PROBLEMS IN CREATING PARALLEL CORPORA

Valisher Azamovich Tangriyev
Teacher of Denau Institute of Entrepreneurship and Pedagogy
v.tangriev@dtpi.uz

Mushtariybonu Shoymardon kizi Yangiboyeva
Student of Denau Institute of Entrepreneurship and Pedagogy

ABSTRACT

This article is devoted to a new type of parallel corpus, linguistic sources, as well as the problems of creating parallel corpora. Corpus linguistics is a branch of computational linguistics that develops general principles for the construction and operation of linguistic corpora. Corpora are divided into monolingual, bilingual, and multilingual categories. Multilingual corpora combine texts written independently in two or more languages in the same thematic area. If different languages coexist in the same area, the emergence of many parallel texts is inevitable. Only texts called "science languages" can provide a sufficient amount of textual material to obtain parallel corpora. For many language pairs, parallel scientific texts can only be obtained through a third language. None of the above text types can be a stable and universal source for a parallel corpus.

Keywords: Corpus Linguistics, Parallel Corpora, Source Language, Translation Language, Parallel Corpus of Scientific Texts, Parallel Corpus of Media Texts, Parallel Corpus of Fiction Texts**.**

INTRODUCTION

Cultural, socio-political, and economic changes in the world today are reflected in language. The development of computer technology and the spread of the Internet in the late 19th and early 20th centuries paved the way for the emergence of modern linguistics in the fields of computer linguistics and corpus linguistics. In today's informed society, any language must become the language of artificial intelligence in order to survive in society. Creating an artificial form of any language is an important task of corpus linguistics.

LITERATURE ANALYSIS AND METHODOLOGY

In the first half of the 1990s, corpus linguistics was formed as a separate part of the science of language. At the same

time, it works closely with computer linguistics, taking advantage of its achievements and enriching them. Since the late 1950s, significant work has been done in corpus linguistics. These include Randolph Quirk's Department of English Language Use Studies, founded in 1959, and Francis and Kuchera's Brown Corpus, published in 1964.

Corpus linguistics is a branch of computational linguistics that develops general principles for the construction and operation of linguistic corpora (text corpora) using computer technology. A linguistic corpus of texts is a set of machine-readable, combined, structured, defined, philologically perfect linguistic data designed to solve specific language problems.

Corpus types include specialized, informative, multilingual, parallel, study, comparative, diachronic, and monitor. According to the criterion of parallelism, corpora are divided into monolingual, bilingual, and multilingual categories. Bilingual and multilingual corpora combine texts written independently in two or more languages in the same thematic area (e.g., a collection of conference proceedings on a specific scientific problem conducted in different countries and in different languages). Such a corpus aids in terminology and is often used by translators. Another option for a bilingual or multilingual corpora is to include original texts written in any source language and translations of these source texts into one or more other languages. Such corpora serve as invaluable resources for comparative research, research on translation theory, and research on human and computer translation.

The parallel text corpus is a relatively new type of linguistic source. The first Parallel Corpus texts are avalanche reports collected in German, French, and Italian in Switzerland, and weather information in English and French in the Canadian media. The first sources of this type appeared in the late 1980s - early 1990s. Over the last decade, a number of projects related to parallel corpus have been launched. For instance, the Anglo-French parallel debate corpus in the Canadian Parliament (Canada-Hansards Anglo-French parallel corpus).

The INTERSECT project at the University of Brighton (International Sample of English Contrasting Texts), Anglo-French Parallel Corpus, including EU Telecommunications Official Documents CRATER (International Telecommunication Union) Trilingual French-Spanish-English Parallel Corpus, 1 million words. This corpus contains texts in the field of telecommunications. The Anglo-Norwegian parallel corpus was created in 1994-1997 at the University of Oslo (Norway) in a

project led by Stig Johansson. The corpus consists of original literary texts in English and Norwegian and their translations into Norwegian and English. The creation of a corpus is currently being expanded, with the new corpus being renamed the Oslo Multilingual Corpus. The original Anglo-Norwegian corpus is filled with German and French texts.

RESULTS AND DISCUSSION

In creating parallel corpora, it is necessary to take into account the factor of intercultural relations, as opposed to a single language and comparative texts. Source language texts are only texts translated into a second language. Thus, if there is no intercultural communication at all, it is not possible to create a parallel corpus. The weaker the connections, the less the cultures are connected, the fewer translations are performed, and the more difficult it is to create a complete parallel corpus. For example, the existence of political and cultural ties between two countries requires the translation of various documents, guidelines, manuals, brochures, etc., from one language to another, i.e., the influx of tourists, small businesses, business relationships, marriages, and so on. An important factor in strengthening ties between countries is that they help to increase interest in another culture rather than the development of "formal" relations. The most interesting thing is that a factor such as geographical proximity does not play as important a role as expected. Although English-speaking countries are not Russia's neighbors (unless we consider the border with the United States along the Bering Strait), the number of texts translated from English into Russian (as well as the number of translations from Russian into English) texts) significantly exceeds the number of translations from other languages. Poland, the Czech Republic, and Slovakia are closer to Russia than Germany and France, and these countries are former partners of Russia in the Warsaw Pact and the Council for Mutual Economic Assistance, but it is clear that Russian is translated from Polish or Czech rather than German or French. If different languages coexist in the same area, the emergence of many parallel texts is inevitable: official texts, documents, instructions, advertising texts, textbooks, translations of fiction, and so on.

The parallel corpus can be compared to the point of intersection of two linguistic cultures. The parallel corpus consists of two sub-corpora: texts in the source language and their translation into one or more other languages. Texts in the source language, although they are primary, are selected based on the source language. In general, the structure of

the source language sub-corpus is determined by the presence or absence of translations into the source language, as well as what texts are being translated.

In general, when creating a parallel corpus, the researcher may have the following language resources at his disposal:

- special texts;
- mass-media texts;
- scientific texts;
- artistic texts.

Documents.These are personal documents (birth certificates, marriage certificates, education documents); business letters, contracts, commercial offers, business plans, licenses; texts of international agreements, materials of diplomatic negotiations, etc. If there are two official languages in the country, for example, Finnish and Swedish in Finland and English-French in Canada, there will be many similar texts. The existence of such parallel texts also depends on the existence of business, diplomatic, and political ties between the two countries. Integration processes in EU countries also lead to the emergence of many documents written in multiple languages. The main problem with creating a corpus from this type of text is the confidentiality of many documents. This problem is solved by removing names, organization names, geographical names, dates, and so on from the texts. Since most of the documents are "ephemeris," meaning that the translation is done once for a single client and the text is deleted after the work is delivered, it is also difficult to obtain such types of text arrays. Another difficulty is that the source code is often in "paper" form. The next problem that can be encountered is the poor quality of many personal documents and business correspondence (the presence of factual errors and vague equivalents in the translator's translation into the native language and grammatical and methodological errors in the translation into another language).

It should be noted that if the target language is used in compiling translation dictionaries or as a source of information for translators, the quality of the translation plays an important role. If corpus compilers plan to check for common interlingual translation errors, the work done in this regard is of great value. The work done to create parallel bodies of documents is not enough. Unfortunately, the texts of the NATO Texts were collected at the University of Mannheim, and the texts of this corpus did not correspond to each other in parallel. Texts of instructions and guides are very common and vary in form and content, especially from the text on food packaging to the tourist brochure. For some language pairs and some sectors of the economy, this type of text

is available in large quantities. For example, Finnish travel brochures are always translated into Swedish and English, mostly German and Russian. The manual for home electronics is always translated into several languages, one of which is usually English. These types of texts are very useful not only for research purposes, but also for developing various practical applications. Technical documentation is an important component of many parallel texts. However, for some texts, finding a parallel to this genre is not as easy as it seems. Electronics are almost never exported from Russia, so it's hard to believe the sheer number of manuals for mobile phones translated from Russian. Finland exports mobile phones, but manuals for their operation are translated into English in Finland itself. In addition, in some cases, the instructions can be structured in English and then translated into Finnish and Swedish. When Finland exported mobile phones, the documents were probably translated from English, not Finnish. Thus, in the above case, we can refer to a pseudo-parallel that is derived from the translated texts. The situation is the same in the translations of products by Japanese and Korean firms and their documents.

The Swedish multilingual text corpus and the Croatian-English corpus of full media texts are included in the parallel language corpora of the media. But there are still a few projects to create parallel corpora of media language.

CONCLUSION

Scientific texts often become the objects of translation, but a number of explanations need to be given here. Scholars who use many scientific texts speak foreign languages. Often, the scholar himself writes in a language familiar to most of his listeners (Latin—in the Middle Ages, French or German—in the 19th century, English—in the present). Therefore, only classical scientific works (Darwin, Marx, Carlisle, Saussure, etc.) have been translated into many languages. Thus, only texts called "science languages" can provide a sufficient amount of textual material to obtain parallel corpora, and for many language pairs, parallel scientific texts can only be obtained through a third language.

None of the above text types can be a stable and universal source for a parallel corpus. As mentioned above, there are a couple of languages in which some types of parallel texts may not exist at all. However, translations of fiction are done even if there are no brochures, technical documents, or translations of scientific texts in these languages. Of course, in many cases, such resources may be limited. Nevertheless, the translation of fiction plays a more

important role in the development of intercultural relationships than the translation of business letters or travel projects.

Poetic translations are very interesting linguistic sources, but the scope of practical application of parallel poetic corpora, in particular, is much more limited than that of prose, which can be used as a lexicographic source. Prose texts include monologues and dialogues, stories and descriptions, normative language and jargon, dialects. Literary texts are undoubtedly a very important source for the parallel corpus.

The following conclusions can be drawn from the analyzed issues:

1. For any language to survive in society, it must be transformed into artificial intelligence, and this is done through corpus linguistics.

2. The factor of intercultural relations should be taken into account when creating parallel corpora.

4. When creating parallel corpora, the translated texts should not deviate from the source texts.

5. When creating parallel corpora, it is necessary to take into account the stylistic features when converting source texts to translated texts.

REFERENCES

1. Захаров В. П., БогдановаС. Ю.Корпусная лингвистика: учебник. 3-е изд., перераб. — СПб.: Изд-во С.-Петерб. ун-та, 2020. — 234 с.
2. Михайлов М. Параллельные корпуса художественных текстов: диссертация.-Финландия . ун-та, 2003. —348с.
3. Azmovich, T. V. (2019). Analysing Some Uzbek Texts Via Corpus Analysis Toolkit-"Antconc". Think India Journal, 22(4), 4690-4700.
4. Mardievna, B. M., Mukhamadjanovna, J. S., Nematovich, N. O., & Azamovich, T. V. (2020). The importance of modern methods and technologies in learning English. Journal of critical reviews, 7(6), 143-148.
5. McEnery, Tony and Wilson, Andrew 2001: Corpus Linguistics: An Introduction. 2nd edition. Edinbourgh: Edinbourgh University
6. Salkie, Raphael 2002: How can linguists profit from parallel corpora? In: Borin, Lars (ed.). Parallel Corpora, Parallel Worlds. Selected papers from a symposium on parallel and comparable corpora at Uppsala University, Sweden, 22–23 April, 1999. Amsterdam — New York, NY: Rodopi. Pp. 93–10 Teubert,
Wolfgang 1996: Comparable or Parallel Corpora. International

Journal of Lexicography. Oxford University Press. 9(3), 238–264.9.
7. Trosterud, Trond 2002: Parallel corpora as tools for investigating and developing minority languages. In: Borin, Lars (ed.). Parallel Corpora, Parallel Worlds. Selected papers from a symposium on parallel and comparable corpora at Uppsala University, Sweden, 22–23 April, 1999. Amsterdam — New York, NY: Rodopi. Pp. 111–122.
8. Ruziyev, K. B. (2020). CORPORATIONS THAT ARE COMPARABLE AND PARALLEL Актуальные научные исследования в современном мире, (11-12), 43-46.
10. Khodjaeva, N. (2021). Teaching grammar and understanding meaning in context.
11. Khodjaeva, N. T. (2019). Some Peculiarities And The Ways Of Giving Instructions On Reading Tests. International Journal of Research, 499-505.
12. Ходжаева, Н. Т., & Бахриддинова, М. Ш. (2020). Стилистические характеристики специальных текстов при информативном переводе. Актуальные проблемы гуманитарных и естественных наук, (6), 95-98.