or economics domains in our case. However, if there are specific collocations extracted in the previous stage from specialized corpora that correspond to the bases extracted from the general language corpora, the lexicographic data will be integrated into one dictionary article. If the collocation candidates are more frequent in the subject field corpora as compared to its frequency in the general language corpora, then the label will be used in the microstructure of the dictionary article.

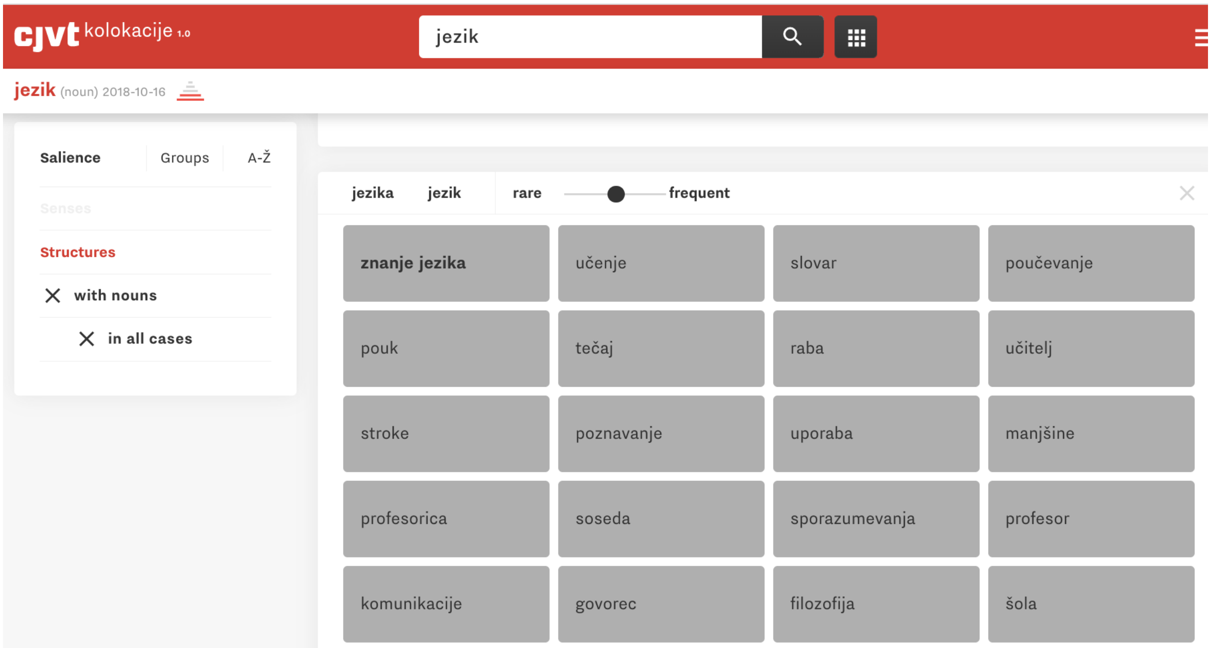
Another consideration would be to introduce sorting and filtering options in the final product, which would allow users to sort the collocations by domain or select only general language collocations, viewing only the data that is most relevant to the users.

6.2.3 Statistical data on collocations

As a separate element, which does not fall within the comment on form and comment on semantics in the microstructure of a dictionary article is statistical association measures relevant for collocations. Accounting for different types of potential users of our dictionary project, in terms of their expertise in corpora statistics, the association measures of collocations will not be displayed by default; instead, users will be able to adjust the amount of information they see on the screen.

As far as frequency of collocations is concerned, the visual solution, which was implemented in Kosem et al., (2019) can be used in our dictionary project. In the Collocations Dictionary of Modern Slovene, the numerical data was transformed into the frequency filter (Figure 13). The function of the filter[30] here is to allow users decide on a priority that is more important to them in a certain situation, thus choosing either frequent or rare collocates.

[30] With an aim to limit the data and facilitate finding the relevant information also other types of filters are used in the Collocations Dictionary of Modern Slovene (Kosem et al., 2019), such as the sense filter, the grammatical relation filter, which allows limiting the results based on the word class of the collocate and subcategories such as case or degree; the preposition filter.



znanje jezika	učenje	slovar	poučevanje
pouk	tečaj	raba	učitelj
stroke	poznavanje	uporaba	manjšine
profesorica	soseda	sporazumevanja	profesor
komunikacije	govorec	filozofija	šola

Figure 13. Frequency of collocations presented as a filter functionality in the Collocations Dictionary of Modern Slovene (Kosem et al., 2019)

In terms of other association measures relevant for collocations, as noted by Evert (2008, p. 31), out of the variety of association measures, several have been treated as de-facto standards, and in computational lexicography these are t-score and MI. Therefore, they will be offered for viewing to the users in the advanced mode.

6.3 Mediostructure

The mediostructure will be represented by a system of cross-referencing, connecting different components of the dictionary. If a word (token) used in the definition paraphrases corresponds to the lemma, an internal hyperlink will be provided to the corresponding dictionary article provided the word (token) is available in the dictionary as a collocation base. This way we will save the users' time during the dictionary consultation procedure and allow them to avoid the need to type the words in the search field or look them up in additional resources.

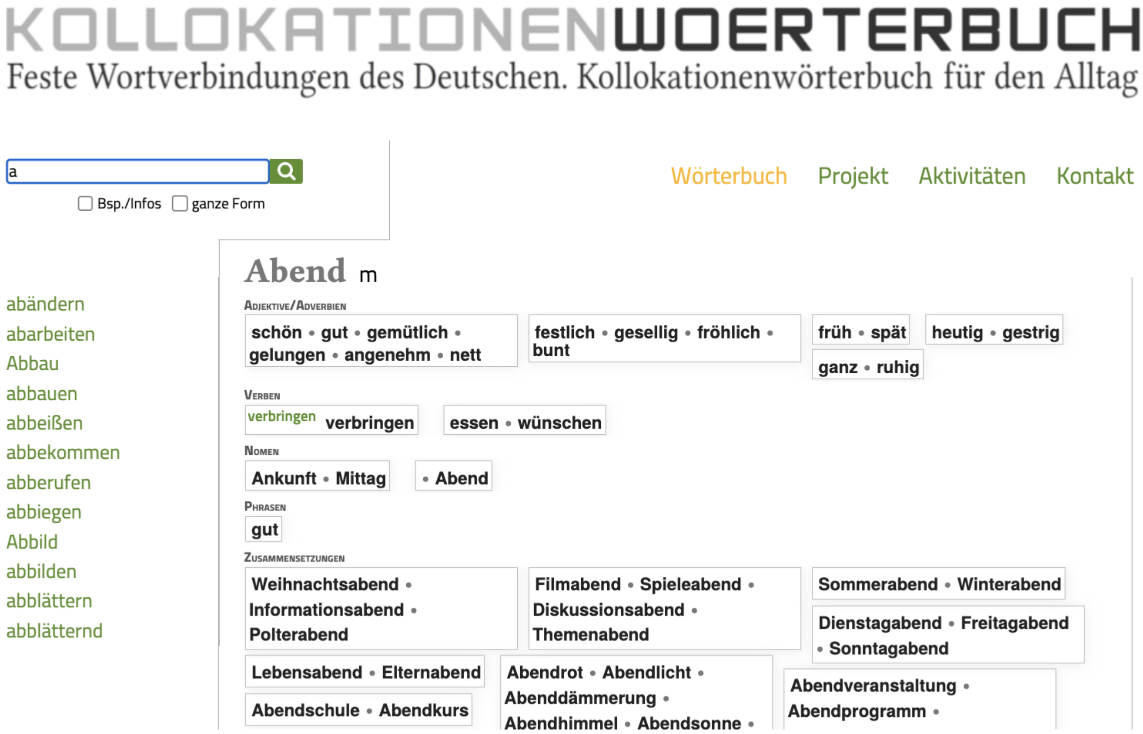
6.4 Access structure

Gouws (2010, p. 102) defines the access structure as "the search route dictionary users take during a dictionary consultation procedure", making a distinction between the outer access structure, i.e., the one that guides a user to the lemma sign and the dictionary article, and the inner access structure, which is used to guide the user to certain data types in a dictionary article. According to Klosa (2013, p. 518), in terms of dictionary outer access structure, online dictionaries can be classified as follows:

- Dictionary with access through scrolling in an article list;
- Dictionary with access via hyperlinked list of headwords;
- Dictionary with access via search options;
- Dictionary with combined access.

While in a printed dictionary access to the dictionary article is provided via the linearly ordered lemmata, in online dictionaries, the search functionality prevails. However, despite the many possibilities the online medium offers for advanced dictionary access and search, it should be noted that access structures that were originally developed for printed dictionaries can also be utilized in online dictionaries in order to maximize the speed and ease of a consultation procedure for users. For instance, the *Kollokationen-Wörterbuch*, the German collocation dictionary by Buhofer et al., (2014), after typing a certain letter or a word, allows users to view on the left panel the available dictionary articles that start with the corresponding letter of the alphabet (Figure 14).

As Gouws (2018, p. 230) notes, "[t]he occurrence of alphabet bars [ ] and partial article stretches [ ] gives the user who is used to an alphabetical way of searching for data in dictionaries a feeling of familiarity", which in turn allows for "a systematic access procedure."



KOLLOKATIONENWOERTERBUCH

Feste Wortverbindungen des Deutschen. Kollokationenwörterbuch für den Alltag

a

□ Bsp./Infos □ ganze Form

Wörterbuch Projekt Aktivitäten Kontakt

abändern
abarbeiten
Abbau
abbauen
abbeißen
abbekommen
abberufen
abbiegen
Abbild
abbilden
abblättern
abblätternd

Abend m

Adjektive/Adverbien

schön • gut • gemütlich • gelungen • angenehm • nett

festlich • gesellig • fröhlich • bunt

früh • spät

heutig • gestrig

ganz • ruhig

Verben

verbringen verbringen

essen • wünschen

Nomen

Ankunft • Mittag

• Abend

Phrasen

gut

Zusammensetzungen

Weihnachtsabend • Informationsabend • Polterabend

Filmabend • Spieleabend • Diskussionsabend • Themenabend

Sommerabend • Winterabend

Dienstagabend • Freitagabend • Sonntagabend

Lebensabend • Elternabend

Abendschule • Abendkurs

Abendrot • Abendlicht • Abenddämmerung • Abendhimmel • Abendsonne •

Abendveranstaltung • Abendprogramm •

Figure 14. Alphabetic lemmata bar in *Kollokationen-Wörterbuch* (Buhofer et al., 2014)

EDU-Col is planned to combine all these access options, including the alphabet bars and list of all lemmata, and search options. In addition to the simple search, advanced search options will allow the users to limit their search:

- Search in a list of domain-specific collocations;
- Search in a list of general language collocations;
- Search for collocations for a particular base;
- Search for collocations where a particular collocate is a part of.

6.5 Outer features

Adapting to online dictionaries the theory of outer texts as discussed in relation to the printed dictionaries, i.e. the elements of a frame structure that form either front or back matter, such as a foreword, user guidelines, etc, Klosa & Gouws (2015) substitute the notion of outer texts by a concept of outer features, which in online dictionaries are not always represented by texts, but instead by more interactive elements.

The role of outer texts in print and similarly online dictionaries should not be underestimated. According to the transtextual approach to lexicographic functions (Gouws, 2007, p. 82), the functions of a dictionary can be achieved not only by finding the information in the central list. Outer texts can also be employed for this purpose and especially function-adhering outer texts that, according to the scholar, can contribute significantly to the lexicographic functions of a certain dictionary. Although Gouws discusses printed dictionaries, the ideas can be also extrapolated to online dictionaries. In line with a variety of types of outer features outlined by Klosa & Gouws (2015, pp. 164-167), the following outer features could be potentially employed in *EDU-Col*:

- user guidelines explaining how to use the advanced search (possibly a video explanation);
- general information, which elaborates on data types present in the microstructure, what some statistical data denotes, etc;
- interactive games on collocations and teaching material targeted at one of the groups of potential users of the dictionary, namely learners of languages.
- pointers to dictionary content, i.e., elements, which can attract the users' attention to the dictionary content (e.g., "Most frequently searched collocations", "Latest submissions", etc).

Such outer features will not only enhance the consultation procedure but will also reinforce the functions of *EDU-Col* giving the users, especially the learners, not only the opportunity to search for collocations, but also provide them with educational and learning content.

6.6 Model dictionary articles

Taking into account the data types we discussed in the previous sections, the summary of the model dictionary articles for the three languages present in our dictionary can be made with the following scheme (Figure 15):

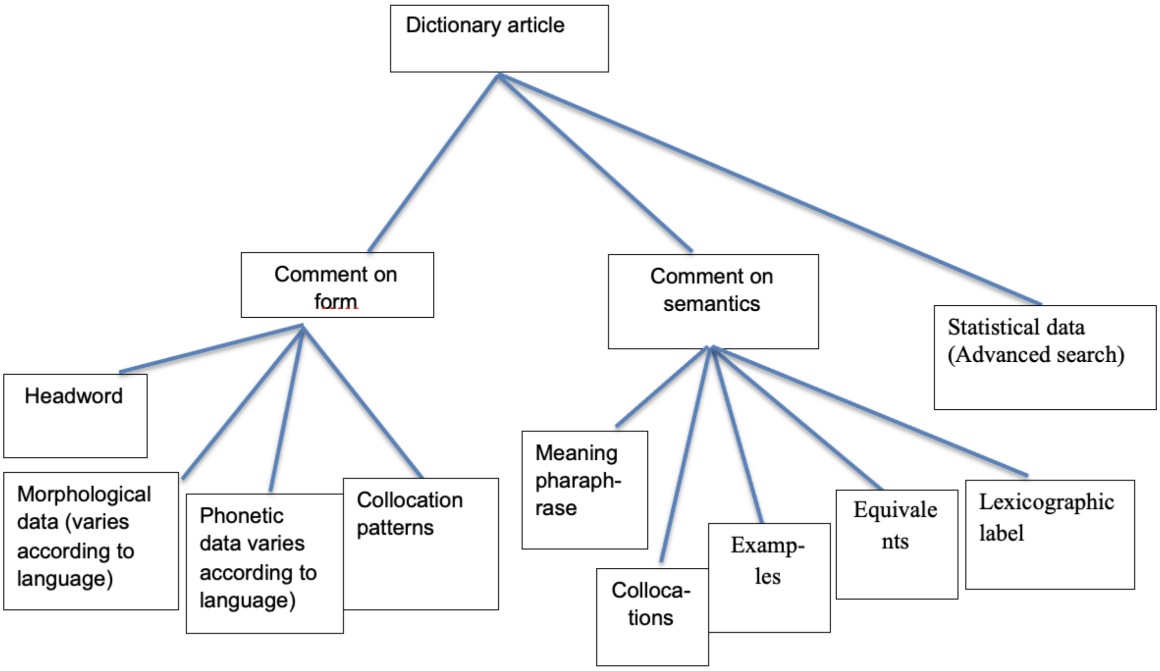


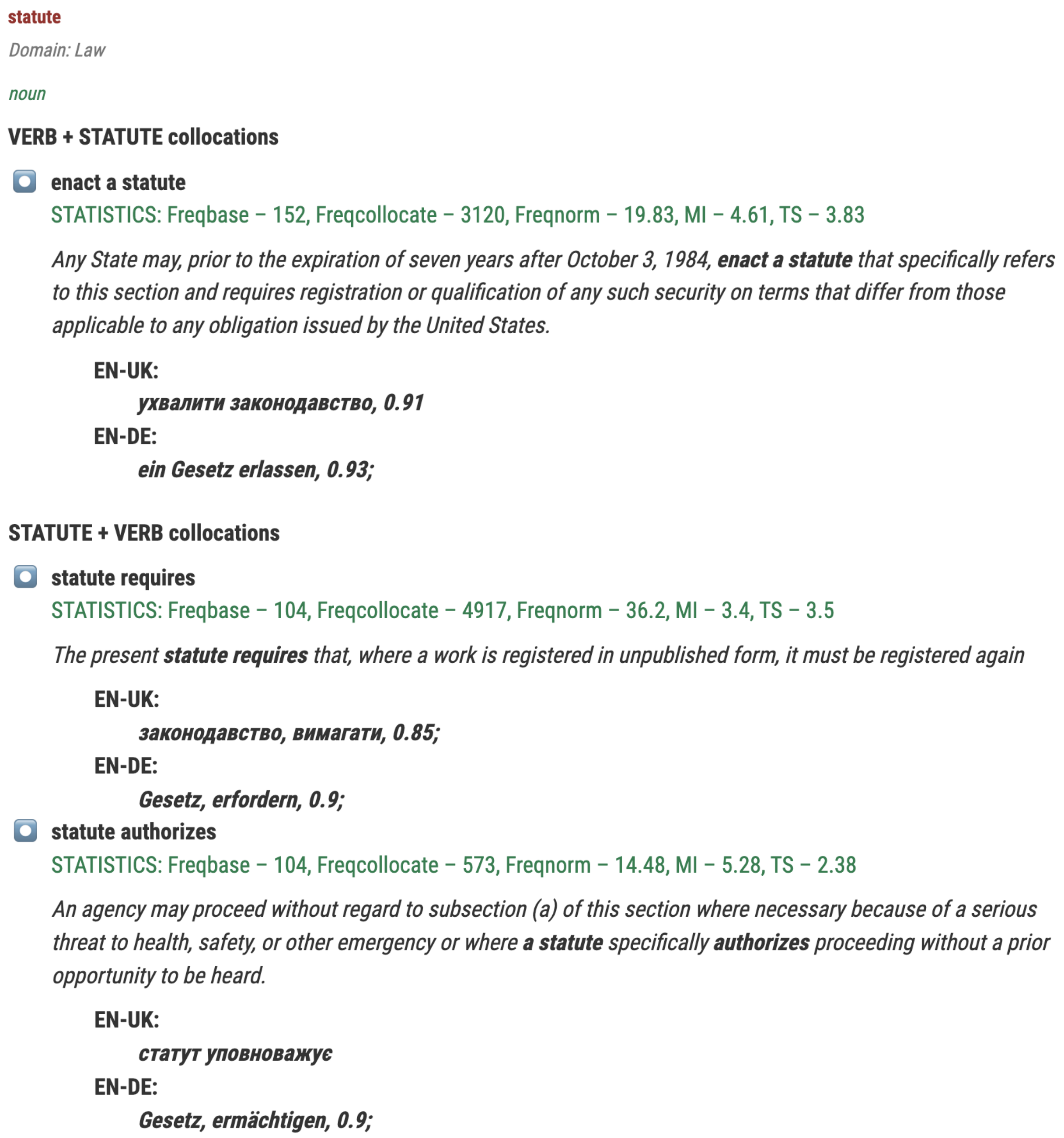
Figure 15. Data types of a model article in EDU-Col

Based on the small English corpora of the legal domain we collected, the data for 3 sample dictionary articles of collocations with the noun, verb, and adjective bases accordingly (Sections 6.6.1, 6.6.2, 6.6.3), whereas the automatically extracted equivalents of collocations will be provided for English-Ukrainian and English-German language pairs. An example of an XML code for the English articles can be seen in Annex 1. The presentation and layout of the dictionary articles are however not final and require further exploration and testing from the design perspective..

Concerning the information types contained in the model dictionary articles presented further, the following comments should be made:

- The frequency indicated for collocations is the frequency normalized per million words.
- The translation confidence score, which is available right next to the equivalent and separated by a comma, is rounded to the decimal number.
- In cases when neither the “weighted” nor “addition” approaches produced acceptable translation candidates in the process of automatic extraction, the translation was provided manually. These instances can be observed for the translation equivalents without the translation confidence score.

6.6.1 An example of a noun as a base collocation article



statute

Domain: Law

noun

VERB + STATUTE collocations

🔘 **enact a statute**

STATISTICS: Freqbase – 152, Freqcollocate – 3120, Freqnorm – 19.83, MI – 4.61, TS – 3.83

Any State may, prior to the expiration of seven years after October 3, 1984, ***enact a statute*** *that specifically refers to this section and requires registration or qualification of any such security on terms that differ from those applicable to any obligation issued by the United States.*

EN-UK:

ухвалити законодавство, 0.91

EN-DE:

ein Gesetz erlassen, 0.93;

STATUTE + VERB collocations

🔘 **statute requires**

STATISTICS: Freqbase – 104, Freqcollocate – 4917, Freqnorm – 36.2, MI – 3.4, TS – 3.5

The present ***statute requires*** *that, where a work is registered in unpublished form, it must be registered again*

EN-UK:

законодавство, вимагати, 0.85;

EN-DE:

Gesetz, erfordern, 0.9;

🔘 **statute authorizes**

STATISTICS: Freqbase – 104, Freqcollocate – 573, Freqnorm – 14.48, MI – 5.28, TS – 2.38

An agency may proceed without regard to subsection (a) of this section where necessary because of a serious threat to health, safety, or other emergency or where ***a statute*** *specifically* ***authorizes*** *proceeding without a prior opportunity to be heard.*

EN-UK:

статут уповноважує

EN-DE:

Gesetz, ermächtigen, 0.9;

ADJECTIVE + STATUTE collocations

- **criminal statute**

  STATISTICS: Freqbase – 676, Freqcollocate – 6291, Freqnorm – 16.32, MI – 3.1, TS – 4.4

  Nothing in this section shall prohibit the attorney general of a State, or other authorized State officer, from proceeding in State or Federal court on the basis of an alleged violation of any ***civil*** *or* ***criminal statute*** *of that State.*

  EN-UK:

  кримінальний статут, 0.84;

  EN-DE:

  strafrechtlich Gesetz, 0.92;

- **civil statute**

  STATISTICS: Freqbase – 676, Freqcollocate – 7965, Freqnorm – 14.37, MI – 2.59, TS – 3.9

  EN-UK:

  цивільний статут, 0.76;

  EN-DE:

  zivil Gesetz, 0.74;

NOUN + preposition + STATUTE collocations

- **violation of a statute**

  STATISTICS: Freqbase – 642, Freqcollocate – 5514, Freqnorm – 27.41, MI – 4.5, TS – 6.69

  Liability should be imposed only for ***violations of statutes*** *or duly issued regulations, after notice and an opportunity to respond.*

  EN-UK:

  законодавство, порушення, 0.82; статут, порушення, 0.79;

  EN-DE:

  Verstoß des Gezetzes, 0.93;

- **provision of a statute**

  STATISTICS: Freqbase – 642, Freqcollocate – 54638, Freqnorm – 20.7, MI – 0.83, TS – 2.66

  It is in accord with the ***provisions of*** *Navy* ***statutes*** *.*

  EN-UK:

  положення статуту

  EN-DE:

  Regelung des Gesetzes, 0.96;

- **effect of a statute**

  STATISTICS: Freqbase – 642 Freqcollocate – 20753 Freqnorm – 12.31 MI – 1.5 TS – 3.04

  No provision of this section shall be construed as altering the validity, interpretation, construction, or ***effect of*** *any State* ***statute*** *.*

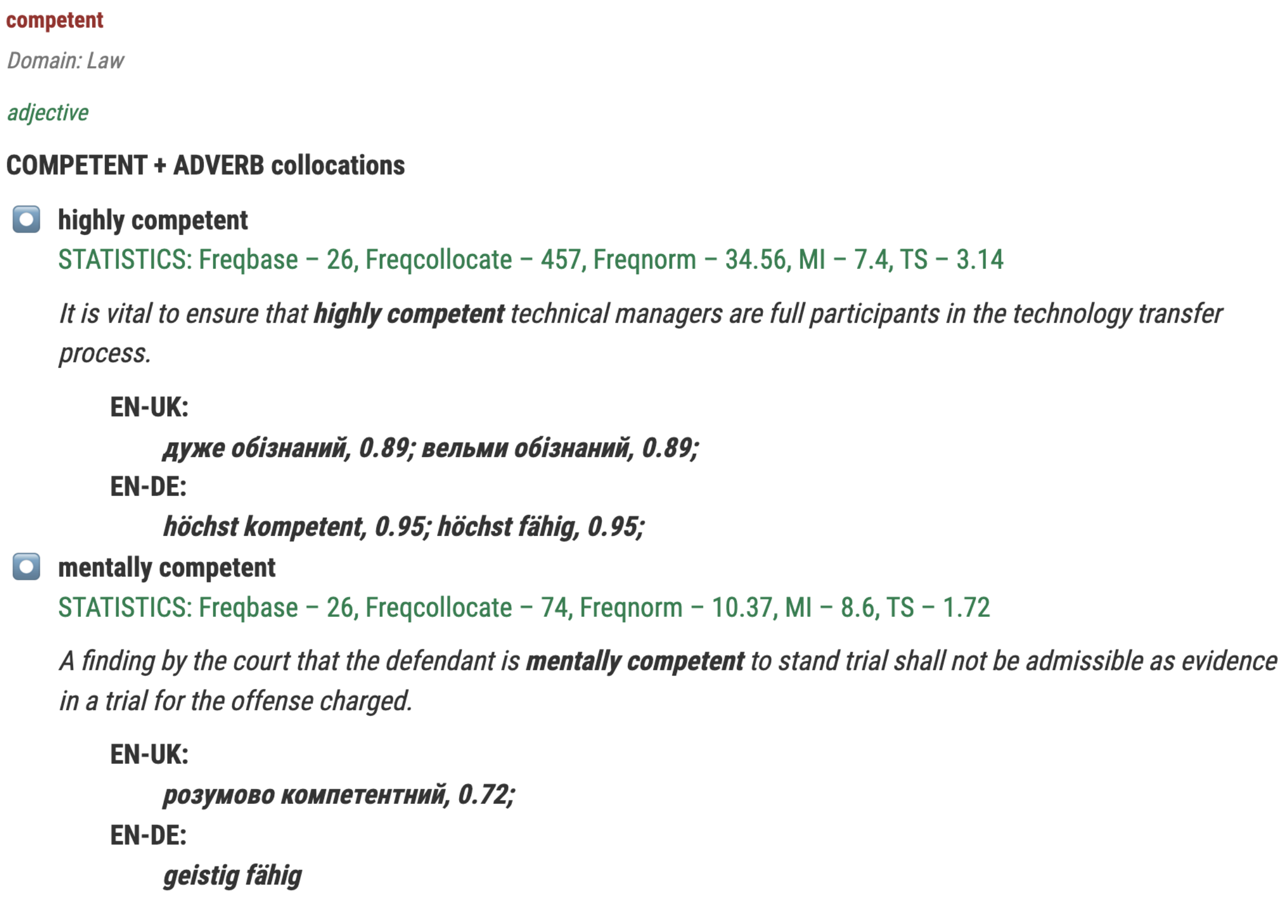
  EN-UK:

  статут, наслідок, 0.78

  EN-DE:

  Wirkung des Gesetzes, 0.94;

6.6.2 An example of an adjective as a base collocation article



competent

Domain: Law

adjective

COMPETENT + ADVERB collocations

◙ **highly competent**

STATISTICS: Freqbase – 26, Freqcollocate – 457, Freqnorm – 34.56, MI – 7.4, TS – 3.14

It is vital to ensure that ***highly competent*** *technical managers are full participants in the technology transfer process.*

EN-UK:

дуже обізнаний, 0.89; вельми обізнаний, 0.89;

EN-DE:

höchst kompetent, 0.95; höchst fähig, 0.95;

◙ **mentally competent**

STATISTICS: Freqbase – 26, Freqcollocate – 74, Freqnorm – 10.37, MI – 8.6, TS – 1.72

A finding by the court that the defendant is ***mentally competent*** *to stand trial shall not be admissible as evidence in a trial for the offense charged.*

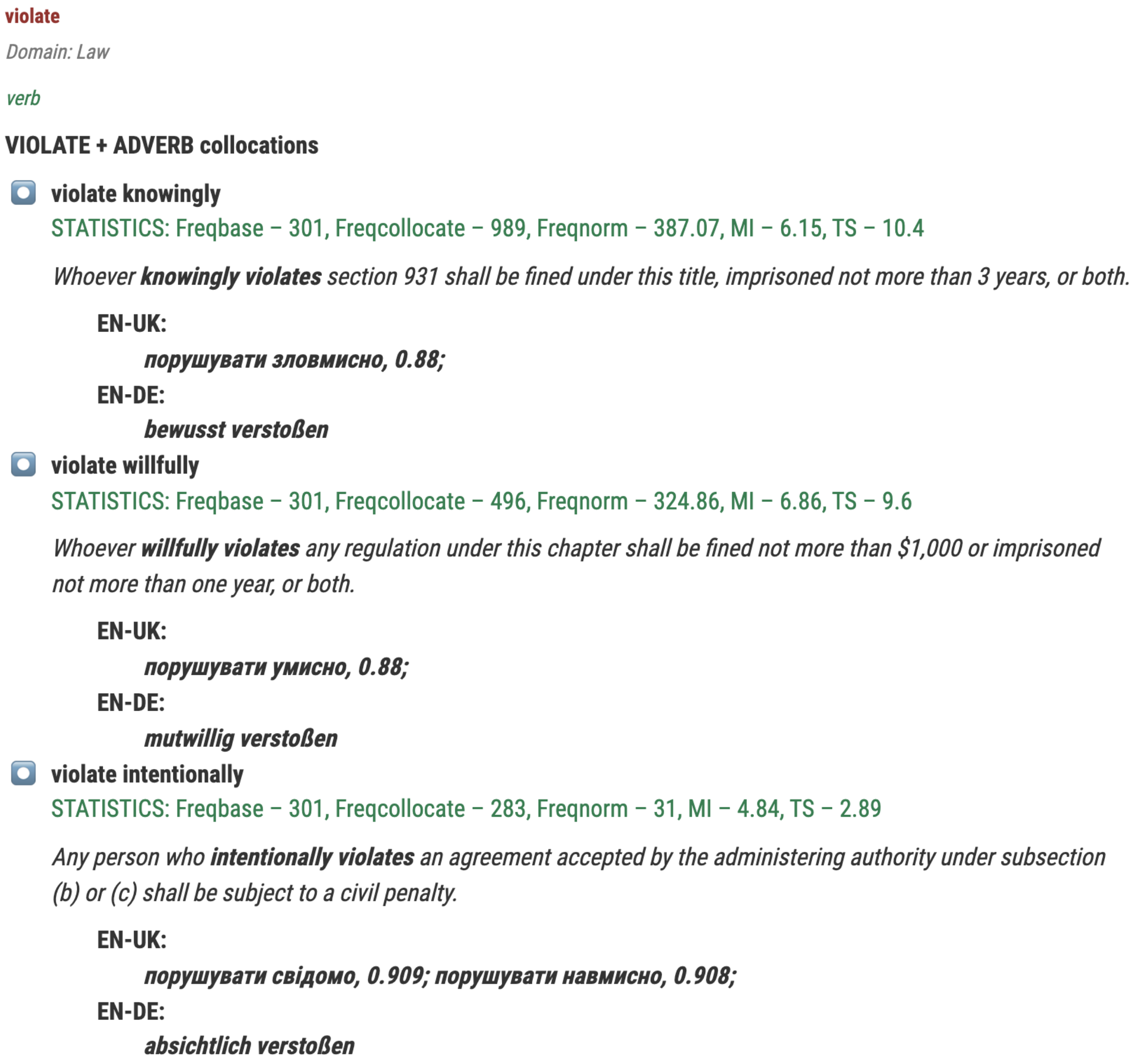
EN-UK:

розумово компетентний, 0.72;

EN-DE:

geistig fähig

6.6.3 An example of a verb as a base collocation article



violate

Domain: Law

verb

VIOLATE + ADVERB collocations

▣ **violate knowingly**
STATISTICS: Freqbase – 301, Freqcollocate – 989, Freqnorm – 387.07, MI – 6.15, TS – 10.4

*Whoever **knowingly violates** section 931 shall be fined under this title, imprisoned not more than 3 years, or both.*

EN-UK:
порушувати зловмисно, 0.88;
EN-DE:
bewusst verstoßen

▣ **violate willfully**
STATISTICS: Freqbase – 301, Freqcollocate – 496, Freqnorm – 324.86, MI – 6.86, TS – 9.6

*Whoever **willfully violates** any regulation under this chapter shall be fined not more than $1,000 or imprisoned not more than one year, or both.*

EN-UK:
порушувати умисно, 0.88;
EN-DE:
mutwillig verstoßen

▣ **violate intentionally**
STATISTICS: Freqbase – 301, Freqcollocate – 283, Freqnorm – 31, MI – 4.84, TS – 2.89

*Any person who **intentionally violates** an agreement accepted by the administering authority under subsection (b) or (c) shall be subject to a civil penalty.*

EN-UK:
порушувати свідомо, 0.909; порушувати навмисно, 0.908;
EN-DE:
absichtlich verstoßen

7. Discussion

7.1 Results

The steps made within the frames of the given thesis included reviewing the existing dictionaries of collocations in English, German and Ukrainian, both monolingual and bilingual, as well as a number of multilingual collocations projects in order to identify how the proposed dictionary will contribute to the realm of practical lexicography. Further, the proposed *EDU-Col* dictionary project was characterized in terms of dictionary types, the target users were defined, and the dictionary functions and usage situations were presented. Moving to the subject matter of *EDU-Col,* the approaches to the interpretation of collocations in linguistics and lexicography were outlined prior to the presentation of the methodology adopted for the dictionary compilation process.

As a practical experiment, a decision was taken to extract several dictionary information types, namely collocation candidates, examples, and translation equivalents for the previously extracted collocations following the outlined approach. In order to fulfill these tasks, we prepared the corpora for the three languages, which will be represented in the dictionary, and processed it with UDPipe to tokenize, lemmatize and syntactically parse the data. Finally, the dictionary structures presentation (macro-, micro-, medio-, and access structures) was contemplated, and the 3 prototype dictionary articles for different types of collocations using the automatically extracted data were created.

The corpora processed in the previous step were subsequently used to obtain sample results of collocation candidates with noun, adjective, and verb collocations bases. Since for this project we restricted the extraction of dictionary data to a single domain, relying on 50 mln token corpora, the number of resulting collocation candidates is rather small for each of the three languages (see Table 6).

Corpus (Legal Domain)	Nouns		Adjectives		Verbs	
	Bases	Collocation candidates	Bases	Collocation candidates	Bases	Collocation candidates
English	6,148	56,915	2,571	9,983	1,761	11,525
Ukrainian	6,662	57,876	4,873	13,717	1,974	13,870
German	13,256	12,190	4,615	6,629	1,815	7,018

Table 6. Number of collocation candidates extracted for validated noun, adjective, and verb bases.

As it was already mentioned throughout the work, these extracted candidates require manual inspection by lexicographers before they can be used for compiling dictionary articles.

As far as the generation of example sentences is concerned, according to the adopted methodology, the number to be extracted for each collocation is 10. However, as it was observed not all collocations received 10 example sentences, possibly due to the fact that the corpora we used are relatively small. Moreover, the sentences in texts from the legal domain (mainly laws, legal codes, etc) that constitute our corpora tend to be rather long and complicated, thus not all of the sentences qualify according to the GDEX approach (in our case "simplified GDEX") applied in the process of examples extraction.

Another aspect to be discussed is the generation of translation equivalents of collocations. As it was mentioned in section 5.2.4 Generation of translation equivalents, we employ the "weighted" and "addition" approaches. Comparing these two approaches on our sample collocations, it was revealed that the "addition approach" works slightly better overall. In particular, we selected 50 collocations from our English corpus (50 collocations of verb-noun pattern, 50 collocations of adjective-noun pattern, and 50 of adjective-adverb structure) and translated them into Ukrainian using the two approaches to evaluate which one achieves better results on our data. The results of the manual inspection as to which translation candidates were acceptable can be seen in Table 7. Better performance of the "addition approach" is presumably related to the fact that at least between the English and Ukrainian languages, most of the extracted collocations from the corpus belonging to the legal domain tend to be congruent and thus require word-for-word equivalents, and the "weighted approach" is known to perform worse on such types of collocations.

Collocation types	EN-UK	
	Weighted approach	Addition approach
verb-noun *(obj)*	28/50 (56%)	38/50 (76%)
adjective-noun (*amod)*	23/50 (46%)	35/50 (70%)
adjective-adverb (*advmod)*	36/50 (72%)	37/50 (74%)

Table 7. Translation candidates as translated using "weighted" and "addition" approaches

In another small experiment, 50 English verb-noun collocations that are non-congruent between English and Ukrainian were selected and translated into Ukrainian using the "weighted" and "addition" approaches. The results produced by the "weighted approach" were expected to be better as the number of acceptable equivalents it generated was just 5 out of 50. However, the collocations selected for this test were of major difficulty in terms of translation as at times two components of a collocation could not be translated literally, not just one. Another reason might be related to the way that the word embeddings we used were mapped., i.e., fully unsupervised, whereas in Garcia et al., (2019c) it was semi-supervised with the help of a list of translated words. In turn, the “addition approach” did not produce any meaningful results in this experiment.

7.1.2 Limitations

The given Master thesis is by no means an extensive enough study to commence the phase of large-scale implementation of the proposed dictionary plan. In the experiment with automatic generation of the dictionary data for domain-specific collocations dictionary articles, we did not account for the extraction of one more type of dictionary data, namely definitions. What can be made further is enriching the corpora semantically in order to organize collocations according to their specific senses. For instance, in Orenha-Ottaiano’s et al., (2021) project, automatic extraction of definitions was performed by enriching each collocation base with the potential senses with the Open Multilingual WordNet (Bond & Foster, 2013).

7.1.3 Challenges faced

The challenges were related to a lack of preconditions important for the execution of the procedures of the chosen methodology. Since the computational processing of large corpora is resource-demanding and most regular personal computers cannot handle the workload, we experimented with generating domain-specific collocations, which allowed using comparatively smaller corpora.

In turn, even the step of the compilation of corpora can be challenging, since the use of collections of texts, where the quality of present material cannot be fully under control, may result in poor quality of the output data, as it was in the first attempt to use an existing corpus of legal texts for Ukrainian.

8. Further work

Following Klosa’s (2013) model of lexicographical process (Figure 16) for a corpus-based online dictionary, we completed certain steps belonging to the phase of preparation, namely conceptional design of the dictionary, phase of data acquisition of primary sources, phase of computerization – annotation, tagging of corpus texts, phase of data processing with the extraction of dictionary information types, and phase of data analysis with writing model dictionary articles. However, the scale of the actions carried out is undoubtedly smaller than they would be in a full-scale lexicographic project.

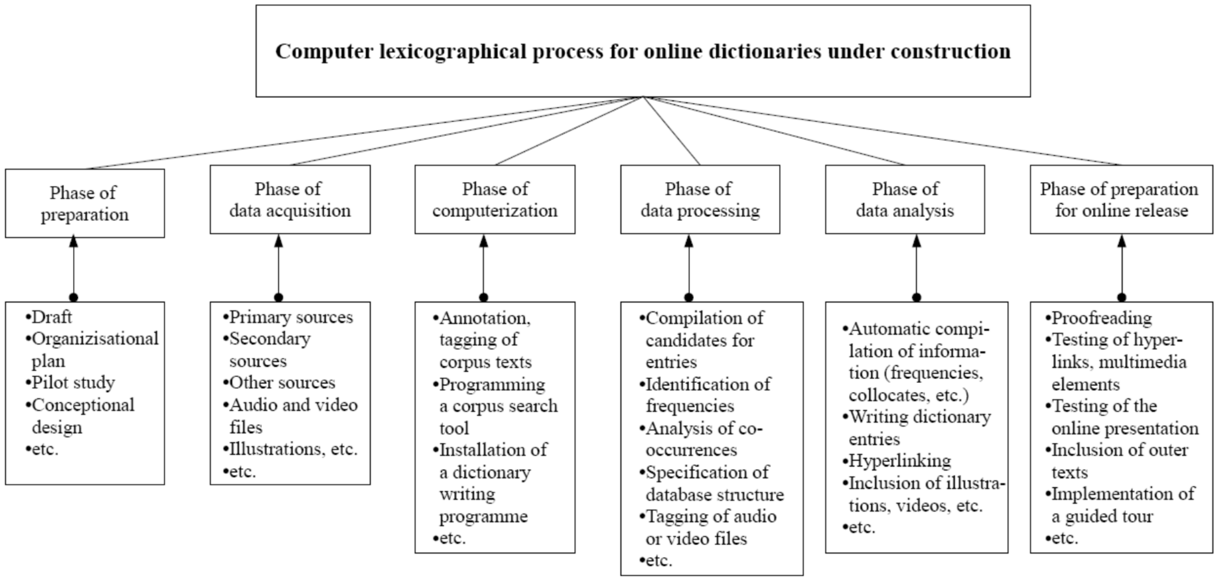

Computer lexicographical process for online dictionaries under construction
Phase of preparation
Phase of data acquisition
Phase of computerization
Phase of data processing
Phase of data analysis
Phase of preparation for online release
•Draft
•Organizisational plan
•Pilot study
•Conceptional design
•etc.
•Primary sources
•Secondary sources
•Other sources
•Audio and video files
•Illustrations, etc.
•etc.
•Annotation, tagging of corpus texts
•Programming a corpus search tool
•Installation of a dictionary writing programme
•etc.
•Compilation of candidates for entries
•Identification of frequencies
•Analysis of co-occurrences
•Specification of database structure
•Tagging of audio or video files
•etc.
•Automatic compilation of information (frequencies, collocates, etc.)
•Writing dictionary entries
•Hyperlinking
•Inclusion of illustrations, videos, etc.
•etc.
•Proofreading
•Testing of hyperlinks, multimedia elements
•Testing of the online presentation
•Inclusion of outer texts
•Implementation of a guided tour
•etc.

Figure 16. Illustration of the computer-lexicographical process for online dictionaries under construction (In Klosa, 2013, p. 520)

As Klosa (2013) argues, the lexicographical process for an “online dictionary under construction”, i.e., a work-in-progress resource, is not linear but circular: “Producing an online dictionary may begin before the phase of writing is finished: online dictionaries can be published step-by-step. Thus, all phases of the computer-lexicographical process (planning – writing – producing) merge [...], giving yet unknown flexibility to the lexicographer” (Klosa 2013, p. 519). Hence, the lines of work that could be possibly undertaken later for our proposed dictionary project include:

1) Tokenizing, lemmatizing and syntactically analyzing the large general language corpora and generating general language collocations for each language;

2) Compiling other domain-specific corpora to obtain collocations belonging to other domains, e.g., economics, environment, etc;

3) Enriching the corpora with semantic information automatic extraction of definitions according to specific senses of collocation bases;

4) Developing a platform/portal which would host not only the dictionary but other resources and features such as:

- functionality, which records all the searches made by the users and allows to review them later;
- interactive exercises on learning colocations for learners.

5) Developing an international survey with learners and translators as a focus group on their needs, which could be used to implement important modifications in the proposed design.

9. Conclusions

This Master's thesis was an effort to contribute to filling the gap in practical multilingual lexicography across three languages - English, German, and Ukrainian with a plan of an online collocations dictionary based on the automatic approach, which relies on computational advances and NLP (Orenha-Ottaiano, 2021, Garcia et al., 2019a, 2019c). In this approach, lexical data, namely collocations, examples, and translation equivalents are automatically generated, and afterward, lexicographers can analyze the data deciding on the final dictionary article contents. The fact that the major part of the work is done automatically, significantly saves time and diminishes the number of human resources that would be otherwise necessary for the dictionary-making processes. In turn, the outlined methodology coupled with the manual validation by lexicographers will allow for maintaining a high quality of lexicographic data.

The method adopted for the compilation of *EDU-Col* project was experimented with for the first time in the context of the three languages covered, with a series of challenges arising in the process.

The present research is by no means fully ready to be implemented into existence, with a number of steps, which would be necessary to be taken such as the survey of potential users, computational preparation of larger corpora for general language collocations, etc. Nevertheless, looking further into the future, if *EDU-Col* was brought to life, it would contribute considerably to the needs of learners, translators as well as native speakers, especially for the Ukrainian speaking target audience since, as compared to the existing resources for English and to a lesser extent German, even Ukrainian monolingual collocation dictionaries based on corpora are scarce or rather non-existent.

References

Adamska-Sałaciak, A. (2010). Why we need bilingual learners' dictionaries. English Learners' Dictionaries at the DSNA. Tel Aviv: K Dictionaries, pp. 121-137.

Altenberg, B & Granger, S. (2001). The grammatical and lexical patterning of MAKE in native and non-native student writing. Applied Linguistics, Volume 22, Issue 2, pp. 173-195.

Anderson, B. E & Fuertes Olivera, P. (2009). The Application of Function Theory to the Classification of English Monolingual Business Dictionaries. In: Lexicographica 25, pp. 213-239.

Artetxe, M., Labaka, G. & Agirre, E. (2018). A robust self-learning method for fully unsu-pervised cross-lingual mappings of word embeddings. In I. Gurevych & Y. Miyao (eds.) Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia, pp. 789–798.

Atkins, B. T. S. & Rundell, M. (2008). The Oxford Guide to Practical Lexicography, Oxford: Oxford University Press.

Bahns, J. & Eldaw. M. (1993). 'Should we Teach EFL Students Collocations?' In System, 21.1, pp. 101–114.

Barnbrook, G., Mason, O. & Krishnamurthy, R. (2013). Collocation: Applications and implications. Houndmills, Basingstoke and New York: Palgrave Macmillan, pp. 254.

Baroni, M., Bernardini, S., Ferraresi A. & Zanchetta, E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. Language Resources and Evaluation 43 (3), pp. 209-226.

Benson, M., Benson, E. & Ilson, R. (1997). The BBI Combinatory Dictionary of English. Amsterdam/Philadelphia: John Benjamins.

Bird, S., Klein, E.,, Loper, E. (2009). Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. Beijing: O'Reilly. ISBN: 978-0-596-51649-9.

Boleda, G. (2020). Distributional Semantics and Linguistic Theory. In *Annu. Rev. Linguist.* 6, pp. 1-22.

Bond, F. & Foster, R. (2013). Linking and extending an open multilingual wordnet. In H. Schuetze, P. Fung & M. Poesio (eds.) Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Sofia, Bulgaria, pp. 1352-1362.

Buhofer A.H., Dräger M., Meier S. & Roth T (Hg.) (2014). Feste Wortverbindungen des Deutschen - Kollokationenwörterbuch für den Alltag, inkl. Beiheft für Selbststudium und Unterricht. Tübingen: Francke Verlag.

Bybyk S.P., Yermolenko S.P. & Pustovit S.Ya. (1998). Slovnyk epitetiv ukrainskoi movy [Dictionary of the epithets of the Ukrainian language]. K.: Dovira, pp. 431.

Cowie, A. P. (1999). English dictionaries for foreign learners: a history. Oxford: Oxford. University Press.

Engelberg, S. & Lemnitzer, L. (2009). Lexikographie und Wörterbuchbenutzung. Tübingen 4., überarbeitete und erweiterte Auflage.

Evert, S., Uhrig, P., Bartsch, S. & Proisl, T. (2017). E-VIEW-affilation–A large-scale evaluation study of association measures for collocation identification. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek & V. Baisa (eds.) Proceedings of eLex 2017–Electronic lexicography in the 21st century: Lexicography from Scratch. Leiden, the Netherlands, pp. 531-549.

Evert, Stefan (2008). Corpora and collocations. In A. Lüdeling and M. Kytö (eds.), Corpus Linguistics. An International Handbook, article 58, pages 1212-1248. Mouton de Gruyter, Berlin. https://www.stephanie-evert.de/PUB/Evert2007HSK_extended_manuscript.pdf

Firth, J.R. (1957). A synopsis of linguistic theory, 1930- 1955. In J.R. Firth, editor, *Studies in Linguistic Analysis.* Basil Blackwell, Oxford, UK.

Firth, J.R. (1951). "Modes of Meaning" Essays and Studies, n.s. 4: 118-49. Rptd. in Firth 1957, pp. 190-214.

Frankenberg-Garcia, A. (2012). Learners' Use of Corpus Examples. International Journal of Lexicography 25.3, p. 273–296.

Frankenberg-Garcia, A. (2014). The Use of Corpus Examples for Language Comprehension and Production. ReCALL 26.2, pp. 128–146.

Garcia, M., García-Salido, M. & Alonso-Ramos, M. (2017). Using bilingual word-embeddings for multilingual collocation extraction. In S. Markantonatou, C. Ramisch, A. Savary & V. Vincze (eds.) Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017). Valencia, Spain, pp. 21-30.

Garcia, M., García-Salido, M. & Alonso-Ramos, M. (2019a). Towards the automatic construction of a multilingual dictionary of collocations using distributional semantics. In I. Kosem, T. Z. Kuhn, M. Correia, J. P. Ferreira, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (eds.) Proceedings of eLex 2019: Smart Lexicography. Sintra, Portugal, pp. 747-762.

Garcia, M., García-Salido, M. & Alonso-Ramos, M. (2019b). A comparison of statistical association measures for identifying dependency-based collocations in various languages. In Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), at the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019). Association for Computational Linguistics, pp. 49-59.

Garcia, M., García-Salido, M. & Alonso-Ramos, M. (2019c). Weighted compositional vectors for translating collocations using monolingual corpora. In G. Corpas Pastor & R. Mitkov (eds.) Computational and Corpus-Based Phraseology. Cham, Switzerland: Springer, pp. 113-128.

Gilquin, G. (2007). To err is not all: What corpus and elicitation can reveal about the use of collocations by learners. Zeitschrift für Anglistik und Amerikanistik, 55(3), pp, 273-291.

Gouws, R. H. (2014). Article Structures: Moving from Printed to e-Dictionaries. In *Lexikos* 24, pp. 155-177.

Gouws, R. H. (2007). A transtextual approach to lexicographic functions. In Lexikos, 17, pp. 77-87.

Gouws, R.H. (2010). The use of an improved access structure in dictionaries. In Lexikos. 11.

Gouws, R.H. & Prinsloo, D.J. (2005). Principles and practice of South African lexicography. Stellenbosch: African Sun MeDIA.

Halliday, M.A.K. & Hasan, R. (1976). Cohesion in English, London.

Hasan, R. (1984). Coherence and cohesive harmony, in J. Flood (ed.): Understanding Reading Comprehension 181-219).

Hausmann, F.J. (1984). "Wortschatzlernen ist Kollokationslernen: Zum Lehren und Lernen französischer Wortverbindungen". Praxis des neusprachlichen Unterrichts, Vol. 31, pp. 395-406.

Hausmann, F.J. (1989). "Le dictionnaire de collocations". In F.J. Hausmann, O. Reichmann, H.E. Wiegand and L. Zgusta, eds, pp. 1010-19.

Hausmann, F.J. (1989). Wörterbuchtypologie. In: F. J. Hausmann, O. Reichmann, H.E. Wiegand & L. Zgusta (Hrsg.) Wörterbücher. Ein internationales Handbuch zur Lexikographie. 1. Teilband. Berlin, New York: de Gruyter (HSK), pp. 968-981.

Hausmann, F.J., Reichmann, O., Wiegand, H.E., & Zgusta, L. (1989). Component Parts and Structures of General Monolingual Dictionaries: A Survey.

Herbst, T. (1996). What are collocations: Sandy beaches or false teeth?. In English Studies, 77:4, pp. 379-393.

Hill J. & Lewis M., Eds. (1997) Dictionary of Selected Collocations. Hove, UK: Language Teaching Publications, pp. 288.

Hollós, Z. (2008). "Kollokationen und weitere typische Mehrwortverbindungen in der ungarischen Lexikographie". In Lexicographica 24, pp. 121–133.

Hollós, Z. (2014). KolleX. Deutsch-ungarisches Kollokationslexikon. Korpusbasiertes Wörterbuch der Kollokationen. Deutsch als Fremdsprache. Szeged.

Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Jones, S. & Sinclair, J. (1974). 'English lexical collocations. A study in computational linguistics'. In Cahiers de lexicologie 24/25, pp. 15-61.

Kilgarriff, A., Rychlý, P., Smrž, P. & Tugwell, D. (2004) Itri-04-08 the sketch engine. Information Technology.

Kilgarriff, A., & Kosem, I. (2011). Corpus tools for lexicographers.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit. J., Rychlý, P. & Suchomel, V. (2014) The Sketch Engine: ten years on. Lexicography, 1, pp. 7-36.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychly, P. (2008). GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In E. Bernal & J. DeCesaris (eds.) Proceedings of the 13th EURALEX International Congress. Barcelona: Institut Universitari de Linguistica Aplicada/Universitat Pompeu Fabra, pp. 425–432.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychly, P. (2008). GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In E. Bernal & J. DeCesaris (eds.) Proceedings of the 13th EURALEX International Congress. Barcelona: Institut Universitari de Linguistica Aplicada/Universitat Pompeu Fabra, pp. 425–432.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychly, P. (2008). GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In E. Bernal & J. DeCesaris (eds.) Proceedings of the 13th EURALEX International Congress. Barcelona: Institut Universitari de Linguistica Aplicada/Universitat Pompeu Fabra, pp. 425–432.

Klosa, A. (2013). The lexicographical process (with special focus on online dictionaries). In Hausmann, F. J. et al. (eds.): Wörterbücher. Ein internationales Handbuch zur Lexikographie. Supplement Volume: Recent Developments with Focus on Electronic and Computational Lexicography. (= Handbücher zur Sprach- und Kommunikationswissenschaft 5/4). Berlin/Boston: de Gruyter, 517-524.

Klosa, A. & Gouws, R. (2015). Outer features in e-dictionaries. In Lexicographica. 31.

Kosem, I., Koppel, K., Kuhn, T. Z., Michelfeit, J. & Tiberius, C. (2019). Identification and automatic extraction of good dictionary examples: the case(s) of GDEX. International Journal of Lexicography, 32(2), pp. 119–137.

Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J. & Laskowski, C. (2019). Collocations Dictionary of Modern Slovene. In Proceedings of the 18th EURALEX

International Congress: lexicography in global contexts. Ljubljana: Ljubljana University Press, Faculty of Arts, pp. 989-997.

Krenn, B. & Evert, S. (2001). Can we do better than frequency? A case study on extracting PP-verb collocations. In Proceedings of the ACL Workshop on Collocations. Association for Computational Linguistics, pp. 39–46.

Laufer, B. (2011). The Contribution of Dictionary Use to the Production and Retention of Collocations in a Second Language. International Journal of Lexicography, 24(1), pp. 29-49.

Mel'ˇcuk, I. (2012). Phraseology in the language, in the dictionary, and in the computer. Yearbook of Phraseology 3(1), pp. 31–56.

Nesselhauf, N. (2004). What Are Collocations? In D. J. Allerton, N. Nesselhauf & P. Skandera, eds., pp. 1-21.

Orenha-Ottaiano, A. (2017). The compilation of an Online Corpus-Based Bilingual Collocations Dictionary: motivations, obstacles and achievements. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek & V. Baisa (eds.) Proceedings of eLex 2017–Electronic lexicography in the 21st century: Lexicography from Scratch. Leiden, the Netherlands, pp. 458-473.

Orenha-Ottaiano, A., García, M.A., Eugenia., M., Silva, O.D., L'Homme, M., Margarita, Ramos, A., Valêncio, C.R., & Tenório, W. (2021). Corpus-based Methodology for an Online Multilingual Collocations Dictionary: First Steps. In Kosem, I., Cukr, M., Jakubíček, M., Kallas, J., Krek, S. & Tiberius, C. (eds.) 2021. Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference. 5–7 July 2021, virtual. Brno: Lexical Computing CZ, s.r.o.

Palmer, H. E. (1933a). Second Interim Report on English Collocations. Tokyo: Kaitakusha.

Palmer, H. E. (1933b). Some notes on construction-patterns. IRET Institute Leaflet 38.

Poulsen, S. (2022). Collocations as a Language Resource. A functional and cognitive study in English phraseology. In Human Cognitive Processing, 71. John Benjamins. xvi, pp. 348.

Quasthoff, U. (2011). Wörterbuch der Kollokationen im Deutschen. Berlin/New York.

Rapp, R. (1999). Automatic Identification of Word Translations from Unrelated English and German Corpora. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics. College Park, Maryland, USA: Association for Computational Linguistics, pp. 519–526.

Ruder, S., Vulić, I. & Søgaard, A. (2019). A Survey of Cross-Lingual Word Embedding Models. Journal of Artificial Intel ligence Research. arXiv preprint arXiv:1706.04902.

Rundell, M. & Atkins, B. (2013). 96. Criteria for the design of corpora for monolingual lexicography. In R. Gouws, U. Heid, W. Schweickard & H. Wiegand (Ed.), Supplementary Volume Dictionaries. An International Encyclopedia of Lexicography. Berlin, Boston: De Gruyter, pp. 1336-1343.

Sakhno, I. P. (1999) Slovnyk spoluchuvanosti sliv ukrainskoi movy: naiuzhyvanisha leksyka [Dictionary of the compatibility of the words of the Ukrainian language: the most commonly used vocabulary]. Dnipropetrovsk: Dnipropetrovsk state university publishing house, pp. 544.

Schierholz, S. (2015). Methods in Lexicography and Dictionary Research. Lexikos, 25, pp. 323-352.

Seretan, V. (2013). On Collocations and Their Interaction with Parsing and Translation. Informatics. 1, pp. 11-31.

Shevchuk, Y., I. (2021). The Ukrainian-English Collocation Dictionary. Hippocrene Books.

Sinclair, J. (1966). Beginning the Study of Lexis. In C.E. Bazell, J.C. Catford, M.A.K. Halliday & R.H. Robins, eds., pp. 288-302.

Straka, M. & Straková, J. (2017).Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada, August 2017.

Straka, M., Hajič, J. & Straková J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, May 2016.

Tarp, S. (2002). Functions in „de Gruyter Wörterbuch Deutsch als Fremdsprache“. In: Wiegand, H. E. (Hg.): Perspektiven der pädagogischen Lexikographie des Deutschen II. Untersuchungen anhand des de Gruyter Wörterbuches Deutsch als Fremdsprache. (Lexicographica. Series Maior 110). Tübingen, pp. 609-619.

Wiegand, H. & Gouws, R. (2013). 4. Macrostructures in printed dictionaries. In R. Gouws, U. Heid, W. Schweickard & H. Wiegand (Ed.), Supplementary Volume Dictionaries. An International Encyclopedia of Lexicography: Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography (pp. 73-110). Berlin, Boston: De Gruyter Mouton.

Wiegand, H. E., Beißwenger, M., Gouws, R., Kammerer, M., Storrer, A & Wolski, W. (2020). Wörterbuch zur Lexikographie und Wörterbuchforschung mit englischen Übersetzungen der Umtexte und Definitionen sowie Äquivalenten in neun Sprachen = Dictionary of lexicography and dictionary research. Band 4 V-Z : Nachträge und Gesamtregister A-H : Symbolverzeichnis : Wörterbuchbasis / herausgegeben und bearbeitet von Herbert Ernst Wiegand, Rufus H. Gouws, Matthias Kammerer, Michael Mann, Werner Wolski. Berlin Boston de Gruyter.

Wiegand, H. (2000). Kleine Schriften: Eine Auswahl aus den Jahren 1970-1999 in zwei Bänden. Bd 1: 1970-1988.

Yermolenko S.Ya., Yermolenko V.I. & Bybyk S.P. (2012) Novyi slovnik epitetiv ukrainskoi movy [New Dictionary of Ukrainian Language epithets]. K.: Hramota, pp. 487.

Zeman, D.; et al., (2021). Universal Dependencies 2.9, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-4611.

Annex 1. XML code of article model (exported from lexonomy.eu)

```xml
<entry lxnm:entryID='6'
    xmlns:lxnm='http://www.lexonomy.eu/'>
    <headword></headword>
    <partOfSpeech></partOfSpeech>
    <pronunciation>
        <UK></UK>
        <US></US>
    </pronunciation>
    <sense1>
        <collocations>
            <grammaticalRelation></grammaticalRelation>
            <Ncollocation>
                <collocation></collocation>
                <statisticalData></statisticalData>
                <example></example>
                <translationDirection>
                    <pair1>
                        <direction></direction>
                        <equivalent></equivalent>
                    </pair1>
                    <pair2>
                        <direction></direction>
                        <equivalent></equivalent>
                    </pair2>
                </translationDirection>
            </Ncollocation>
        </collocations>
        <collocations>
            <grammaticalRelation></grammaticalRelation>
            <Ncollocation>
                <collocation></collocation>
                <statisticalData></statisticalData>
                <example></example>
                <translationDirection>
                    <pair1>
                        <direction></direction>
                        <equivalent></equivalent>
                    </pair1>
                    <pair2>
                        <direction></direction>
                        <equivalent></equivalent>
                    </pair2>
                </translationDirection>
            </Ncollocation>
        </collocations>
    </sense1>
    <sense2>
        <collocations>
            <grammaticalRelation>
                <Ncollocation>
                    <collocation></collocation>
                    <statisticalData></statisticalData>
                    <example></example>
                    <translationDirection>
                        <pair1>
                            <direction></direction>
                            <equivalent></equivalent>
                        </pair1>
                        <pair2>
                            <direction></direction>
                            <equivalent></equivalent>
                        </pair2>
                    </translationDirection>
                </Ncollocation>
            </grammaticalRelation>
            <Ncollocation></Ncollocation>
        </collocations>
        <collocations>
            <grammaticalRelation>
                <Ncollocation>
                    <collocation></collocation>
                    <statisticalData></statisticalData>
                    <example></example>
                    <translationDirection>
                        <pair1>
                            <direction></direction>
                            <equivalent></equivalent>
                        </pair1>
                        <pair2>
                            <direction></direction>
                            <equivalent></equivalent>
                        </pair2>
                    </translationDirection>
                </Ncollocation>
            </grammaticalRelation>
            <Ncollocation></Ncollocation>
        </collocations>
    </sense2>
</entry>
```

Authorization for the Distribution

Declaration of Originality

I declare herewith, that this work is my own original work.

Furthermore, I confirm that:

− I have referenced in accordance with the guidelines all sources and references (print and internet) in text and the References section;
− All data and findings in the work have not been falsified or embellished;
− This work has not been previously used either for other courses or exams;
− This work has not been published.

University of Santiago de Compostela, July, 2022
Full Name: Mariia Polova