Features and Differences of the Parallel Corpus of English and Uzbek Languages

Norov Jamshid Nurulloyevich
A teacher of Academic Lyceum of Bukhara Engineering Technological Institute, Bukhara city

Abstract:

A parallel corpus consists of texts that have been translated one / more than the original. Which topic to PC, the choice of text in the genre depends on the purpose of the compiler. When choosing a text for the Uzbek-English PC, it is advisable to collect translations from Uzbek into English and direct translations from English into Uzbek. Because some units may lose their value in indirect translation, the PC cannot fully perform its function, so texts that translate directly from the original to the PC are included.

Key words: A parallel corpus, national corpus, Uzbek linguistics, corpus linguistics, dialectology, quantitative and qualitative analysis.

I. Introduction

The national corpus means the national language treasure. It is widely used by linguists, lexicographers, computer linguists, programmers, editors, translators, journalists, publishers, scientists, teachers, students and any other professional. Language corpus is an undeniable tool for language research and practical problem solving. It is different from a normal electronic library. The purpose of the e-library is to provide a full range of works of art and journalism that reflect the socio-political, spiritual and economic life of the people. The fact that the library texts are not processed from the linguistic point of view is inconvenient for research. Because the e-library is not created to prepare a database of research materials, but to consolidate the national spiritual heritage. The language corpus, unlike the electronic library, is a collection of useful and interesting texts necessary for the study and research of a language.

II. Literature review

Much has been done in world linguistics in the field of corpus linguistics and the creation of language corpora. In many countries, such corpus began to be formed in the 80s of the twentieth century. They serve a variety of purposes and tasks. The Bank of English and the British National Corpus (BNC) have developed projects in the UK, and the Russian National Corpus in Russia. For example, the volume of the National Corpus of the Russian language is now 149 million words. In recent years, the development of the Internet has led to the emergence of a corpus of virtual texts. That is, Internet searches sites, electronic libraries; virtual encyclopedias serve as a corpus. The genre and thematic diversity of the case depends on the interests of the Internet user. The creation of a national corpus in the Uzbek language is still being delayed. Many dictionaries have been created in Uzbek, but no serious research has been done on the development of the language corpus and its theoretical basis.

III. Analysis

The science of corpus linguistics has created its own object and work material, which is the basis for its recognition as an independent linguistic science. The main goal of the science we are studying is the linguistic description of the language system. The main directions of corpus linguistics:

The main directions of modern corpus linguistics are:

- First, it is the creation of dictionaries and lexicographic research, all modern English dictionaries are based on the corpus (Collins, Webster, MacMillan, etc.);
- Secondly, to get accurate information about the lexical structure of languages by studying the corpus, to determine the frequency of use of words. The importance of corpus in lexicology is that no tool can match the corpus to determine the period and frequency of use of a word.

As a result of the search to determine the frequency of a particular word on the basis of the corpus, using diagrams and graphs, the ordinal number of the word is inversely proportional to its frequency, because the word in the second ordinal number is less than the first digit. It is clear that less is used than the third. No frequency dictionary can provide accurate body information because language is constantly changing and the frequency of words is relative. Based on such a practice of the corpus as the Sipfa law, the chances of identifying frequently used words in any language are now high. The first body of computer-generated texts was the Brown Corpus (UK, Brown Corpus, VS), created in 1961 at Brown University, each containing 500 text fragments of 2,000 words. In the 1970s, a frequency dictionary of the Russian language was created on the basis of a body of texts containing 1 million words. In the 1980s, a corpus of texts in Russian was also created at the University of Uppsala in Sweden.

Later, as a result of the development of computer lexicography, there was a need for a large body of texts. That is, 1 million words is not enough for an electronic dictionary database. On this basis, a large body of texts began to be created. In many countries, such corpus began to be formed in the 80s of the twentieth century. They serve a variety of purposes and tasks. The Bank of English and the British National Corpus (BNC) have developed projects in the UK, and the Russian National Corpus in Russia. For example, the volume of the National Corpus of the Russian language is now 149 million words. In recent years, the development of the Internet has led to the emergence of a corpus of virtual texts. That is, search sites on the Internet, electronic libraries, and virtual encyclopedias serve as a corpus. The genre and thematic diversity of the case depends on the interests of the Internet user. For example, in science, Wikipedia is used as a corpus of large volumes of text. The corpus is especially important in teaching and learning native and foreign languages. The fact that today the world language system is focused on the corpus is another proof of our opinion. Therefore, it is important that a number of micro corpus are being formed, such as educational buildings, dialect texts, poetic texts, oral, scientific, official texts, parallel buildings. The issue of teaching English, German, French and Russian as foreign languages is being studied separately in the methodology. There are also corpuses for language teaching, including the Learning Corpus of the Russian Language and the Learner Corpus of English. The importance of the language corpus in the process of working with representatives of foreign languages increases several times.

It should be noted that the first Russian language corpuses were created by Russian language researchers in Europe, not in Russia. There are monolingual and multilingual versions of the case material, depending on how many languages it is presented in. Corpus specialists (mostly translators) have always been interested in creating a multilingual corpus. From the earliest days of the corpus, English, Finnish, French, German, Greek, Norwegian, Spanish, Swedish, and so on. bilingual corporations for languages began to emerge. Such a corpus is also called bitexts. There is no barrier to making the corpus trilingual, quadruple or more, rather than bilingual. Experts also divide the body into monolithic, bilingual and multilingual types in terms of parallelism. In a monolingual corpus, when a language variant and dialects are contrasted, a bilingual and multilingual corpus consists of a set of texts written in different languages within the same subject. For example, it may include conference proceedings in different languages in different countries on a particular scientific problem.

IV. Discussion

Multilingual corporations are often used by translators. Another aspect of the multilingual corpus is the original text and the translated text. This type of corpus serves as a very important resource in comparative research, translation theory, and the study of computer translation.

There are 2 types of multilingual corpus:

1) text corpus with translation of each other;

2) a bilingual text corpus related to the same topic.

The first type of corpus is called a parallel corpus and is used to study various aspects of a particular translation. For example, there is a body of texts of the Canadian Parliament (English-French). The parallel body is further divided into two types - aligned and not aligned. The term “customized” means that there is a clear, interrelated relationship between translation units in the corpus. The advantage of such a corpus is that it is convenient to find out how this or that sentence is translated. This type of corpus is important for the translator because it has a unique resource - memory translation memory. The function of the "incompatible" corpus is to match the text with its translation, to show which unit in the translation corresponds to which unit. The adjustment can be done automatically or manually. The first method is easy, but there are many mistakes. For example, in the process of translation, a simple sentence can be given as a compound sentence. In this case, it is difficult to determine which building is original. An example of a multilingual harmonized corpus is the EU's Acquis Communautaire database. The second type of corpus is called the translation corpora and is important for studying the expression of the same idea in different languages. The value of a parallel corpus is determined by its size and the amount of languages. Acquis Communautaire is the largest parallel corpus in the world, which is characterized by the free use of the corpus and the presence of rare language pairs such as Maltese-Estonian and Slovenian-Finnish.

These housings can be used for the following purposes:

1) creation of typical translation methods and transformation;

2) study of statistics of automatic translation system;

3) creation of monolingual and multilingual dictionaries;

4) study and evaluation of data storage and transmission programs;

5) automatic verification of translation accuracy;

6) facilitate the work of the translator through the breadth of the possibility of equivalent selection.

The importance of the parallel corpus, which includes the world's languages, shows the urgency of the issue of creating a common corpus of Turkic languages. The common corpus of Turkic languages is important in that it serves as a tool for studying the rich sources of textual, comparative linguistics, translation theory, literature, related interlinguistic relations, and language vocabulary. The creation of such a corpus will not only ensure the development of languages belonging to the Turkic language family, but also guarantee the preservation of a small number of Turkic languages whose users. It is natural that the joint (parallel) corpus of Turkic languages will serve as the most modern educational tool for teaching the common monuments of Turkic languages – west Avesto, Orkhon-Enasay monuments, and epics of Turkic peoples to the children of Turkic-speaking peoples. Many of the world's languages have their own national corpus, which differs in the level of excellence and the ability to scientifically process the text. The English-language Brown Corpus, the Lancaster-Oslo / Bergen (LOB) Corpus, the London-Lund Corpus, the American Heritage Corpus for Lexicographic Studies, the Lancaster English Speaking Corpus, and the Diachronic Corpus. The existence of famous corpus of English texts such as Helsinki Corpus, International Corpus of English Learners for Linguodidactical Research, Bank of England, British National Corpus, International English Corpus, American National Corpus as the latest generation of English corpus in the development of national and state language shows the importance and place of the national corpus. In addition to the national corpus of English, Spanish, Chinese, Arabic, French, Russian, German, Polish, Polish-Ukrainian, Czech, Slovak, Serbian, Croatian, Bosnian, Bulgarian, Bulgarian, Macedonian, Scottish, Dutch, Dutch-French, Swedish, Dutch, Norwegian, Icelandic, Faroese, Medieval French, Italian, Portuguese, Romanian, Lithuanian, Latvian, Greek, Eastern Armenian, Ossetian, Albanian, Hindi, Gypsy, Hittite, Finnish, Uralic, Estonian, There are corpuses of Veps, Hungarian, Udmurt, Georgian, Anglo-Georgian, Lezgin, Turkish, Tatar, Bashkir, Crimean Tatar, Kalmyk, Buryat, Mongol, Hebrew, Amharic, Japanese, ancient Japanese, Baman, Esperanto. In world computer linguistics, the existence of a national language corpus is seen as a criterion for the survival of a language and its transformation into a computer language.

The problem of giving morphemes that have been transformed in translation in a parallel corpus. In practical translation, giving a morpheme with a morpheme is a rare occurrence, which in most cases has theoretical significance. N.Kambarov agrees with L.S.Barkhudarov that all language units can be a unit of translation, noting that such methods are used in translation.

According to N. Kambarov, there are many words in English that describe a person who performs an action. If he is spreading *dinner,* he is a diner in English; if he spends a lot of money, *spender*; if he collects money, he is called a *saver*. In Uzbek it is formed with the help of the suffix *-chi, -vchi*. For example: *writer, student, weaver, teacher, florist, herdsman,* etc. The English *-er* morpheme is more productive than the Uzbek *-chi* morpheme. When the translation refers to a person engaged in an activity in English, and it is made with the suffix *-er*, it may not always be translated with the suffix -*chi* in Uzbek because it has no equivalent.

Theoretically, English morphemes such as *-er, -or, -cy, -man, -ship, -ment, -hood* have different lexical equivalents in Uzbek. However, it is not always possible to achieve equivalence in translation. The reason is that, for example, in annotated and bilingual dictionaries there are always words that are not part of the sentence, i.e. the equivalent of the words in the system. As they are used in speech, their shape, style, text types, and structures used in speech change. The above additions change in the process of translation under the influence of word order, genre, and style in the sentence: transformation, more precisely, occurs in the phenomenon of transposition.

Mr. Gray is a **World Banker** turned Oxford professor. Translation: *Mr. Gray previously **worked at the World Bank** and now teaches at Oxford University.*

The second form of translation is that *Mr. Gray was formerly **a World Bank employee** and now works at Oxford University.* This phrase cannot be translated as “World Banker” because to say so would be a violation of the grammatical norms of the Uzbek language. The word ***banker*** in English can be translated into Uzbek as a *banker* out of context, but the text and its content required the use of the word ***employee***.

In this regard, we fully agree with N. Kambarov. Indeed, such morphemes are common. For example: *The president understands that **lower** taxes may not improve wages for low earners on their own.* Translation: *Prezident soliqlarni kamaytirish bilan **kam haq to'lanadigan ishchila**rning daromadini oshirish yetarli emasligini tushunadi*. In this sentence, too, the name is given in the Uzbek translation using a combination of four words. The translation was done by pictorial way, word addition and periphrasis methods.

Such US companies as American Electric Power and Duke Energy are big coal-burners. Translation: *AQSHning “Amerika Elektrik Pauer” hamda “Dyuk Enerji” kabi kompaniyalari juda ko'p miqdorda* *ko'mir yoqadi.* In the translation of the above sentence from English into Uzbek, the method of figurative translation and word addition was also used. There is a free phrase "charcoal burner" in Uzbek, but the stylistic feature and functional style of the text do not require the use of this phrase.

Hence, the English morpheme can theoretically be given with examples in separate dictionaries, but it is advisable to translate it taking into account the nature and genre of the text in the translation process.

V. Conclusion

The above translation shows that Uzbek word combinations are widely used to convey the meaning of morphemes in English through figurative translation. From the point of view of translation theory, horse-forming morphemes in English are translated in the following ways:

1. Morpheme using morpheme.
2. Morpheme using words.
3. Morpheme using phrase.
4. Morpheme figuratively or verbally.

N. Qambarov evaluates the translation of morphemes as a manifestation of grammatical transformation and proposes to call it a morphemic translation transformation. The extent to which one of the methods of morphemic transformation is used in translation depends directly on the text, the units of language in it, and the characteristics of the languagesin contact with the translation.

References:

1. Aligned Hansards of the 36th Parliament of Canada Release 2001-1a // Information Sciences Institute: URL: http://www.isi.edu/natural-language/download/hansard/
2. British National Corpus (BNC) [Электронный ресурс] // British National Corpus: [сайт]. URL: http:// www .natcorp.ox.ac.uk/
3. Collins M., Koehn P., Kucerova I. Clause restructuring for statistical machine translation // Proceedings of the Association for Computational Linguistics (2005) [Электронный ресурс] // Faculty of Humanities - McMaster University: [сайт]. URL: www.humanities.mcmaster.ca/~kucerov/ ACL2005.pdf; Melamed I. Bitext Maps and Alignment via Pattern Recognition // Computational Linguistics. 1999. Vol. 25 (1).
4. Crime and Punishment by Fyodor Dostoyevsky [Электронный ресурс] // Project Gutenberg: URL: http://www.gutenberg.org/ebooks/2554
5. cyclowiki.org/wiki/Корпус_параллельных_болгарских_и_русских_тексто
6. Hoshimov U. Affairs of life. Stories (Translator: O.M.Muminov). – Tashkent, 2013. – 164 p.
7. Krave M. F. Converbs in Contrast: Russian converb constructions and their English and Norwegian counterparts. PhD thesis. – 2011, University of Oslo.
8. Loiseau S., Sitchinava D. V., Zalizniak A. A., Zatsman I.M. Information technologies for creating the database of equivalent verbal forms in the Russian-French multivariant parallel corpus // Информатика и её применения. − 2013. Т.7. Вып. 2. С. 100-109. DOI: https://doi.org/10.14357/19922264140210. (Параметры загрузки: IP: 213.230.114.157. 17 августа 2019 г., 20:50:00)
9. MaltParser // MaltParser: [сайт]. URL: http://www.maltparser.org/
10. The Collected Tales of Nikolai Gogol / translator Pevear R., Volokhonsky L. New York: Pantheon Books, 1998. – 435 p.
11. The Lady with the Dog and Other Stories by Anton Pavlovich Chekhov // Project Gutenberg: URL: http://www.gutenberg.org/ebooks/13415