Yan Kapranov,
Doctor of Sciences (Philology), Associate Professor, Professor
Kyiv National Linguistics University (Ukraine);
Associate Professor at the University of Economics and Human Sciences
in Warsaw (Poland)

RESOURCES AND OPPORTUNITIES OF OPUS IN TRANSLATION STUDIES: SPECIAL REFERENCE TO EDUCATIONAL AND SCIENTIFIC ACTIVITIES

It should be noted that requirements of the ISO 18587:2017 EU Standard were involved, emphasizing pre-editing, inter-editing, and post-editing of the machine translation. The students are provided with innovative EU translation technologies based on the texts belonging to different discourses. This confirms the hypothesis that EU innovative translation technologies help any translator save time, money, and in some cases, health. These tools allow everyone to learn about Europe, its economy, history, culture, education, politics, and values, especially when you do not speak European languages.

In this context, the concept of multilingualism/multiculturalism is actualized, which is more philosophical, although it is simultaneously related to both language and culture.

In general, the variety of languages of the modern EU is now an objective reality. In creating the EU, a conscious decision was made not to introduce the primary language but to preserve different languages and provide them equal rights. As we can see, the EU was conceived as multilingual or multicultural, and there is even the Commissioner for Education, Training, Culture, and Multilingualism.

At the same time, multilingualism/multiculturalism in the EU complicates interactions between ordinary people and academic institutions, commercial structures, and states in general. In addition, protecting cultural heritage and preserving language diversity requires constant and significant investment because European institutions spend a lot of current budgets on translation and interpretation services. Due to the fact that the EU considers language diversity an integral part of cultural heritage, the main principle of preserving the equality of languages is fixed in its fundamental treaties.

When the first computers appeared, the idea concerning their usage for solving polymorphic issues arose. This issue was relevant at different stages of the development of computers and technological progress: from the time of the first computer technologies that were unsuitable for the full implementation of automatic translation to the most modern ones, even digital ones that are still not perfect and perfect machine translation tools, which indicates the vital role of the translator-post-editor in the quality of inter- and transcultural relations.

The field of Artificial Intelligence (AI) continues to attract attention of many scholars working in Corpus Linguistics. Chomsky himself categorically did not recognize this approach. In support of his skeptical views, scholars cite his extremely radical objections (for example, Professor Alla Korolyova at the international conference on corpus linguistics at the National Pedagogical Drahomanov University),

quoting his answer on this matter, given in a 2004 interview when asked about his attitude toward corpus linguistics. The answer was as follows: “Corpus linguistics does not mean anything" (cit. in Andor [1]). Moreover, his supporter, Professor Lees, as far back as 1962, at a conference at Brown University, declared that "creating a corpus is a waste of time and government money. Moreover, a native speaker can provide more examples of any phenomenon of English grammar in ten minutes than can be found in many millions of words of random texts".

Chomsky, a founder of American generative linguistics, is convinced that **the corpus approach** is reduced to a trivial observation of a large amount of data and "is not a method of scientific knowledge, and therefore cannot provide either a successful solution to cognitive and practical problems or the accumulation new knowledge."

However, already in the middle of the 20th century in linguistics, early generativist illusions were replaced by text-centric and discourse-centric trends, which are based on the understanding that the study of any fragment of the language system should be carried out using a representative number of texts of the corresponding language, which, in principle, there is a corpus. However, the criteria for such representativeness are still under development.

Furthermore, therefore, judging by the long discussion around the main question, what is the corpus approach as a direction of applied research: a requirement of the time or a new temporary trend in linguistic studies? We must state that although it is still far from its completion, it is already possible to provide a preliminary convincing and promising answer to this question, which fits into the context of the views of Plungyan, who assures that without corpus linguistics, modern linguistic science will undergo a significant regression.

Moreover, already preliminary observations and critical analysis of the scientific literature on this problem give reason to assume that three approaches to the assessment and significance of the corpus direction for the study of linguistic phenomena have been formed in modern science: 1) radical-categorical, 2) moderate-skeptical, and 3) scientific-promising. In world science, the debate is not about the attitude of scholars to the corpus but about the approaches to working with this linguistic resource and its reliability as a search engine.

Moreover, both frequency phenomena and occasionally used units can be studied and interpreted on corpus materials. Comparing and analyzing the data obtained through different corpora, it is possible to identify linguistic variability and patterns of language changes, predict the further development of the phenomenon under study, etc.

According to Plungyan (2008), **the corpus-based approach** makes the results more empirically relevant. The scientist assumes that the fundamental novelty of the results of corpus research gives grounds for the development of absolutely innovative "corpus dictionaries" and "corpus grammars," conclude and verified about a specific fixed corpus. In turn, the corpus nature of dictionaries and grammar increases their reliability and verification, thus preventing subjectivity and incompleteness. The creation of analyzers and specialized dictionaries for automated corpus mark-up

(morphological, syntactic, thematic) is technologically possible only within the framework of corpus linguistics (cit. in Boriskina [2, p. 27]).

OPUS is a free corpus system in open access (URL: https://opus.nlpl.eu/), which contains corpora of texts from L1 and L2 to L3...Ln from various Internet resources and which is constantly replenished. All texts are converted and aligned according to the methodology of corpus linguistics. The name of the corpus resource OPUS (English ... *the open parallel corpus*) was formed from the English word CORPUS by omitting the letters C and R.

The key characteristics of the OPUS corpus resource include the following: if the first characteristic is **multilingualism** because OPUS contains more than 90 European / non-European languages, then the second is **parallelism** because OPUS contains a large number of parallel text corpora. The multilingual nature of the corpus makes it necessary to process its documents in language-specific ways, so work is currently underway to create special processing programs for all languages included in OPUS.

In the article "Parallel Data, Tools and Interfaces in OPUS," Tiedemann notes that OPUS contains more than 3,800 language pairs, which is more than 40 billion tokens in 2.7 billion parallel units [6, p. 2216]. In addition, it is worth emphasizing that OPUS also provides tools for parallel processing and monolingual L1 data, as well as several options for searching data, making it a unique resource for research activities of any direction.

Tiedemann proposed a model illustrating the scope of the 100 most numerous language pairs included in the selection of OPUS text corpora. The model shows that these subcorpora significantly exceed the 100-million-word mark, which is high even for data-intensive natural language processing (NLP) [6, p. 2216].

Today, according to Tiedemann, the Spanish-English pair of 36 million parallel sentences containing approximately 500 million tokens remains the language pair with the largest volume of parallel data. Despite the fact that a large number of these popular language pairs are mainly traditional languages, among the top 100, there are also various language pairs that, on the contrary, have a lower resource potential. These are parallel texts with such pairs as Bulgarian-Hungarian and Romanian-Turkish. Moreover, they contain more than 100 million words, which are the most rarely used [6, p. 2216].

As for the subject matter of the texts, Tiedemann notes that "the largest domains covered by OPUS are legislative and administrative texts (mainly from the European Union and related organizations), translation of subtitles for films and data on the localization of software projects open source. There are also many nonfiction texts and other smaller texts from various online sources" [5].

The advantage of the updated versions of OPUS is the ability to download texts in different formats for all subcorpora. Thus, all data is provided in native XML format (using the XCES Align DTD for sentence alignment), Translation Memory eXchange (TMX) format, and plain text format (for Moses/GIZA++). In addition, a unique interface was developed for searching for specific language resources (see Fig. 3). A Wiki resource (http://opus.lingfil.uu.se/trac) was also developed with additional

information about the corpus. Furthermore, the website and OPUS-related data are stored on a separate dedicated server to reduce interference with other processes (http://opus.lingfil.uu.se) and users [6, p. 2216).

In conclusion, it should be mentioned that the verification of one-, two- and three-component L1 and L2 lexical constructions in OPUS proved its effectiveness in checking equivalents / differentiated equivalents in L2 texts, which helps the translator ensure the correctness of translations.

REFERENCES

1. Andor, J. (2004). The master and his performance: An interview with Noam Chomsky. *Intercultural Pragmatics*, *1*(1), 93-111.http://doi.org/10.1515/iprg.2004.009

2. Boriskina, O. O. (2015). Korpusnoye issledovaniye yazyka: moda ili neobkhodimost'? [Corpus study of language: Fashion or necessity?]. *Vestnik VGU. Seriya: Lingvistika i mezhkul'turnaya kommunikatsiya*, *3*, 24-27. http://www.vestnik.vsu.ru/pdf/lingvo/2015/03/2015-03-03.pdf

3. Plungyan, V. A. (2008). Korpus kak instrument I kak ideologiya: O nekotorykh urokakh sovremennoy korpusnoy lingvistiki [Corpus as a tool and as an ideology: On some lessons of modern corpus linguistics]. *Russkiy yazyk v nauchnom osveshchenii*, *2*(16), 7-20. http://www.philology.ru/linguistics2/plungyan08.htm

4. Plungyan, V. A., & Rakhilina, Y. V. (2009). Novyye vozmozhnosti natsional'nogo korpusa russkogo yazyka kak unikalnogo internet-resursa [New possibilities of the national corpus of the Russian language as a unique Internet resource]. *Russkiy yazyk I literatura v obrazovatelnoy sisteme Armenii: problem I perspektivy*, 21-35. https://mail.brusov.am/docs/SMO-final.pdf#page=21.

5. Tiedemann, J. (2009). News from opus – a collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, & R. Mitkov, editors, *Recent Advances in Natural Language Processing*, vol. V, pp. 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria.

6. Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *LREC Conferences*, pp. 2214-2218.