PARALLEL CORPUSES IN LINGUISTICS ON THE EXAMPLE OF UZBEK, RUSSIAN AND ENGLISH.

Abdurakhmanova Mukaddas Tursunalievna, Abzhalova Manzura Abdurashetovna, Akramjanova Madina Ismatullaevna

Article Info

Article History

Received: October 27, 2021

Accepted: December 08, 2021

Keywords *parallel corpus, machine translation, translation quality, bilingual dictionary, Uzbek-English and Russian-English languages, literary texts, linguistics.*

DOI: 10.52123/zenodo.564443

(ISSN:1068-3844)

Abstract

Annotation: *A parallel corpus is a bilingual (monolingual) corpus, that is, the original text and its translation into some other language, and these two texts not only lie next to each other, but must be aligned: individual fragments of the original must coincide with the corresponding fragments of the translation. This is what makes it possible to use the parallel corpus as a research tool. In language and speech, there are such phenomena, the functioning of which is widespread at almost all levels of the language.*

Relevance of the topic: A parallel corpus or corpus of parallel translations is a corpus consisting of texts in one language together with its translation into another language or languages.

The creation of a parallel corpus includes several stages: text alignment, text markup, search interface design.

The alignment procedure is used primarily to ensure that the fragment corresponding to the original is found in the translation. After that, the same fragments of parallel texts are compared with each other.

The question that arises at the very initial stage is what, in fact, needs to be leveled. You can align word by word, but this often turns out to be almost impossible for a number of reasons: the sets of tokens, stable expressions in different languages do not match. Also, tests are aligned by sentences, but in this case, pitfalls may be encountered: the number of sentences or paragraphs may also not coincide.

The creation and use of parallel corpora seems to be expedient both from a practical point of view and from the point of view of the development of corpus linguistics - one of the most promising linguistic directions. The corpus of parallel texts can be effectively used both in various linguistic studies of a comparative nature and in studies on translation theory, comparative literary studies, cultural studies, and automatic text processing. In this work, we will be primarily interested in the actual linguistic - and including lexicographic - aspects of the use of parallel corpora.

By turning to the corpus, the creators of bilingual dictionaries get a very simple and effective tool for collecting material and empirically testing their hypotheses regarding interlanguage equivalence. The value of this tool is determined by the fact that in linguistics the stage of collecting material is the most time consuming and least creative, and the corpus of parallel texts allows you to significantly save time and effort for the creative stage of work.

Traditional comparative lexicology and bilingual lexicography are characterized by an orientation towards comparing more or less isolated linguistic structures. A negative consequence of such an orientation is insufficient consideration of the usus, that is, those features of the syntactic and compatible behavior of language units that cannot be explained by their systemic features. So, in principle, it is known that this or that structure of one language cannot be translated into another in all contexts with the help of its standard equivalent. In certain contexts, L2 resorts to other ways of describing the relevant situation. It is also known that there are no productive rules by which such deviations from "standard equivalence" could be derived from some more general principles. The only way to describe such deviations is to carefully record them on authentic material. This is the only way to build comprehensive comparative descriptions and create dictionaries that meet modern requirements.

The Parallel Text Corpus is the most adequate tool for performing these tasks. A particular linguistic structure of interest to a researcher can be found in all contexts presented in the corpus with their translations into the corresponding language. Thus, the researcher gets at his disposal a set of authentic contexts representing the structure of interest to him in its natural environment, as well as the most diverse equivalents of this structure in the language of the target. Since these equivalents also turn out to be embedded in natural contexts, based on the materials obtained using the parallel corpus, conclusions can be drawn about the dependence of the choice of an equivalent on the type of context. Such results are almost always at odds with the information that we can glean from traditional dictionaries created in the "pre-corpus era", thus being non-trivial.

An important parameter by which languages can differ from each other is the degree of usage of certain expressions. So, some expression "A" of the L1 language can be translated into the L2 language in a standard way with the help of the expression "B", which is quite correct from the point of view of the norms of this language. Thus, the expressions "A" and "B" are equivalent within the framework of the respective languages. However, their functional equivalence is often incomplete. In particular, this is the case when one of the expressions turns out to be much more common in its language than its translated equivalent in its own. Such cases are well traced when comparing the original texts with their translations into other languages; compare, for example, [Yokoyama 2012: 13].

The study of a certain linguistic phenomenon based on the corpus of parallel texts (especially if the phenomena of the L2 language are considered) can be, in a number of parameters, opposed to the study of this phenomenon on the basis of a large corpus of original texts. The difference between original texts and translations lies both in volume (millions of words of original texts produced every day, versus a relatively small number of texts translated from foreign languages), and in the nature of authorship (that is, in the degree of originality and creative freedom when generating a text), as well as in a cultural context (translated texts are usually immersed in the culture of the source language). All these factors contribute to the distinction between the original and translated texts in a number of ways.

There are no one-to-one correspondences between specific words with similar meanings in different languages (which are considered to be equivalent). In principle, this applies to many semantic word classes. Each specific word has its own unique set of compatibility restrictions and preferences. This idea is one of the traditional theoretical positions of linguistics. Something new that brings the appeal to the corpora of parallel texts to the solution of such problems is the possibility of empirical testing of the corresponding hypotheses on representative material.

Work on parallel corpuses with English was started much earlier than with German. Accordingly, the volume of the Anglo-Russian and Russian-English corpuses far exceeds the volume of the German-Russian and, in particular, the Russian-German. Now in the RNC there are 146 English-Russian and 77 Russian-English texts available online. For more details, see the site http: // www. ruscorpora. ru / mycorpora-para. html. The following parallel texts (works of Russian literature with their English translations) are under preparation:

Братья Карамазовы – The Brothers Karamazov;

Белые ночи – White Nights;

Бесы – The Possessed or, The Devils;

Преступление и наказание – Crime and Punishment;

Звонок – The Doorbell;

Письмо в Россию – A Letter that Never Reached Russia;

Подлец – An Affair of Honor;

Бритва – Razor;

Возвращение Чорба – The Return of Chorb;

Драка – The Fight;

Пассажир – The Passenger;

Путеводитель по Берлину – A Guide to Berlin;

Рождество – Christmas;

Ужас – Terror;

Сказка – A Nursery Tale.

The purpose and objectives of the study. The purpose of the work is theoretical understanding and practical study of the application of elements of corpus linguistics in the creation of English-language linguodidactic materials, as well as the nature of the functioning of the linguistic corpus in an English textbook.

Accordingly, the objectives of our research are defined:

- Collect a theoretical base on the topic "corpus linguistics" and describe its key principles;

- Study the classification and structure of language corpora;

- To trace the use of corpus linguistics in various promising areas, in particular in the field of linguodidactics;

- Consider the main elements of the technology for constructing a textbook on the principles of corpus linguistics;

- Determine the parameters for a comparative analysis of the lexical content of traditional and corpus-oriented textbooks;

- On the basis of the analysis carried out, develop a number of recommendations for the creation of linguodidactic English-language materials.

The subject of the research is presented by the technologies of corpus methods in the development of English-language linguodidactic materials.

The object of the research is corpus linguistics as the science of systematized arrays of linguistic data.

The theoretical and practical significance of the work. The theoretical basis of the research was the work of domestic and foreign scientists: N. B. Gvishiani [22], S. O. Savchuk [24], V. P. Zakharov [23], D. Lich [25] and others. These authors considered corpus linguistics not only as a direction of linguistics, but also explored the possibility of using corpus instruments in linguodidactics.

The practical significance of our work lies in the development of recommendations for the creation of English-language linguodidactic materials, as well as the selection and editing of their lexical content.

Research methodology and methods. As the initial material for the analysis, a large volume of text material from poetic and prose works of classical and modern Uzbek literature, as well as data from translation and explanatory dictionaries of the Uzbek language, was selected.

The work was carried out in line with interdisciplinary research, which is based on the principle of consistency, which determined the use of such general scientific methods as analysis of literature and articles, study and generalization of domestic and foreign practice, comparison, classification and generalization. Empirical research methods, first of all, include the method of strategic sampling, with the help of which a quantitative analysis of the lexical content of textbooks was carried out, which served as the fundamental basis of this research.

The methodological basis of the work was the works of scientists V. Dressler [10], Z.S. Harris [21], E. Agricola [1], O.I. Moskalskaya [17] and others. Computational linguistics and translation theory, such as A.V. Zubov. [12], Karpov V.A. [14], Lekomtsev Yu.K. [15], M.V. Morozkina E.A. [16], and others. Machine translation and its practical application: Borisevich A.D. [4], Belonogov G. G. [2], Nelyubin L.L. [18], and others. Translation modeling: V.G. Gak [6], Yu.I. Gurova [7], S.V. Evteev [11], and others; as well as scientific developments of large IT companies: Microsoft, Google, Yandex, ABBYY.

There are a huge number of languages on the globe (about 7000), and each of them has some features in common with other languages, which we find only in a separate language. If we set ourselves the task of identifying all common features or characteristics of the structure of a certain number of languages, then we will receive the sum of features that will distinguish this group of languages from some other group where these features will not be present.

Common structural features are found in a wide variety of languages that have no genetic relationship. Thus, a definitive phrase, in which an adjective precedes a noun without any agreement with it, is found in English, Turkic, Mongolian languages, in Japanese and Chinese. Comparative typology is called the term "general typology", which deals with the study of common problems associated with identifying the sum of similar and distinctive features that characterize the systems of individual languages of the world. Currently, typological research significantly expands the boundaries of linguistic research, taking them beyond the framework of genetically related languages, makes it possible to attract a wide range of languages of different structures, enriching the material attracted for research and thereby allows solving broad general linguistic problems.

At one time I.A. Baudouin de Courtenay [3] wrote: “We can compare languages completely regardless of their relationship, from any historical connections between them. We constantly find the same properties, the same changes, the same historical processes and rebirths in languages that are historically and geographically alien to each other". The general typology is opposed by a particular typology, which deals with the study of problems of a more limited nature. Depending on the more specific and specific tasks and objects being studied, a particular typology includes a historical or diachronic typology, which is faced with the tasks of studying historical changes in the typology of states of individual languages, the typology of the structure of individual languages and groups of languages, for example, the transition of languages from a synthetic type to an analytical type or a change in the structure of grammatical categories that characterize a given part of speech in the ancient, middle or new period of the history of the language.

At the present stage of its development, theoretical linguistics is characterized by an increased interest in the complex of problems associated with the comparative and typological analysis of the language. At the same time, the focus is on the issues of comparative analysis of both related and unrelated languages.

For comparative typological studies, it is of fundamental importance to establish both structural identities and structural differences of constitutive units and relations between them in two or more compared languages. At the same time, structural identity can be a consequence of both a primordial unified model for the entire group of related languages, and a consequence of a secondary, later convergence of related and unrelated languages due to the action of intrasystemic and extralinguistic factors.

In recent years, comparative, or constructive, linguistics has been developing more and more. She usually studies two languages, rarely more than two, comparing individual elements of one language with the corresponding elements of another. The comparison is carried out synchronously on the basis of a descriptive analysis of different levels of these languages: phonological, grammatical, word-formation, lexico-phraseological, stylistic. The results of a comparative analysis of the two languages show in which elements of these levels they differ.

The Uzbek language and its culture developed as a result of the fusion of dialects and dialects of different local populations, and therefore it is widespread outside the Republic of Uzbekistan, as well as in the Kyrgyz Republic, Tajikistan, Kazakhstan, Turkmenistan, in the northern part of Afghanistan.

E. D. Polivanov [19, p. 194], highlighting three genetically different dialects in the composition of the Uzbek language, outlined the characteristic linguistic features both for each of the three groups and for their subgroups of dialects. Kh. Daniyarov [9] asserts that Uzbeks as a nation were formed in the 11th century from three Turkic ethnic components.

1) Karluko - Chigil (this includes the Turko - Barlas groups),

2) Kypchak (since ancient times they were called "Uzbeks"),

3) Oguz.

The word "Uzbek" appeared in the XI century and even earlier, with its help the people were called, making up the majority of what is called "Turk".

Dialectally, the modern literary Uzbek language is heterogeneous, and, in turn, it is not identical in detail to any of the living dialects and dialects.

It should be noted that the main task of this work is not to consider the entire history of the development of the English and Uzbek languages we are comparing. We are faced with a more definite task: comparison of the functioning of the phenomenon of the parallel corpus in linguistics of two different-structured languages - Uzbek and English at the present stage of their development. Comparison of individual phenomena of the syntactic structure of the Uzbek and English languages is a very difficult task. The difficulty of comparing the linguistic phenomena of the syntactic structure of languages of different grammatical structures is rightly noted by many linguists.

Modern English is one of the Germanic languages belonging to the West Germanic group. It has a large number of territorial dialects: in Great Britain - the Scottish dialect, a group of northern, central, southern and southwestern dialects; in the United States -Eastern-English, mid-Atlantic (central), southeastern, mid-western groups. Dialectal variation is much more pronounced than in the United States, where the central dialect becomes the basis of the literary norm.

Conclusion.

In the tasks of teaching translation, parallel corpora of texts can be considered as abstract information and provide samples of professional translation when studying the techniques and methods of translation. In the tasks of teaching a foreign language, such corpora make it possible to select possible equivalents of the studied vocabulary, to trace its meanings and functions in certain contexts.

Currently, corpora (or parallel texts) of fiction are especially widespread, although for teaching translation at a university it is necessary to develop corpora of different genres and styles and, first of all, focus on scientific, technical, journalistic and business texts.

The analysis of text corpora, methods and developments in corpus linguistics are a promising direction in the field of teaching foreign languages. The world practice of the development of this area proves the effectiveness of this kind of applications, although at present the possibilities of the methods of corpus linguistics in Uzbekistan have not yet been properly implemented in applied linguistics, linguistic teaching, teaching the native and foreign languages.

List of used literature.

- Agricola, E. Micro - medio - and macrostructure as a meaningful basis of the dictionary // E. Agricola. Questions of linguistics. 1984. Issue 2. p. 72-82.
- Belonogov, G. G. Automated processing of scientific and technical information. Linguistic aspects. - M .: All-Russian Institute of Scientific and Technical Information, 1984.
- Baudouin de Courtenay, I.A. On the mixed nature of all languages / I.A. Baudouin de Courtenay. Selected Works on General Linguistics. Volume 1. -M., 1963.-371 p.
- Borisevich A.D. Automatic translation of English construction texts. // In the book: LAAT. p. 279-285.
- Buntman and others. 2014 -, Zaliznyak Anna A., M.,,. Information technologies of corpus studies: principles of constructing cross-linguistic databases // Informatics and its applications. Volume. 8, issue 2, 2014. p. 98-110.
- Gak V. G Language transformations. Some aspects of linguistic science at the end of the XX century. From a situation to a statement, - Edition by Librokom, 2009. - 167 p.
- Gurova Yu.I. Phraseologisms of ancient origin. - St. Petersburg Humanitarian University of Trade Unions. - Innovative views of young scientists '2015. - 5 p.
- Dobrovolsky 2009 - Corpus of parallel texts in the study of culturally specific vocabulary // National corpus of the Russian language: 2006-2008. New results and prospects. SPb .: Nestor-History, 2009. p. 383-401.
- Doniyorov, X. O’z tariximizni bilamizmi? "O’zbek so’zining kelib chiqishi va qo’llanishi haqida.
- Dressler, V. Syntax of text // New in foreign linguistics. / V. Dressler. - M .: Progress, 1978. - Issue VIII: Linguistics of the text. - with. 111-137.
- Evteev S.V. Linguocultural model of translation. // Bulletin of the Bryansk State University. № 2. 2014, p. 3.
- Zubov A.V., Zubova I.I. Information technology in linguistics: a textbook for university students. - M .: Academy, 2004.208 p.
- Yokoyama 2012 - Towards the metatheory of translation: translation as a discourse // Man about language - language about man. Collection of articles of memory. M .: Azbukovnik, 2012. p. 152-159.
- Karpov V.A. Language as a system. - Edition 3 of the USSR Academy of Sciences. 2009.304 p.
- Lekomtsev, Yu.K. An introduction to the formal language of linguistics. - Edition 3 of the USSR Academy of Sciences. 1983.264 p.
- Morozkina, E. A. Hermeneutics in philology, linguistics and translation studies / E. A. Morozkina // Bulletin of the Bashkir University. - 2012. - T.17. - No. 1. - p. 154-157.
- Moskalskaya, O. I. Grammar of the text. / O.I. Moskalskaya. M .: Higher school, 1981 .- 184 p.
- Nelyubin L.L. Translation and Applied Linguistics. - M .: Higher school, 1998 .-- 207 p.

- Polivanov, E. D. Uzbek dialectology and Uzbek literary language. / E.D. Polivanov. - Tashkent, 1933.-194 p.
- Semenov A.S., Modern information technologies and translation M .: Publishing center "Academy", 2008. P. 11
- Harris, Z.S. Joint occurrence and transformation in the linguistic structure // New in linguistics. / Z.S. Harris M .: Foreign Literature Publishing House, 1962. - 209 p.
- Gvishiani NB Workshop on corpus linguistics: a textbook in English / NB Gvishiani. - M.: Higher school, 2008 - 191 p.
- Zakharov V.P. Corpus linguistics: a textbook for students of humanitarian universities / V.P. Zakharov, S.Yu. Bogdanova. - Irkutsk: Irkutsk State Linguistic University, 2011 - 136 p.
- Savchuk S. O. National corpus of the Russian language: prospects for use in linguistic research and teaching / S. O. Savchuk // Bulletin - Vladivostok, 2011 - № 02 (03). - p. 62–67.
- Leech G. Computers in English language research / G. Leech, A. Beale // Cambridge language teaching surveys. – Cambridge: Cambridge University Press, 1985 – Vol. 17(3). – P. 5–18.

Author Information

Abdurakhmanova Mukaddas Tursunalievna
associate professor of the National University of Uzbekistan, candidate of philological sciences.

Abzhalova Manzura Abdurashetovna
Associate Professor, Tashkent State University of Uzbek Language and Literature, PhD

Akramjanova Madina Ismatullaevna
Master's student at the National University of Uzbekistan