JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 7, ISSUE 17, 2020

THEORETICAL FUNDAMENTALS OF UZBEK-ENGLISH PARALLEL CORPUS

Mengliev Bakhtiyor Rajabovich, Rustam Abdurasulovich Karimov

Professor of Tashkent State Uzbek language and Literature University named after Alisher Navoiy
A senior teacher of Foreign Languages Faculty Bukhara State University

Received: 21 March 2020 Revised and Accepted: 25 June 2020

Abstract – The article deals with the development of linguistics, in particular, the science of translation in general and in different directions, the use of modern linguistic sources and methods, including large sources collected in parallel corpus: scientific, socio-political, legal, economic texts, their translation into foreign languages. There is also information about the structure of the parallel corpus of Uzbek and English and its theoretical foundations.

Key words: development of linguistics, science of translation, modern linguistic sources and methods, parallel corpus, economic texts.

I.Introduction

The widespread use of the parallel corpus opens up new opportunities for access to bilingual linguistic information: the subsequent application of the data obtained; analyze and generalize it; assists in the implementation of human and computer translation, compiling a dictionary and learning foreign languages. The key to using these materials collected from a variety of sources is this text / unit layout.

Firstly, compare the text section / chapter, paragraph, sentence, phrase, and words selected for research purposes with the plain text; morphological, syntactic, semantic analysis, the task of comparing the state of semantic-syntactic connection directly or in the component model.

The size and level of the corpus annotation (unit mark) is determined by the researcher's goal. The large-volume parallel corpus, chosen in proportion to the subject, chronology, and genre, serves to study the general features of the language. The specialized parallel corpus mainly helps to solve specific-analytical questions in specific situations of translation, specific types of text, creation of authorship style, and so on.

II.Literature review

Below we present the views of S.B. Potemkin on the problem of creating a parallel corpus of Russian classics, their translation into English, and their solution. The scientist begins the work of creating a parallel corpus by **developing the principles of text selection.**

According to S.B. Potemkin, the modern Russian language was created by the great artists of the XIX century; the lexical, grammatical core is preserved, although in some cases deviations from the literary norm are observed. Teaching Russian to a foreign student, translating foreign literature into Russian, and vice versa, requires a corpus specialist to study Russian classical literature and its translation by mature translators. Interest in the Russian language arose in the late nineteenth and early twentieth centuries, and to this day interest in both the study of the Russian language and the translation of Russian classical literature into a foreign language is growing.

Russian classics that serve as material for the creation of the Russian-English parallel corpus: N.V. Gogol's "Petersburg News", F.M. Dostoevsky's "Crime and Punishment", the electronic form of a relatively old translation of A.P. Chekhov's stories made by Constance Garnett, are used. This means that the author has an electronic format in the selection of the text, relying on the finished translation works.

III.Analysis

Experiments show that, unlike multilingual corporations such as the British National Corpus, the Russian National Corpus, and the late 20th-century Russian newspaper corpus, the parallel corpus consists only of translated texts. They are relatively small in size and balanced. For example, the size of the Russian-British parallel corpus within the Russian-speaking national corpus is 10 percent of the size of the main corpus. The reason for this is very simple: it is easy to find the speech of language speakers in different areas: education, culture, social life, and so on. Translated texts are selected from texts pertaining to a particular area of social life. The text of the translation is collected on a limited scale - from foreign language teaching, various invitations, meeting speeches and conference materials. In fact, the existing parallel buildings are designed for a separate field of study (official working paper, works of art with translations, technical manuals, etc.).

Automatic processing of parallel corpus differs from other corpus: the main, technical problem of automatic corpus is the marking of the text in the original and translated as a unit of corpus at the level of speech, phrase and word. It is difficult to do this based on the automatic search model in NP. To do this, it will be necessary to compare

ISSN- 2394-5125 VOL 7, ISSUE 17, 2020

different statistical methods, frequency. Very rare word search methods, empirical and linguistic considerations are used in a particular text.

As a result of the efforts and research of modern linguists and programmers, convenient methods have been developed to equate the original and translated text units for a special parallel corpus of languages with a large linguistic base, mainly European languages such as Chinese and Japanese.

The copyright of such corpus, the closedness of their use, in many cases, prevents their free use for educational, research purposes. Such corpuses, consisting of the works of one author, form the bulk of the corpuses used in education and translation.

For example, the electronic corpus of agiographic texts of ancient Russian monuments is one of such means. Despite its small size, the productivity of such a hull is equal to that of a large hull with one tongue. The information obtained from them has some parameters that are not present in the general corpus - the frequency of word use, the presence of a lexicon with a specific feature, which has an advantage in determining the grammatical structure.

The main problem in building a parallel corpus is the complexity of the process. Often, it is easier to re-translate a text than to adapt the finished translation to the original text sentences. Over the last decade, several hundred sentences have had to be adapted to translation and originality for research purposes in order to automatically equate translation and originality units. Despite the successes achieved, an error occurs as a result of the automatic equalization of a rare word / compound. Therefore, the development of a method / algorithm that can provide manual equalization quality, leading to cost reduction, remains relevant.

The second important step in the formation of the corpus is **to align the text within the sentence,** i.e. in a parallel corpus the bilingual sentence is adapted to its content by its equivalent. Among the tasks to be performed is, first of all, the alignment of text units. Text units can be sentences, phrases, and words in a parallel corpus, as noted earlier. But to equate the bilingual texts of the parallel corpus is to determine the boundary of the sentences in the text; the original and the translation must begin by defining the boundaries of the sentences that express each other in the text. In many cases, for example, for a PC created for educational purposes, alignment at the sentence boundary alone will suffice. After all, the equation in the next stage (at the level of phrase and sentence) does not take place without this initial stage.

In the previous chapter of the work, the idea of the inconsistency (or vice versa) of the boundary between the original and the translated text was discussed. Sometimes in a translation, the sentences, even the whole paragraph, may be omitted or their boundaries may not be correct: in fact, the word in the first sentence falls into another sentence in the translated text. Usually, the inequality of the boundaries of words and phrases is observed in the translation of a work of art.

When the unit of equation of the corpus is a sentence, a purely structural (sentence length, word size) and statistical (based on the word being expressed) method is used. This method does not require an excellent dictionary base and can be used for languages that do not have a large dictionary resource.

IV.Discussion

Depending on the length of the sentence, the equation is very sensitive to the interval or punctuation, and the use of a single inappropriate space or punctuation leads to an error in the equation. The statistical method also gives an erroneous result in some cases: it requires subsequent manual verification and adjustment. For scientific texts, the transcription method is often used: because most terms are derived from one source - Greek, Latin, and later English, French, and German.

Such a comparison of the term is important for the next stage of equalization. The use of a bilingual dictionary in equations is rare; such a source is mainly applied to special texts (Anglo-French Protocols of the Canadian Parliament, EU legal texts).

According to S.B. Potemkin, there are several obstacles to the proposed method of equalization:

1) the boundaries of text sentences in Russian and English correspond;

2) there are no significant (more than 200 words) spaces in parallel texts;

3) The length of parallel texts is not large (40,000 word usage).

This method is performed using a two-sided Russian-English translation dictionary with a search engine of 1.5 million equivalent pairs in the text being analyzed. Firstly, low-frequency - words that occur once in a text are equated. The translation equivalent is determined for each such word in the original text. It is natural that such a low-frequency word is also rare in the translated text. If more than one equivalent of a word in the original text is identified, they are limited; such a unit is also bypassed if an exceptional case of conjugation occurs in the identified equivalent sentence. Such a filter results in a unique equivalent pair. This pair forms the main point of the structure - the anchor: they are interconnected in the text and form a sentence. In the next step, the translation is limited to the identified pairs / "pieces" of text.

These passages are treated as new parallel text, and the practice of finding the starting point is repeated; the return will continue until a new anchor is found. In practice, the process of re-analysis did not exceed 6 times. The translation of each word in the original text piece (which can be abbreviated as a paragraph) is searched for equivalent in the dictionary of the language being translated for the text.

ISSN- 2394-5125 VOL 7, ISSUE 17, 2020

Finding such an equivalent is important for the process of equalizing each sentence. The compatibility level is set next to the original text cell in the table. The paragraph is completed from beginning to end by filling in a side-by-side matrix cell. With such a critical approach, the search is done with standard methods of dynamic programming. For example, lemmatization for the Russian part is performed on the basis of a dictionary of words using the program StarLing.

For obsolete, obsolete words, manual word-changing paradigms were created, used in lemmatization, and a dictionary of word forms was supplemented. Due to the presence of many word forms in bilingual dictionaries, English text lemmatization was not performed. As a result of equalization of the English and Russian versions of the story "Anna on the shoulder" by A.P. Chekhov, 223 sentences were found in Russian and 139 sentences in English. Out of 182 pairs, 165 (90.5%) were fully and correctly translated, 16 sentences (9%) were part (or reverse) of the translation, and 1 sentence (0.5%) was compared with an erroneous translation.

Here is a typical example of a single sentence in the original text being translated as two sentences: "Nima gap kasalmisiz?" <> “What’s this?” “Are you ill?”. Such errors are easily corrected in the process of lexical analysis. But there is a relatively complicated process: a few sentences are represented by 2-3 sentences, but their boundaries do not match: “Eh, u uxlayapti-ku! – xitob qildi u, u baribir uxlayapti!" <> “My goodness; how he sleeps! - she cried indignantly: And he is always asleep ”. In this case, it is necessary to equate the sentences without setting boundaries.

After this type of repetitive iteration on the text, the undetected 5% of the sentence was manually equalized. When equating parallel sentences, a smaller unit than the sentence can be made on the basis of statistical models of machine translation: it is possible to translate any element of the original sentence - the word into a second language. In a parallel corpus, the translation of the original word into a second language depends on its frequency: in a parallel corpus, the frequency of both must be the same. It is accepted for a relatively common equivalent translation. This approach also has a number of shortcomings: this shortcoming is related to low-frequency wording. Typically, in a small-volume corpus, the compound is represented by a word in different places in the word order.

About half of the corpus vocabulary consists of a low-frequency, low-frequency construction (e.g., 10 in a million word usage). It is well known that rare, emergencies do not provide sufficient information for statistical analysis. On the other hand, 5-10% of vocabulary consists of high-frequency words; such words occur in any position of the corpus; if the equations are based on statistics, they can be compared with different units. There is another problem with word-level equalization. It is a matter of the order of the word in the original and in the translation.

In most cases, there is a correspondence in the order of sentences in English and Russian, but a situation of local inversion also occurs. Inconsistency in word order remains a problematic, pending issue in the statistical equation method. Of course, this idea was said a few years ago. Today, this problem is being solved in the field of world corpus linguistics and automatic translation. A fixed compound that finds its expression in a dictionary should be treated as follows: its translation equivalent would be a word / compound in the dictionary.

For example, in the parallel phrase “Birdan uni negadir odamlar torta boshladi” <> “But now at all once he felt a desire to be with other people” the original word birdan translated with the combination all at once; such a translation is also reflected in the Russian-English dictionary. In a bilingual dictionary, if the order of verbal expression and translation is provided, the transition to the fragmentation stage can be equated with the original and a certain syntagm / compound in the translation.

In the original and translated sentences, which have the same word order, there is an inverted fragment: "изредка толко" <> "only occasionally". It is preferable not to include such inversions in the dynamic programming algorithm, as they require a separate approach. The inversion fragment is extracted, entered into the comparative fragment algorithm, and checked by a critical search engine.

Let's go back to the previous example. The critical search engine will break the original sentence into the following fragments:

1. Ammo endi == But now
2. Endi uni birdan == now all at once
3. Birdan nimadir ... ga tortdi == all at once he felt a desire to be with
4. odamlarga == with other people

The boundary between words and phrases is an issue that needs to be addressed, which is reflected in translated dictionaries, but not in equivalent dictionaries. It is appropriate that such units be expected to be included in the author's language dictionary as an example of translation used for educational purposes. These units are used for educational purposes; then included in the dictionary of the language of the author's works.

The following are some of the "new pairs" found in the translation of N.V. Gogol's "Stories of St. Petersburg" at the stage of equalization of words:

jinoyat <> evil deed; ko'rsatma bermoq <> enjoin; muzlatilgan <> frozen; olib bormoq <> raise; qorovul <> watchmen; qabul qilmoq <> conceive; o’z <> your; qoldirmoq <> neglect; talab qilmoq <> compel; harakat <> impulse; o’ylanmoq <> begin to think; tanishtirmoq <> recur; rangpar <> poor; eshitdi <> hearkened; vaqtichog’lik <> divert; joylashishi <> state; yoqimli <> delightful; ma’lumki <> as every one knows.

ISSN- 2394-5125 VOL 7, ISSUE 17, 2020

V.Conclusion

The widespread use of the parallel corpus opens up new possibilities for access to bilingual linguistic data: the subsequent application of the obtained data; analyze and generalize it; assists in the implementation of human and computer translation, compiling a dictionary and learning foreign languages. The key to using these materials collected from various sources is the linguistic and extralinguistic layout of this text (corpus unit).

The size and level of the corpus annotation (unit mark) is determined by the purpose of the researcher. The large-volume parallel corpus, chosen in proportion to the subject, chronology, and genre, serves to study the general features of the language.

VI.References:

1. The Collected Tales of Nikolai Gogol / translator Pevear R., Volokhonsky L. New York: Pantheon Books, 1998. – 435 p.
2. Crime and Punishment by Fyodor Dostoyevsky [Электронный ресурс] // Project Gutenberg: [сайт]. URL: http://www.gutenberg.org/ebooks/2554
3. The Lady with the Dog and Other Stories by Anton Pavlovich Chekhov [Электронный ресурс] // Project Gutenberg: [сайт]. URL: http://www.gutenberg.org/ebooks/13415
4. British National Corpus (BNC) [Электронный ресурс] // British National Corpus: [сайт]. URL: http:// www .natcorp.ox.ac.uk/
5. Dzemyda G., Sakalauskas L. Optimization and knowledge-based technologies // Informatica. 2009. Vol. 20(2) – P. 165-172.
6. Tiedemann J. News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces // Recent Advances in Natural Language Processing. 2009. Vol V. – P. 237-248.
7. Brown P.F., Della Pietra V.J., Della Pietra S.A., Mercer R.L. The mathematics of statistical machine translation: parameter estimation // Computational Linguistics. 1993. Vol. 19(2). – P. 263-311;
8. Marcu D., Wong W. A phrase-based, joint probability model for statistical machine translation // Proceedings of the Conference on Empirical Methods in Natural Language Processing. Philadelphia, 2002. – P. 87-99;
9. Och F.J., Ney H. Discriminative training and maximum entropy models for statistical machine translation // ACL Anthology: [site]. URL: http://acl.ldc.upenn.edu/P/P02/P02-1038.pdf; Toutanova K., Ilhan H.T., Manning C.D. Extensions to HMM-based statistical word alignment models // Proceedings of Empirical Methods in Natural Langauge Processing. Philadelphia, 2003. – P. 87-94.
10. P.F.Brown, V.J.Della Pietra, S.A.Della Pietra, R.L.Mercer. The mathematics of statistical machine translation: parameter estimation // Computational Linguistics. 1993. Vol. 19(2).
11. Collins M., Koehn P., Kucerova I. Clause restructuring for statistical machine translation // Proceedings of the Association for Computational Linguistics (2005) // Faculty of Humanities - McMaster University: [site]. URL: www.humanities.mcmaster.ca/~kucerov/ ACL2005.pdf; Melamed I. Bitext Maps and Alignment via Pattern Recognition // Computational Linguistics. 1999. Vol. 25 (1). – P.107-130.
12. Gale W.A., Kenneth W.C. A Program for Aligning Sentences in Bilingual Corpora // Computational Linguistics. 1993. Vol. 9(1).
13. P.F.Brown, V.J.Della Pietra, S.A.Della Pietra, R.L.Mercer. The mathematics of statistical machine translation: parameter estimation // Computational Linguistics. 1993. Vol. 19(2)