CORPUS-DRIVEN STUDIES AND CORPUS-BASED TRANSLATION: PRAGMATIC POTENTIAL

Svitlana Matvieieva
National Pedagogical Dragomanov University, Ukraine
sam141175@gmail.com

ABSTRACT:

The article deals with the language corpus and its potential for corpus-driven studies and corpus-based translation. The author analyzes the phenomenon of the language corpus as a theoretical and practical tool for Linguistics, Lexicography etc. and the importance of parallel corpora for translation and Translation Studies. The research is made on the material of the pilot English-Ukrainian parallel corpus.

Key words: English, parallel corpus, text, translation, Ukrainian.

INTRODUCTION

Corpus Linguistics is one of the most promising and priority areas of modern Linguistics. Today, corpus data are widely used in many areas of Humanities, such as Lexicography, Translation Studies, Sociolinguistics, Journalism, Pedagogy and the like. At the same time, certain aspects of the creation and practical use of corpora for solving separate issues associated with the translation of texts of various genres require further study.

LITERATURE REVIEW

Recently, problems of Corpus Linguistics have received much attention in world Linguistics. A large contribution to the study of Corpus Linguistics issues was made by such Ukrainian researchers as N. Darchuk [1], O. Demska [2], V. Zhukovska [3], V. Shyrokov [4], and other. World Linguistics pays special attention to parallel corpora (Y. Adesam [5], T. McEnery & R. Xiao [6], W. Teubert [7] and other) and their use in practical translation and Translation Studies (M. Baker [8], L. Bowker [9] and other).

MATERIALS AND METHODS.

According to V. Zhukovska, Corpus Linguistics is "a branch of Applied Linguistics engaged in defining general principles for constructing, processing and operating data from linguistic corpora (or text corpora) using modern computer technologies, developing methods for collecting real language phenomena — written and oral texts, as well as the ways to save and analyze them" [3]. Following A. Demska, we consider the text corpus to be "in a certain way an organized electronic collection of written and oral texts of an arbitrary natural language, which has inherent obligatory features and is intended to solve specific linguistic tasks" [2]. So, in a general sense, the linguistic corpus is a digitally presented, large, unified, structured, annotated and philologically competent collection of texts in (a) natural language(s), supported by a control system – universal software for searching and processing a variety of linguistic information.

Often linguistic corpora are used in morphological, lexicological and syntactic studies. Corpus data bases allow searching and statistical calculation of the use of given roots, affixes and flections, thereby the methods of creating a language unit can be studied. Linguistic corpora allow to get data on specific forms of the words and on the whole grammatical categories. For example, linguists can conduct system-structural studies of the morphology of the verb: the volume of implementation of the verb paradigm and the peculiarities of using verb forms in modern discourses; categories of aspects and the aspectual oppositions in modern texts; categories of time and phenomena of transposition and so on.

A significant number of dictionaries today are based on data from linguistic corpora (for example, Cambridge, Collins, Longman, Macmillan, Merriam-Webster, Oxford and other). Such dictionaries are built on the material of various (mono- and multilingual) corpora. Moreover, for accuracy, precision, and representativeness of the latest trends in the language and the provision of relevant information on the applicability and compatibility of words, these dictionaries are constantly updated. Also, corpora are often addressed when making reference literature – grammar books, educational dictionaries and reference books, which include not only vocabulary, but also grammatical information. "At the moment, there are already several corpus-based English grammars, for example, general (Collins COBUILD English Grammar, 1990; Longman Grammar of Spoken and Written English, 1999) and those specialized in certain parts of speech, in particular, the verb (Collins COBUILD Grammar Patterns 1: Verbs, 1996; Dieter Mindt an Empirical Grammar of the English Verb)" [3].

Linguistic corpora serve as material for studying the frequency of use of linguistic units as a whole, and their direct functioning in the texts of different styles. For example, based on the English corpus *Longman Spoken and Written English (LSWE) Corpus*, the American linguist Douglas Biber [10] identified 12 of the most common verbs in the English language: *say, get, go, know, think, see, make, come, take, want, give, and mean*. They are especially common in colloquial style, where they make up almost 45% of all lexical verbs. Besides there are corpus-based technologies for study specifics of structure and using various types of grammatical constructions, for example, in past and perfect forms, which are used in different periods of language development. The study of grammatical category of aspectuality was conducted using the corpus approach. It is proved that the simple form of the verb in English is used 20 times more than a progressive, or a continuous one, although there are a number of verbs that are mainly found in their continuous form, namely: *bleeding, chasing, shopping, starving, joking, kidding, and moaning* [10].

The parallel corpus requires special attention of translators. The *parallel corpus* is understood as "a translation repository" [7], "source texts and their translations" [11], "a corpus that contains source texts and their translations" [6]. The Ukrainian researchers define this phenomenon as "the unity of a subset of original texts and a subset of their translations into another language(s)" [2], "a corpus consisting of at least two subcorpora, one of which is the source one, and the other contains texts-translations of the source corpus" [3].

Recently the pilot English-Ukrainian parallel legal corpus has been build. This corpus is made of the European Court decisions in English ("European Court of Human Rights" [12]) and their translations into Ukrainian ("ECHR: Cases, Opinions, Matter. Ukrainian Aspect" [13]). These texts are organized by the author of the article into the pilot English-Ukrainian aligned parallel legal corpus for research (the volume of the Corpus is 595 305 words (322 135 – in English, 273 170 – in Ukrainian; these numbers are used as reference values for all the calculations).

Normally, the corpus analysis procedure includes three main steps: identification of language data using categorical analysis, correlation of speech data using statistical methods and intelligent interpretation of the results [3]. 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. They provide a unique tool for studying the language, thanks to which we can search in large text arrays, obtain data on language units and phenomena of various language levels (phonetic, morphological, lexical-semantic and syntactic): to study the frequency of word forms, lexemes, grammatical categories, syntactic constructions; to determine atypical grammatical phenomena and constructions; to find the

closest lexical and grammatical surroundings of a word, with which we can analyze the use of a word in all its collocations ("word combinations which have developed an idiomatic semantic relation based on their frequent co-occurrence" [14]), colligations ("morphologically and syntactically motivated conditions for the ability of linguistic elements to be combined" [14]), and syntaxemes ("a basic semantic-syntactical element of a language" [15]).

RESULTS AND DISCUSSION

We have conducted the research of translation of the terminological unites used in the legal discourse. The study shows that corpus analysis provides objective results in various fields, including the specifics of pragmatic contexts of the use of various language units and their translation with respect to the particulars of the context of the target language.

Analysis of the data of linguistic corpora helps analyze the contextual use of units, especially synonymic ones, their frequency compatibility with other words, and clearly define their semantics, which is crucial for the success of the translation of any text. One of the main tasks of engaging parallel corpora is to establish the frequency and specifics of using certain language units in the source text (in English) and their correspondences in the target text (in Ukrainian). Also, the corpus allows the linguist to conduct research and verify various lexical and grammatical factors affecting the valency of the words.

The usefulness and realism of the corpus directly depends on the degree of its balance and representativeness. For this, it, firstly, should be of a sufficiently large volume, and secondly, have a clear structure and annotation for more convenient analysis, that is, characterize the text as a whole; separate one word from another, highlight the boundaries of a phrase, sentence, text, and assign certain linguistic information to text units. The richer and more diverse the annotation, the higher the scientific and educational value of the corpus.

CONCLUSION

So, the priority goals of using corpora in linguistic research is the ability to quickly process large text volumes and obtain objective quantitative results. Moreover, the combination of corpus-driven studies with traditional vocabulary analysis makes it possible to obtain systematized and accurate data.

Corpus studies allow to analyze various language units at any language level in their real environment for a more accurate and detailed description of their lexical and grammatical features. Using the corpus approach to the study of language data allows the analysis of the language and its units in real use. The study of language units through the corpus makes it possible to obtain accurate data on the lexical composition of the language and the compatibility of its units, which, of course, has a positive impact on the process and result of translation.

REFERENCES

1. Дарчук, Н.П. (2013). *Комп'ютерне анотування українського тексту: результати і перспективи*. Київ: Освіта України.
2. Демська, О.М. (2011). Текстовий корпус: ідея іншої форми. Київ: НаУКМА.
3. Жуковська, В.В. (2013). Вступ до корпусної лінгвістики. Житомир: Вид-во ЖДУ ім. М. Франка.
4. Широков, В.А., Бугаков, О.В., Грязнухіна, Т.О., Костишин, О.М., Кригін, М.Ю. (2005). Корпусна лінгвістика. Київ: Довіра.

5. Adesam, Y. (2012). The Multilingual Forest – Investigating High-quality Parallel Corpus Development. Stockholm.
6. McEnery, A.M., Xiao, R.Z. (2007). Parallel and Comparable Corpora: What are they up to? Incorporating Corpora: Translation and the Linguist. Clevedon, UK: Multilingual Matters.
7. Teubert, W. (2007). Text Corpora and Multilingual Lexicography. Benjamins Current Topics.
8. Baker, M. (1995). Corpora in Translation Studies. An Overview and Suggestions for Future Research. Target, 7 (2), pp. 223-243.
9. Bowker, L. (1998). Using Specialized Monolingual Native-Language Corpora as a Translation Resource: A Pilot Study. Meta: Translators' Journal, Vol. 43, No.4, pp. 631-651.
10. Biber, D. (2001). Using corpus-based methods to investigate grammar and use: some case studies on the use of verbs in English. Corpus Linguistics in North America, pp. 101–115.
11. Anderman, G., Rogers, M. (2018). Incorporating Corpora. The Linguist and the Translator. Clevedon: Buffalo. Toronto.
12. European Court of Human Rights. [Online]. Available: https://www.echr.coe.int.
13. ECHR: Cases, Opinions, Matters. Ukrainian Aspect. [Online]. Available: https://www.echr.com.ua.
14. Bussmann, H. (1998). Routledge Dictionary of Language and Linguistics. London, New York: Routledge.
15. YourDictionary. [Online]. Available: https://www.yourdictionary.com.