ktu 1922 faculty of social sciences, arts and humanities

SAL 40/2022
Research Journal
Studies about Languages
pp. 61-74
ISSN 1648-2824 (print)
ISSN 2029-7203 (online)
DOI 10.5755/j01.sal.40.1.30735

LINGUISTICS / KALBOTYRA

English-Ukrainian Parallel Corpus: Prerequisites for Building and Practical Use in Translation Studies

Received 02/2022 | Accepted 05/2022

crossref http://dx.doi.org/10.5755/j01.sal.40.1.30735

HOW TO CITE: Matvieieva, S. A., Lemish, N. Ye., Zernetska, A. A., Babych, V. O., & Torgovets, M. A. (2022). English-Ukrainian parallel corpus: Prerequisites for building and practical use in Translation Studies. *Studies about Languages / Kalbų studijos*, 40, 61–74. http://doi.org/10.5755/j01.sal.40.1.30735

English-Ukrainian Parallel Corpus: Prerequisites for Building and Practical Use in Translation Studies

Anglų-ukrainiečių kalbų lygiagrečiojo tekstyno kūrimo ir praktinio naudojimo vertimo studijose prielaidos

SVITLANA A. MATVIEIEVA, National Pedagogical Dragomanov University, Ukraine

NATALIYA YE. LEMISH, National Pedagogical Dragomanov University, Ukraine

ALLA A. ZERNETSKA, National Pedagogical Dragomanov University, Ukraine

VOLODYMYR O. BABYCH, International European University, Ukraine

MARYNA A. TORGOVETS, Private Higher Educational Institution “Institute for Ecology Economy and Law”, Ukraine

Abstract

Consistent demand for highly professional translators determines continuous attempts of researchers and programmers to develop and propose reliable tools for both improvement of translation quality and facilitation of translators’ work. Last ten years have brought the parallel and comparable corpora into the focus of Ukrainian scientists’ attention. The aim of the paper is to specify the prerequisites for building the English-Ukrainian parallel corpus and describe its application in Translation Studies. A parallel corpus as a separate type of linguistic corpora cannot be built without alignment that enables placing and extracting corresponding sentences/paragraphs of source and target texts in one space. To create parallel corpora, it is necessary to perform additional text preparation. The Sketch Engine system (an example of a web-oriented system for work with corpora) can offer the solution for annotation with Excel. However, Sketch Engine lacks artificial intelligence techniques for further word processing. There is probability that employment of a neural network in the future will enable text alignment in parallel corpora instead of system users. Data from parallel corpora can be used in translation lexicography, comparative lexico-grammatical works, studies in the theory and practice of translation, language teaching, and development of machine translation systems. Corpus-based translation analysis is extremely relevant to identifying translation solutions that can only be explored on the basis of translation products. It is stipulated by rather frequent absence of dictionary equivalents in most contexts and ready evidence of possible translation variants in parallel corpora that provide the usage of a language unit in a wide range of contexts.

KEYWORDS: English-Ukrainian parallel corpus, annotation, text alignment, translation studies.

Introduction

Translation is a tool that allows keeping culture and passing it to the next generations. It has never been an easy task to do. A translator shall have a solid background and an excellent command of at least two languages (source and target ones). He/she shall posses lively mind and broad outlook, intelligence and erudition. The vocabulary shall be as rich as it is required for rendering a wide variety of topics/themes, translation may concern. Equivalence and adequacy of translation also need deep knowledge of genre/style nuances, intercultural communication specifics, and business etiquette peculiarities. A translator is a person who connects different worlds, people, and nations. In addition, a translator shall be able to analyze the data, to concentrate on and explain some specific information, to make necessary comments, to be careful, stress-resistant, to have a good memory. The translator's task is "to mediate between the two cultures" (Brown, 2006, p. 664); playing a powerful role, a translator is "the only one in the communicative 'game' of translation who knows both the source and the target cultures" (ibid, p. 665). The given list is not full.

For a long time, traditional dictionaries have been the only source of linguistic information for a translator. Intensified processes of computerization and informatization have expanded the range of such tools to completely atypical sources of information. Today, diversification processes are observed in all areas of translation, which is facilitated, among other things, by expanding the sources of reference information that translators use to solve language problems. One such source is a computer database consisting of aligned texts in the original language and their translations.

Versatile computer programs and automated machine translation systems are not expensive nowadays (some are even free), they work fast, are ready to use, but the quality of the final product still requires the human mind to check and correct, as an intelligent and specially educated person can apply an individual approach and consider the cognitive and social context of the target audience. However, it does not mean that all those programs and systems are useless. Yes, they are useful when you know how to combine their results with personal contribution.

Our observation brings into the focus of researchers (translators inclusive) a new possibility which is corpora (corpora of texts) investigation. "A corpus of texts" is an expression known nowadays not only to professionals of a narrow specialty but also to a wide variety of scholars/researchers, translators, and university students. One of the simplest and easiest for understanding is the definition by McCarthy: "A corpus is a collection of texts, written or spoken, usually stored in a computer database" (McCarthy, 2004). A wider interpretation is given by Evans who writes that a corpus is "a principled collection of naturally occurring texts which are stored on a computer to permit investigation using special software" (Evans, 2004). But both wordings in no way outline the linguistic part of a corpus, the rules of its creation, criteria to be satisfied, or practical use. This paper aims at specifying the prerequisites for building the English-Ukrainian parallel corpus with further description of its practical use in Translation Studies.

Achievements of Corpus Linguistics and Prospects for Translation Studies in Ukraine

The active development of corpus linguistics began in the 1960s, and in Ukraine only at the beginning of the 21st century. However, despite such a relatively short period of time, corpus linguistics has become one of the most promising branches of applied linguistics. Data from existing corpora have found application in lexicography, translation studies, stylistics, forensic linguistics, language variation studies, sociolinguistics, language description, interpretation of literary texts, methods of teaching and learning a foreign language (Baker, 2006, p. 2–3). Corpus linguistics as a branch of applied linguistics "deals with determining the general principles of construction, processing and operating data of linguistic corpora (text corpora) using modern computer technology, developing methods for collecting actual linguistic facts – written and oral texts, and ways to preserve them and analyse" (Zhukovska, 2013, p. 9).

In modern corpus linguistics, there are two conventional areas of research:

1 theory and practice of building corpora: type of the corpus, its purpose, scope, parameterization of the subject area, representativeness, structuring and principles of selection of basic units, storage, etc.;

2 study of linguistic corpora, i.e. the study of language using corpus methods (Kopotev & Mustaioky, 2008, p. 12).

Thus, the object of corpus linguistics is the corpus of texts, both as a source material for linguistic research and as a result of activities in this area of research. *The corpus of texts*, in a broad sense, means "any collection of written or oral texts used for the purpose of language research", and in a narrower sense – "a collection of texts in electronic form representing a particular language" (Bobkova, 2014, p. 11).

The text corpus differs greatly from other collections of texts in electronic format, such as electronic libraries or archives with the following features: representativeness, authenticity, selectivity, balance, machine readability (Zhukovska, 2013, p. 55). The corpus is a clearly structured and annotated set of texts of a certain model (by genre, author, period of time, etc.).

There are different types of text corpora that can be used for translation. In general, three types of corpora are considered to be the most effective for translation purposes: parallel ("original, source language-texts in language A and their translated versions in language B" (Baker, 1995, p. 230)), comparable ("consist of two separate collections of texts in the same language: one corpus consists of original texts in the language in question and the other consists of translations in that language from a given source language or languages" (Baker, 1995, p. 234)), and multilingual ("sets of two or more monolingual corpora in different languages, built up either in the same or different institutions on the basis of similar design criteria" (Baker, 1995, p. 232)) corpora.

Parallel corpora (free access inclusive) are better represented on a global scale. For instance, in 2019 the whole issue of Studies in Corpus Linguistics (three parts with 301 pp.) was devoted to new resources and applications of parallel corpora in contrastive and translation studies (Doval et al., 2019). In Part 1 "Parallel Corpora: Background and Processing", Spanish researchers published their results related to current achievements and challenges of parallel corpora (Irene Doval, M. Teresa Sánchez Nieto), current practices in corpus-based translation studies based on comparable parallel corpora (Lidun Hareide), the potentials and limitations of parallel corpora use in translation research (Josep Marco), useful and usability of parallel corpora (Rosa Rabadán), innovations in parallel corpus alignment and retrieval (Martin Volk). Part 2 deals with creation, annotation and access of parallel corpora, namely InterCorp – a parallel corpus of 40 languages (Petr Čermák), Corpus PaGeS as a multifunctional resource for language learning, translation and cross-linguistic research (Irene Doval, Santiago Fernández Lanza, Tomás Jiménez Juliá, Elsa Liste Lamas, Barbara Lübke), EPTIC as a many-sided, multi-purpose corpus of EU parliament proceedings (Adriano Ferraresi, Silvia Bernardini), CLUVI, WordNet and SemCor – following the process of parallel corpora enrichment with multimedia and lexical semantics (Xavier Gómez Guinovart), MULTINOT – issues and challenges in discourse annotation (Julia Lavid López), PEST – a parallel electronic corpus of state treaties (Mikhail Mikhailov, Liia Santalahti, Julia Souma), The COVALT PAR_ES Corpus (EN/FR/DE>ES) – indexation and analysis of a parallel corpus with CQPweb (Teresa Molés-Cases, Ulrike Oster), P-ACTRES 2.0 – a parallel corpus for cross-linguistic research (Hugo Sanjurio-Gonsález, Marlén Izquierdo), Basque corpora – an overview and the extraction of multi-word expressions (Zuriñe Sanz-Villar). Part 3 describes various tools and applications of parallel corpora, for example, strategies in automatic building of bilingual dictionaries – with a pivot language and existing bilingual dictionaries + string similarity and cognate extraction (Pablo Gamallo), using distributional semantics to discover bilingual collocations (Marcos Garcia, Marcos García-Salido, Margarita Alonso-Ramos), normalization of abbreviations and shorthand forms in French text messages (Parijat Ghoshai). The results of the abovementioned researches can significantly facilitate the creation of multilingual/parallel corpora by those scholars who are new to a corpus linguistics, its opportunities and advantages it provides.

Currently, there are no free, ready to use parallel corpora in Ukraine albeit it is known that the Ukrainian Language and Information Fund of National Academy of Sciences (Luchik, 2017; Shyrokov, 2011) and several universities (among which Kyiv Mohyla Academy, Taras Shevchenko National University of Kyiv, Kyiv National Linguistic University, National Pedagogical Dragomanov University, Lviv Polytechnics) are working on building/developing various parallel corpora based on different language pairs and linguistic genres.

When working with the corpus, special search engines are used – corpus managers – which represent the search results in the form of a concordance, i.e., "a list in which the search unit is presented in a contextual environment from fragments of various texts and statistics such as wordform tokens, grammatical categories, features of compatibility, management, etc.” (Maslova, 2016, p. 279).

For example, **Table 1** presents the concordance of the results of the search for the term *a perpetrator* in the English-Ukrainian parallel legal corpus (a pilot corpus compiled by Matvieieva in 2020).

Table 1 Fragment of the English-Ukrainian parallel legal corpus with the term *a perpetrator* and its context in source text in English and its Ukrainian translation

Sentence number	Source language (English)	Target language (Ukrainian)
128	They typically should *order* **a perpetrator** *to vacate* the residence of the victim for a sufficient period of time and *prohibit* **the perpetrator** *from entering* the residence or *contacting* the victim.	Зазвичай вони повинні *зобов’язувати* **винного** *звільнити* місце проживання жертви протягом достатнього періоду часу та *забороняти* **йому** *в’їжджати* до місця проживання або *контактувати* з жертвою.
387	In terms of content, protection orders may *order* **the perpetrator** *to vacate* the family home, *stay* a specified distance away from the victim and her children (and other people if appropriate) and some specific places and prohibit **the perpetrator** *from contacting* the victim.	Щодо змісту, захисні приписи можуть *наказати* **кривднику** *звільнити* сімейний будинок, *перебувати* на певній відстані від потерпілої та її дітей (та інших людей, якщо це доречно) та деяких конкретних місць, а також *заборонити* **йому** *контактувати* з жертвою.

The corpus analysis procedure includes three steps: identification of language data with categorial analysis, establishment of the ratio of language data with statistical methods, and data mining (Zhukovska, 2013). Search possibilities of corpus managers include searching for specific word forms: word forms by lemmas, a group of word forms, word forms by a set of morphological features, and so on. Employment of the corpus approach allows not only to study lexical units, but also to obtain data on the frequency of word forms, of tokens, of grammatical categories, on the common use of lexical units, on specificity of their compatibility, on unit management, etc. The use of monolingual corpora also allows us to study the frequency of discrete language unit usage in a specific discourse, the frequency of a separate language unit meaning actualization/realization, the frequency of certain language unit usage depending on time, author, scope, etc.

However, the most useful for translators remain parallel and comparable corpora, which is determined by the practical opportunities they provide.

Parallel Corpus. Corpus Annotation. Parallel Treebank

The Encyclopaedia of Language and Linguistics defines a parallel corpus as “a collection of documents that exist in both the source language and the translation language, on which the system can be trained and tested” (Brown, 2006, p. 412).

The main idea of parallel corpora is to provide users not only with the contexts in one language but with parallel concordances enabling them to search a unit both in a source and in a corresponding target language sentence. According to Teubert, parallel corpora can be treated as “translation repositories” (Teubert, 2007, p. 128). This conception summarises and reveals the essence of this linguistic phenomenon, interpreting a parallel corpus as a set of original texts written in any source language, and translations of these source texts into one or more other languages.

In addition to all of the named characteristics of any corpus, the parallel corpus has its own specific features. First of all, to enable the prospective use of corpus texts for translation purposes, the criterion of the quality of translated texts, which make up the text array of any parallel corpus, becomes of crucial importance. The text arrays of the majority of the currently existing parallel corpora consist of the originals and translations of literary works.

Another peculiar feature of a parallel corpus is its alignment. The corpus texts should not just be collected, but aligned: separate fragments of the source text must coincide with the corresponding fragments of the target text. The base of the parallel corpus should be compiled and organized in such a way as to provide a direct user access to two (several) subcorpora at the same time, that is, the system must simultaneously display ordered units of the source text and their translation equivalents, aligned by paragraphs and sentences and designed in accordance with linguistic and extra-linguistic information applied by the compilers, such as a part of speech, a type of grammatical form, a syntactic function, a type of links in a word combination, and the like. Such requirements are caused by the need to establish clear algorithmic standards for computer processing of language material with its subsequent use to create software products and databases that will become effective tools in the hands of translators. Aligned at paragraph and sentence levels, corpus texts need further annotation and markup. The depth of corpus annotation is determined by the purpose and tasks of every corpus.

Preceding our discussion of *corpus annotation*, it is useful to stress that the latter one (i.e. corpus annotation) differs from its contiguous term *markup*. *Annotation* can mean 5 things: 1) a critical or explanatory commentary or analysis, 2) a comment added to a text, 3) the process of writing such comment or commentary, 4) (computing) metadata added to a document or program, 5) (genetics) information relating to the genetic structure of sequences of bases" (Annotation vs Markup, 2021), first 4 of which can be applied to corpus annotation as "the product of the human mind's understanding of the text" (Leech, 1997, p. 2). Meanwhile, *markup* is primarily interpreted as "the notation that is used to indicate how text should be displayed" (Annotation vs Markup, 2021), or a combination of each text structure + the components of a corpus.

Corpora are used to extract/decode linguistic information. But to extract this information from a corpus it must first be encoded in it. Such encoding or "adding interpretative, linguistic information to an electronic corpus of spoken and/or written language data" (Leech, 1997, p. 2) is called corpus annotation. In case we mean both interpretative linguistic analysis + textual/contextual information, then it is a broad sense of corpus annotation or corpus annotation + markup. Encoding solely linguistic analysis (covering part-of-speech tagging + syntactic parsing) is usually referred to as a narrow sense of corpus annotation or just corpus annotation.

As soon as availability of corpus annotation facilitates the extraction of necessary data from and speeds up the work with corpus (Kübler & Zinsmeister, 2015), McEnery (2003), following Leech (1997), observes at least 4 its advantages, namely:

1 easier extraction of information from annotated corpora;

2 reusability of annotated corpora;

3 multifunctionality of such corpora;

4 explicity of a linguistic analysis, to be scrutinized and criticized.

Despite several disadvantages / criticisms, available in some papers (e.g., Sinclair, 1991; Hunston, 2002), that mostly concern clustering of corpora, imposing a linguistic analysis on the researchers, lack of accuracy and consistency of corpus annotation, it should be underlined that lack of annotation or failure to do it only signals of difficulty or even impossibility to annotate a corpus. We may argue that it is indeed the researcher's task to interpret the results of corpus annotation implementation for a certain collection of texts and suggest the corrections or prove them to be verified.

The most commonly used types of corpus annotation are based on the existence of different language levels (phonological, morphological, lexical/semantic, syntactic, textual/discoursal), as well as 3 major constituents of any language sign (a form, a meaning, a function). In this respect there are 8 possible types of corpus annotation: phonetic/phonemic (syllable boundaries) and/or prosodic; morphological (stems and affixes); lexical (parts-of-speech (POS) tagging and lemmas/lemmatization); semantic (semantic fields); syntactic (parsing, treebanking/bracketing); coreference (anaphoric relations); pragmatic (speech acts); stylistic (speech and thought presentation).

Each type of annotation requires special knowledge to be encoded. Some of these types are more common (e.g. lexical, syntactic), some are rare (e.g. coreference, pragmatic). A standalone annotation type is represented by alignment, as it is used in multilingual corpora. Parallel corpora cannot be built without alignment that enables placing and extracting corresponding sentences / paragraphs of source and target texts in one space.

Experience with unmarked parallel corpora has demonstrated the need to apply deeper annotation and build parallel treebanks – "a particular kind of annotated corpus where each sentence is mapped to a special type of graph, a tree which represents its syntactic structure" (Volk et al., 2017, p. 7).

Recently, there have been many attempts to annotate parallel corpora syntactically to construct parallel treebanks. Such treebanks became a new paradigm in parallel corpora creation and use. The basic idea of parallel treebanks is to provide a translator or a researcher with the lexically and syntactically annotated parallel corpus – a very special tool for information extraction and analysis. “Parallel treebanks provide an advantage to ordinary parallel corpora in that they can be used for inducing reliable structural correspondence between the languages in question” (Brown, 2006, p. 112).

For today, building parallel treebanks is a labour-intensive and time-consuming process. A parallel treebank can be created automatically, but without manual correction it often results in errors and mismatches; therefore, segmentation and tagging still need reviewing and post-editing (Green et al., 2013). Applying various translation strategies and using numerous translation shifts, translators make automatic word alignment hardly possible. At the same time there is a conception of word co-occurrence frequencies applied in corpus theory recently, promising to solve this problem to some extent.

It is a well-known fact that to create parallel corpora, special software is required, which allows to align paragraphs and sentences of a source text with their translated versions into target language(s). Besides, automatic processing and annotation of texts for parallel corpora have specific features.

Opportunities of Programming Languages for Parallel Corpora Creation

Software development is a complex and difficult task because it is multifactorial. Developers with little experience frequently tend to apply the principle of ‘golden hammer’. This principle deals with using one and the same technology that they know well for all the tasks. However, when developing information systems that have high requirements to the quality of software products, it is necessary to take into account the peculiarities of programming languages, databases, and approaches to software product development.

Ten years ago, Desktop Application was already the most popular trend in software development. A good example here can be tlCorpus Concordance Software (see **Fig. 1**). This program can be installed on the computer running by Windows and MacOS operational systems. On the one hand, this opportunity is very attractive – to work with this program you need no access to the Internet, on purchasing such a program it becomes your property. Running MacOS, Delphi, C++, C#, Java, and Objective C were most often used to write such applications. Many systems have survived till this day only in the form of a desktop version, i.e. Desktop Application.

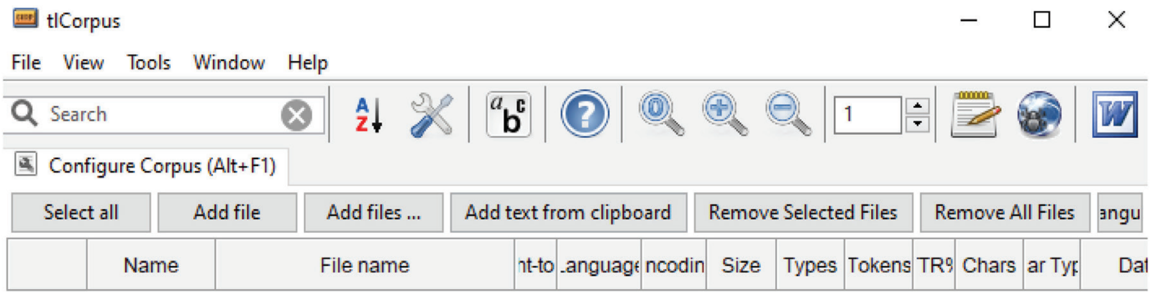

Fig. 1 The main window of tlCorpus Concordance Software

Today the most popular area of software development is the creation of services. For instance, Microsoft long ago created Office 360, Google provides access to its office software counterpart. This model of software sales allows companies not to think about checking compliance with the license agreement, and enables clients to use the purchased software on several devices. Undoubtedly, with the development of the Internet, this opportunity has become especially relevant.

This trend is also relevant to the development of systems for work with various corpora. Firstly, we need to understand that it is not sufficient just to collect texts. Texts should be marked and annotated, with statistics collected. In case of parallel corpora, it is necessary to make alignment. It is obvious that the tasks of obtaining data, data access and other ordinary operations can be best done using compiled and strictly typed programming languages, such as C#, C++, and Java. C++ has a lot of memory management feasibilities. One of the features of C++ is the opportunity to access RAM (random access memory) by passing the cash of the processor (CPU – central processor unit). There is lack of such possibility in the C# programming language. It is possible that in the future C++ will add C# and Java. Nevertheless, the .NET Core and .NET 5 platforms can successfully compete with C++ as a programming language for development of the system core. .NET Core and .NET 5 web applications can deploy to servers running both Windows and Linux. In the case of a server on the Linux platform, we have an advantage in conventionally free licenses, which is definitely an advantage for startups. The Java programming language is much weaker in performance than C++, C, and C#. The Java programming language uses Java Virtual Machine. This intermediate layer is the cause of the performance issue. Java cannot deal with RAM without using the CPU cache. C and C++ languages do not use any virtual machines. NET 6 contains a lot of features of C++ for improved performance (https://codersera.com/blog/reasons-why-dot-net-is-better-than-java/).

An important part of work with corpora is their practical use. Work with textual data requires mathematical and algorithmic training. For a long time, scientists / scholars have been looking for a kind of programming language that could be easy in learning and simple in using. The Python programming language corresponds to these principles. At the moment, it is considered the number 1 programming language in working with data. An unexpected advantage of this programming language was dynamic typing. This programming language is based on an interpreter. Python is currently being studied at the majority of US universities (for example, Stacksocial, 2014), where it is very popular (Tagliaferri, 2022). Python is used to build Facebook neural networks in the PyTorch library. Google applies it in its BERT neural network (BERT, 2019) and in the open source library for machine learning TensorFlow (Why TensorFlow, 2019).

After selecting the programming languages, it is necessary to determine the databases that the system will use to store data. In this case, it is possible to use relational and non-relational (NoSQL – Not only SQL) DBMS (database management system). Oracle, MS SQL Server, MySQL, MariaDB, and PostgreSQL are the most popular relational databases. Among them, recently PostgreSQL has been gaining the most popularity due to its ability to create its own data types and also because of being free of charge. The first NoSQL databases were created much earlier than relational ones, but gained considerable popularity only after 2009, when the term NoSQL was actually proposed. NoSQL databases in turn consist of document-oriented store, key-value store, graph, and others. The most well-known NoSQL database is the document-oriented MongoDB. It is often used in the development of Internet sites. One of the most famous sites that started using it was the New York Times. When developing a system for work with corpora, it is advisable to use graph databases. One of the most famous examples of graph databases is neo4J. This key-value database can be used to store links, and it favourably differs in speed. A popular key-value database is Redis.

The last, but not less important issue is use of technologies of data visualization. Web-based systems visualize data in a browser. During the last 5 years, the leaders among frontend development technologies are React from Facebook and Angular from Google. Making choice between these two technologies is quite a challenge. Both technologies have all the necessary functionality to develop systems for work with corpora. Thus, to develop a large, multifunctional system for work with corpora, it is necessary to take exactly those technologies that are suitable for each specific task.

To make a good system it is not sufficient to have a set of technologies. The CAP theorem (see **Fig. 2**) states that a system cannot have three properties at once: data consistency, availability, and partition resistance. That is why developers must sacrifice one of these properties in the process of designing information systems.

One great example of a web-oriented system for work with corpora is the Sketch Engine (see **Fig. 3**). Two programming languages were used for its development: C++ and Python. As already mentioned, their combination gives the optimum result in terms of time spent on the development and quality of the software product. It is also known that this platform uses jQuery to visualize data. The choice in favour of this development technology can be justified by several factors, the historical one and not very big amount of work on data visualization compared to the tasks used on the server side. The Sketch Engine was first released in 2003, when Angular and React were not yet available.

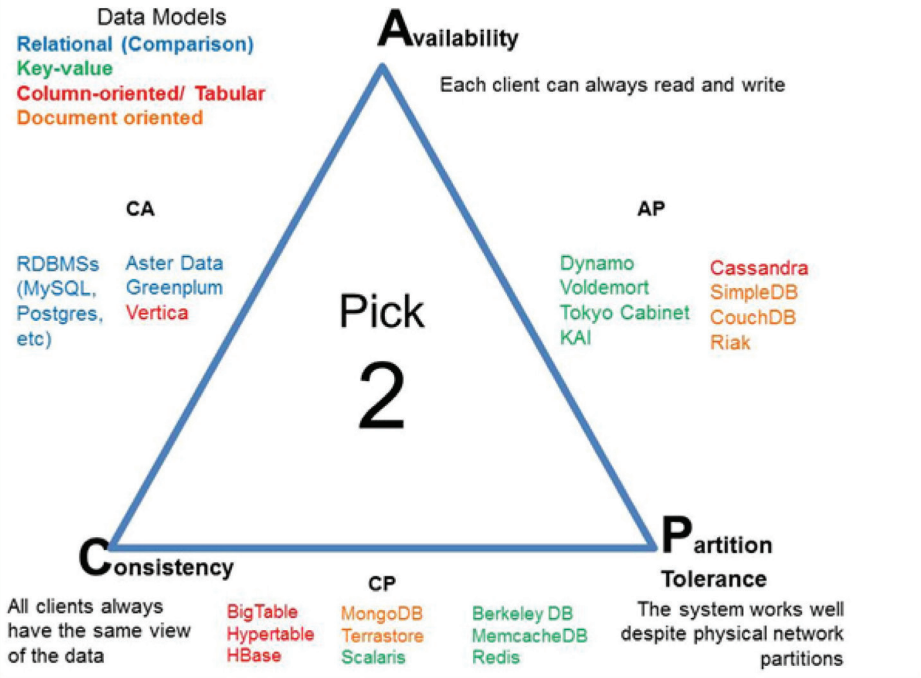

Fig. 2 Visualization of CAP-theorem (Perwej, 2017)

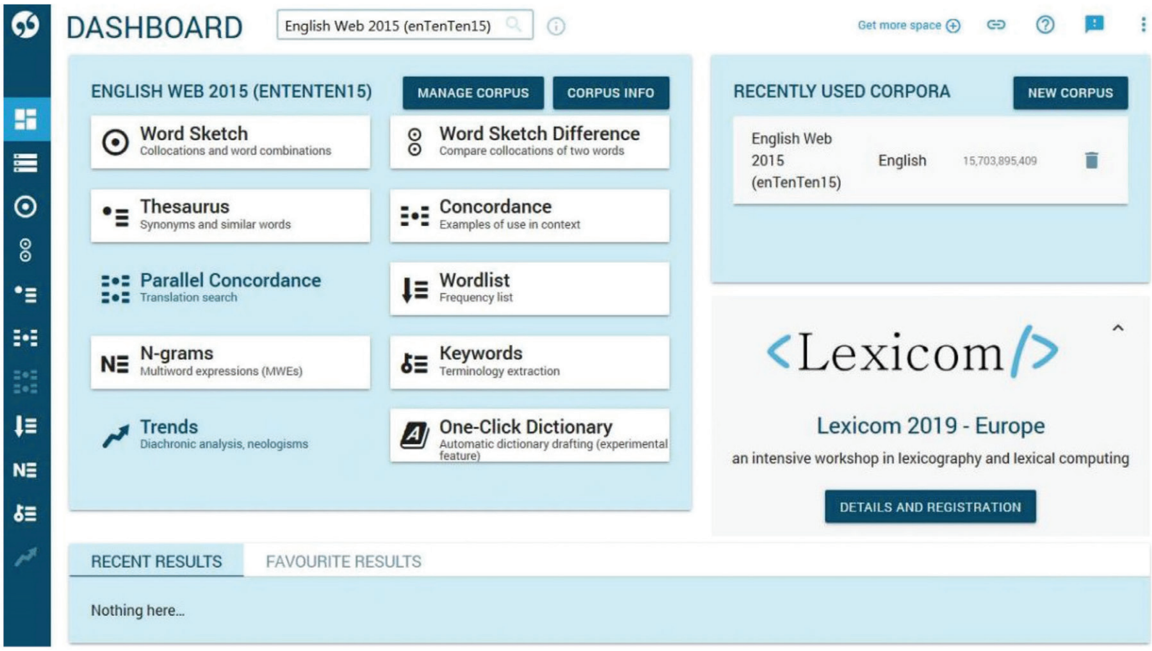
DASHBOARD
English Web 2015 (enTenTen15)
Get more space
ENGLISH WEB 2015 (ENTENTEN15)
MANAGE CORPUS
CORPUS INFO
Word Sketch
Collocations and word combinations
Word Sketch Difference
Compare collocations of two words
Thesaurus
Synonyms and similar words
Concordance
Examples of use in context
Parallel Concordance
Translation search
Wordlist
Frequency list
N-grams
Multiword expressions (MWEs)
Keywords
Terminology extraction
Trends
Diachronic analysis, neologisms
One-Click Dictionary
Automatic dictionary drafting (experimental feature)
RECENTLY USED CORPORA
NEW CORPUS
English Web 2015 (enTenTen15)
English
15,703,895,409
<Lexicom/>
Lexicom 2019 - Europe
an intensive workshop in lexicography and lexical computing
DETAILS AND REGISTRATION
RECENT RESULTS
FAVOURITE RESULTS
Nothing here...

Fig. 3 The main window of Sketch Engine (https://www.sketchengine.eu/)

Over a long period of development, the system has received many interesting functions which are recommended to be borrowed by other system developers to work with corpora. It is worth noting that the system interface is made in blue and white. It is in these colours that the social network Facebook is made. This colour scheme combined with an intuitive interface provided a user-friendly interface.

There are several ways to download texts to Sketch Engine. The way of downloading texts from Internet sites is noteworthy: when you enter the main page of the site, the Sketch Engine itself 'walks through' the pages. Then, with the help of Bing, it clears the ads and saves the text. The developers of the system obviously had technical reasons to make delays between loading each page. As the system can be used by a large number of people at the same time, there could be problems with a lack of resources on the server side when loading multiple pages in parallel. That is why this solution made it easy to download texts from sites and the servers will not have problems with lack of resources. It is also possible to manually enter texts from the keyboard and download text files. This wide functionality for replenishment of materials of the corpus is by all means a big advantage of this system.

After downloading texts Sketch Engine analyses and processes them. Due to this, users of the platform have the opportunity to receive statistics and data on regular expressions; as well as to use Corpus Query Language and Lexonomy Dictionary Writing System.

There is one task that is almost impossible to automate in parallel corpora – text alignment. This task is extraordinary, because, as stated before, a sentence from one language can be rendered with several sentences in another language and vice versa. Therefore, to create parallel corpora, it is necessary to perform additional text preparation. The Sketch Engine system offers markup/annotating with Excel. This is a convenient option for marking/annotating texts. Among other options for markup/annotating text alignment, it is possible to select markup/annotating directly in the browser, but it requires significant efforts on the part of the frontend developer. Xml or json format can satisfy this need.

Sketch Engine lacks artificial intelligence techniques for further word processing. Nowadays in Ukraine, parallel corpora are created manually. There is probability that in the future developers will add a neural network that will perform text alignment in parallel corpora instead of system users. This improvement would obviously make the job easier.

Parallel Corpora for Translation Purposes

From the standpoint of theoretical translation studies, researchers are interested in corpora options to study translation processes (Alves & Vale, 2017; Baker, 1995; Bowker, 1998; Hareide, 2019; Kruger, 2011, etc.), taking into account the cognitive nature of the translator, i.e., the focus is on analytical work with existing corpora. From a practical point of view, there are many more similar issues, among which the most crucial are those that concern the effective use of corpora to create new texts and corpora of the highest quality.

Today the effectiveness and importance of parallel corpora with aligned and annotated texts in two or more languages for translation is out of question. This can be explained by a number of reasons, including the updated and actual practical purpose of such corpora, as they can be used for:

1. identification of typical translation techniques and transformations of document connotations;
2. training statistical schemes for automatic translation;
3. creation of monolingual and multilingual dictionaries;
4. selection of possible equivalents (Demianchuk, 2016, p. 106).

Parallel corpora offer textual databases as a source of translation equivalents for various language units (words, word combinations of various types, sentences) and of the systematic study of language norms. Obviously, a large amount of empirical material, living language units functioning in real language practice and reflecting real language environment allow the translator to check and verify translation options and solutions, using a proven evidence base. Corpus studies allow to rule out possible subjective factors in the research and get as close as possible to the objective study of language, as they are based "mainly on actual "live" language material, but not on linguistic intuition and introspection" (Zhukovska, 2013, p. 5). Parallel corpora provide translators with the information on the methods used by other translators in practice. Employment of such resources can bring the effectiveness of translation work to a qualitatively new level, "completely change the results of vocabulary work and rearrange the lists of equivalents presented in traditional bilingual and multilingual dictionaries" (Matvieieva et al., 2021, p. 474).

Such corpus-based technologies provide useful tools for translators, e.g.:

1. choosing the best translation equivalent;
2. checking the contextual use of a particular unit or pattern in a natural, authentic environment with wide language data;
3. forming the set of translation models / equivalents;
4. compiling multilingual specialized dictionaries for particular translation needs;
5. improving translation quality and accuracy;
6. both creating new and improving the current approaches to machine translation technology (Lemish et al., 2020, p. 255).

Together with linguistic information, the data of the text corpus includes a large amount of meta-information, such as data of the author and translator of the text, its genre, dates of writing and translating, etc., which allows the translator to specify the functioning and translation of specific linguistic units in a specific language work. Parallel corpus is a crucial technology for a translator, as "the main task of the translator is to reproduce the original text in the target language with maximum preservation of functional, semantic, and figurative components of the source text. Any text is a complex system of explicit and implicit meanings, which significantly complicates the translation. The implicitness of meanings and the multiplicity of interpretations require the translator to conduct an in-depth analysis of the original text and choose a translation strategy for each specific text" (Matvieieva, 2021, p. 469–470). The fulfilment of this task can be fully ensured by involving a parallel corpus in the translation process.

Among other uses, parallel corpora are employed for automatic extraction of words and collocations, their mathematical analysis and decision making regarding the most adequate translatable correspondence, equivalent scenario of each discourse. Lemish states that "corpus-based processing of linguistic material gives the most reliable results and impartial conclusions about the state and issues of the studied languages both at a separate stage of their development and in comparison" (Lemish, 2017, p. 140). Besides, "since the corpus-based studies are built primarily on an empirical approach to the analysis of language material, this allows us to achieve maximum objectivity in language learning, excluding the subjective views of the researcher" (Matvieieva, 2020, p. 168).

One more "area of application of corpus technology is the study of translation norms in certain socio-cultural and historical contexts, as well as the nature of the compliance of texts with regulatory requirements, or deviations from them" (Avdeiev, 2019, p. 143). Besides, data from parallel corpora can be used in translation lexicography, comparative lexico-grammatical works, studying the theory and practice of translation, teaching languages, as well as for the development of machine translation systems (Tyshchenko-Monastyrska et al., 2011).

Recently, English dictionaries based on corpora and corpus data have appeared. One of the first dictionaries is the *Collins COBUILD English Language Dictionary* (Collins), which was first published in 1987 and "this was the first dictionary to be based in full on research carried out using the Collins Corpus, a database of over 4.5 billion words of written and spoken English which provides evidence of how English is really used" (Collins). Accordingly, the authors took into account many factors that had not been previously considered when compiling dictionaries. Nowadays, "there are software tools and other methods for efficiently extracting the information that corpora hold. Modern corpus search software gives an overall picture of a word by displaying it on the screen in a way that shows how it combines with other words. It shows the search word together with its collocates – the words it combines with most frequently – and tells you how significant these combinations are. Each collocational or grammatical chunk displayed can be expanded, allowing you to examine it in more detail if necessary" (Collins COBUILD English Language Dictionary). This dictionary has completely changed the approach to compiling dictionaries and introduced a new type of corpus dictionaries and reference materials for those who study English. This approach is now used in the compilation of many English dictionaries (e.g., Longman, Oxford, etc.). However, a significant number of dictionaries today are based on the material of their own, closed to public access corpora

(Zhukovska, 2013, p. 107). In addition, these dictionaries are constantly updated to reflect the latest trends in language and to provide the most up-to-date information on the usability and coherence of words. Corpora are also often referred to when compiling reference books – grammars, educational dictionaries, and reference books, which include not only vocabulary but also grammatical information (Zhukovska, 2013, p. 107). There are currently several corpus-based English grammars (e.g., Collins COBUILD English Grammar, 1990; Longman Grammar of Spoken and Written English, 1999, etc.).

Concluding Remarks

Corpus-based translation analysis is extremely relevant to identifying translation solutions that can only be explored on the basis of translation products. It is stipulated by frequent absence of dictionary equivalents in many contexts and ready evidence of possible translation variants in parallel corpora that provide the usage of a language unit in a wide range of contexts.

The combination of the efforts of professionals in the field of linguistics and programming provides practicing translators with unique tools for processing, analysing and using large volumes of language material in different languages. One of such tools is the parallel corpora technology. Various programming languages, databases, and software used for the information systems development enable improving automated proceeding of language material that helps translators and brings us closer to solving machine translation problems.

The active application of the corpus approach in linguistics and the use of the achievements of corpus linguistics in language theory and translation practice warrant complicated solutions for a number of issues that have not received sufficient attention from researchers to date. Among such issues is the problem of creating and using parallel corpora with Ukrainian as one of the corpus languages. Unfortunately, work on the creation of an English-Ukrainian parallel corpus is at an early stage in Ukraine. Experience of the researchers from other countries and support of knowledgeable programmers will definitely help realize such task, which in its turn will significantly contribute into both teaching / learning foreign languages and developing / improving translation skills.

References

1 Alves F., & Vale D. C. (2017). On drafting and revision in translation: A corpus linguistics oriented analysis of translation process data. Annotation, exploitation and evaluation of parallel corpora. Berlin: Language Science Press, 81-101.

2 Annotation vs Markup - What's the difference? Retrieved March 15, 2021, from https://wikidiff.com/markup/annotation/.

3 Avdeiev, A. A. (2019). Tehnologiia parallelnyh korpusov tekstov i ieio ispolzovanie v protsesse obucheniia perevodu [Technology of Parallel Test Corpora and its Use in Translation Teaching]. Nauchnyi zhurnal "Sovremennyie lingvisticheskie i metodiko-didakticheskie issledovaniia" [Scientific Journal "Modern Linguistic, Methodical, and Didactic Studies"], 3, 140-151. [In Russian].

4 Baker, M. (1995). Corpora in Translation Studies. An Overview and Suggestions for Future Research. Target, 7 (2), 223-243. https://doi.org/10.1075/target.7.2.03bak

5 Baker, P., 2006. Using Corpora in Discourse Analysis. London: Continuum. https://doi.org/10.5040/9781350933996

6 BERT - state-of-the-art yazykovaia model for 104 languages [BERT - state-of-the-art the language models for 104 languages] (2019). Retrieved March 15, 2021, from https://habr.com/ru/post/436878/.

7 Bobkova, T.V. (2014). Korpus tekstiv: osnovni aspekty vyznachennia [Corpus of Texts: Basic Approaches to its Defining]. Naukovyi visnyk kafedry Yunesko KNLU. Seriia Filolohiia, Pedahohika, Psykholohiia [Scientific Messenger of the UNESCO Department of Kyiv National Linguistic University. Philology. Pedagogy. Psychology], 29, 11-20. [In Ukrainian].

8 Bowker, L. (1998). Using Specialized Monolingual Native-Language Corpora as a Translation Resource: A Pilot Study. Meta: Translators' Journal, 43(4), 631-651. https://doi.org/10.7202/002134ar

9 Brown, K. (Ed.) (2006). Encyclopaedia of Language & Linguistics. Boston: Elsevier.

10 Collins COBUILD English Language Dictionary. Retrieved February 12, 2022, from https://www.collinsdictionary.com.

11 Demianchuk, Yu. I. (2016). Riznovydy korpusu tekstiv u protsesi perekladu dokumentiv ofitsiino-dilovoho styliu [Text Corpora Types in Official Documents Translation]. Naukovyi visnyk DDPU imeni I. Franka. Seriia

"Filolohichni nauky". Movoznavstvo [Research Journal of Drohobych Ivan Franko State Pedagogical University. Series "Philology" (Linguistics)], 5, t. 1, 104-107. [In Ukrainian].

12 Doval, I., & Nieto, M. T. S. (Eds.). (2019). Parallel corpora for contrastive and translation studies: New resources and applications (Vol. 90). John Benjamins Publishing Company. https://doi.org/10.1075/scl.90

13 Evans, D. (2004). Corpus building and investigation for the Humanities: An on-line information pack about corpus investigation techniques for the Humanities. Retrieved March 15, 2021, from https://www.birmingham.ac.uk/research/activity/corpus/publications/introduction-corpus-investigative-techniques.aspx/

14 Green, S., Heer, J., Manning, C. D. (2013). The efficacy of human post-editing for language translation (pp. 439-448). Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Retrieved March 15, 2021, from https://dl.acm.org/doi/10.1145/2470654.2470718.

15 Hareide, L. (2019). Comparable parallel corpora: A critical review of current practices in corpus-based translation studies. Parallel corpora for Contrastive and Translation Studies: New resources and applications (pp. 19-38). John Benjamins. https://doi.org/10.1075/scl.90.02har

16 Hunston, S. (2002). Corpora in Applied Linguistics. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139524773

17 Kopotev, M. V. & Mustaioky, A. (2008). Sovremennaia korpusnaia rusistika [Modern Corpus Russian Studies]. Instrumentarii rusistiki: korpusnye podkhody [Tools of Russian Studies: Corpus Approaches]. Slavica Helsingiensia Series, 34, 7-24. [In Russian].

18 Kruger, A., Wallmach, K., & Munday, J. (ed.) (2011). Corpus-based Translation Studies. Research and Applications. London and New York: Bloomsbury.

19 Kübler, S., & Zinsmeister, H. (2015). Corpus Linguistics and Linguistically Annotated Corpora. London: Bloomsbury.

20 Leech, G. (1997). Introducing corpus annotation. In R. Garside, G. Leech, & T. McEnery (Eds.), Corpus Annotation Linguistic Information from Computer Text Corpora. London: Longman, 1-18.

21 Lemish, N. Ye. (2017). Korpusna metarozmitka spetsialnyh tekstiv z linhvoantropohenezu [Corpus Mark-up of Specialized Texts on Linguoanthropogenesis]. Naukovyi chasopys NPU im. M.P. Drahomanova. Seriia 9 "Suchasni tendentsii rozvytku mov" [Scientific Journal of National Pedagogical Dragomanov University. Series 9. Current Trends in Language Development], 15, 139-152. [In Ukrainian].

22 Lemish, N. Ye., Aleksieieva, O. M., Denysova, S. P., Matvieieva, S. A., & Zernetska, A. A. (2020). Linguistic corpora technology as a didactic tool in training future translators. Information Technologies and Learning Tools, 79(5), 242-259. https://doi.org/10.33407/itlt.v79i5.3626

23 Luchik, A., & Ostapova, I. (2017). Syntahmatychna parametryzatsiia ekvivalentiv slova u paradyhmi korpusnoi linhvistyky [Syntagmatic Parametrization of word equivalents in Corpus Linguistics Paradigm]. Human. Computer. Communication (pp. 33-37). Lviv. [in Ukrainian].

24 Maslova, T. B. (2016). Instrumenty korpusnoi linhvistyky u navchanni i doslidzhenni inozemnykh mov [Corpus Linguistics Tools in Foreign Languages Teaching and Studying]. Smart-osvita: resursy ta perspektyvy: materialy II Mizhnarodnoi naukovo-metodychnoi konferentsii [Smart-education: resources and prospects: papers of II International Scientific and Methodical Conference]. (pp. 277-280). Kyiv: Kyivskyi natsionalnyi torhovelno-ekonomichnyi universytet. [In Ukrainian].

25 Matvieieva, S.A. (2020). Corpus-driven studies and corpus-based translation: pragmatic potential. International Engineering Journal for Research & Development, 5(4), 167-170.

26 Matvieieva, S., Lemish, N., & Zernetska, A. (2021). English cognitive verbs and their translation into Ukrainian: A corpus-based approach. SLAVIA časopis pro slovanskou filologii, 90(4), 453-476.

27 McCarthy, M. (2004). Touchstone. From Corpus to Course Book. Cambridge: Cambridge University Press. What is Corpus. Retrieved March 15, 2021, from https://21centurytext.wordpress.com/home-2/special-section-window-to-corpus/what-is-corpus/.

28 McEnery, A. (2003). Corpus linguistics. In R. Mitkov (Ed.), The Oxford Handbook of Computational Linguistics. Oxford: Oxford University Press, (pp. 448-463).

29 Perwej, Yu. (2017). International transaction of electrical and computer engineers system. An Experiential Study of the Big Data, 4(1), 14-25.

30 Shyrokov, V. A. (2011). Kompiuterna leksykografiia [Computational Lexicography]. Kyiv: Naukova dumka. [in Ukrainian].

31 Sinclair, J. M. (1991). Corpus, Concordance, Collocation. Oxford: Oxford University Press.

32 Stacksocial (July 9, 2014). The most popular coding language at top US universities. Retrieved March 15, 2021, from http://blog.stacksocial.com/popular-coding-language/.

33 Tagliaferri, C. (January 15, 2022). Programming languages ranking: Top 9 in 2022. Retrieved March 15, 2021, from https://distantjob.com/blog/programming-languages-rank.

34 Teubert, W. (2007). Text Corpora and Multilingual Lexicography. Benjamins Current Topics. https://doi.org/10.1075/bct.8

35 Tyshchenko-Monastyrska, O. O., Shvedova, M. O., & Sichinava, D. V. (2011). Paralelni ukrainsko-rosiiskyi ta rosiisko-ukrainskyi korpusy [Parallel Ukrainian-Russian and Russian-Ukrainian Corpora]. Leksykohrafichnyi biuleten [Lexicographic Bulletin], 20, 35-38. [In Ukrainian].

36 Volk, M., Marek, T., Samuelsson, Y. (2017). Building and querying parallel treebanks. In Silvia Hansen-Schirra, Stella Neumann & Oliver Čulo (Eds.), Annotation, exploitation and evaluation of parallel corpora, 9-35, Berlin: Language Science Press.

37 Why TensorFlow (2019). Retrieved March 15, 2022, from https://www.tensorflow.org/.

38 Zhukovska, V.V. (2013). Vstup do korpusnoi linhvistyky: navchalnyi posibnyk [Introduction to corpus linguistics: a textbook]. Zhytomyr: Vyd-vo ZhDU im. I. Franka. [In Ukrainian].

Summary

Svitlana A. Matvieieva, Nataliya Ye. Lemish, Alla A. Zernetska, Volodymyr O. Babych, Maryna A. Torgovets. Anglų-ukrainiečių kalbų lygiagrečiojo tekstyno kūrimo ir praktinio naudojimo vertimo studijose prielaidos

Nuolatinė profesionalių vertėjų paklausa lemia tyrėjų ir programuotojų bandymus kurti ir pasiūlyti patikimas priemones tiek vertimo kokybei gerinti, tiek vertėjų darbui palengvinti. Pastarajame dešimtmetyje lygiagretieji ir lyginamieji tekstynai atsidūrė ir Ukrainos mokslininkų dėmesio centre. Šio tyrimo tikslas – išryškinti anglų-ukrainiečių kalbų lygiagrečiojo tekstyno kūrimo prielaidas ir aprašyti jo taikymą vertimo studijose. Lygiagretusis tekstynas, kaip atskiras kalbos tekstynų tipas, negali būti sudarytas be lygiagretinimo, leidžiančio vienoje erdvėje sudėti ir ištraukti atitinkamus šaltinio ir tikslinio teksto sakinius / pastraipas. Norint sukurti lygiagrečiuosius tekstynus, būtina papildomai paruošti tekstą. „Sketch Engine“ (internetinė darbo su tekstynais platforma) gali pasiūlyti sprendimą anotavimui su „Excel“ programa. Tačiau „Sketch Engine“ trūksta dirbtinio intelekto metodų tolesniam teksto apdorojimui. Yra tikimybė, kad neuroninio tinklo panaudojimas ateityje leis tekstą lygiagretinti pačiuose tekstynuose. Duomenys iš lygiagrečiųjų tekstynų gali būti naudojami vertimo leksikografijos tyrimuose, lyginamuosiuose leksikologijos ir gramatikos darbuose, vertimo teorijos ir praktikos studijose, kalbų mokyme, mašininio vertimo sistemų kūrime. Tekstynų pagrindu atliekama vertimo analizė yra labai svarbi, siekiant identifikuoti vertimo sprendinius, kuriuos galima nustatyti tik remiantis vertimo produktais. Tai sąlygoja gana dažnas žodyno atitikmenų nebuvimas daugelyje kontekstų ir įrodymai apie galimus vertimo variantus lygiagrečiuosiuose tekstynuose, kurie parodo kalbos vienetų realią vartoseną.

About the Authors

SVITLANA A. MATVIEIEVA

Prof. dr., Department of Applied Language Studies, Comparative Linguistics, and Translation, National Pedagogical Dragomanov University, Kyiv, Ukraine

Research interests
Corpus linguistics, translation, comparative linguistics, cognitive linguistics

Address
Pyrohova st. 9, 01601 Kyiv, Ukraine

E-mail s.a.matvyeyeva@npu.edu.ua

Orcid iD
0000-0002-8357-9366

NATALIYA YE. LEMISH

Prof. dr., Department of Applied Language Studies, Comparative Linguistics, and Translation, National Pedagogical Dragomanov University, Kyiv, Ukraine

Research interests
Corpus linguistics, translation studies, comparative linguistics, intercultural communication

Address
Pyrohova st. 9, 01601 Kyiv, Ukraine

E-mail n.ye.lemish@npu.edu.ua

Orcid iD
0000-0001-5321-4705

ALLA A. ZERNETSKA

Prof. dr., Department of Applied Language Studies, Comparative Linguistics, and Translation, Dean of Foreign Philology Faculty, National Pedagogical Dragomanov University, Kyiv, Ukraine

Research interests
Applied linguistics, foreign language teaching technology, neurolinguistics

Address
Pyrohova st. 9, 01601 Kyiv, Ukraine

E-mail a.a.zernetska@npu.edu.ua

Orcid iD
0000-0002-8500-5884

VOLODYMYR O. BABYCH

Lecturer, Department of Information Technology, European IT School, International European University, Kyiv, Ukraine

Research interests
Databases, artificial intelligence, corpus linguistics

Address
Academika Glushkova prosp. 42 V, 03187 Kyiv, Ukraine

E-mail
volodymyrbabych64@gmail.com

Orcid iD
0000-0001-8788-9225

MARYNA A. TORGOVETS

Assistant prof. dr., Department of Social and Humanitarian Education, Private Higher Educational Institution “Institute for Ecology Economy and Law”, Kyiv, Ukraine

Research interests
Foreign language acquisition, translation, language training methods

Address
Turhenevska st. 11, 01054 Kyiv, Ukraine

E-mail m_maryna@ukr.net

Orcid iD
0000-0002-5878-5387