International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

Vol. 31 No. 1 February 2022, pp. 270-275

Interpretation Of Lexicographic Research In The Creation Of Parallel Corpus

Anorkhon Nasivali qizi, Akhmedova

Independent researcher, UzSWLU

Abstract – In recent years, the development of corpus linguistics and the increasing focus on statistical methods of processing linguistic materials have led to a number of studies related to the use of parallel or similar texts in different languages. This article is devoted to creating parallel corpus and interpretation of lexicographic research. It is expedient to pay special attention to one of the components of corpus linguistics, which is becoming an integral part of translation studies - parallel texts.

Keywords – Parallel Corpus, Corpus Linguistics, The System Of Automatic Language Processing, Multilingual Parallel Corpus, Machine Translation, Parallel Texts, Separate, Combine, Delete, Add Or Change The Order Of The Sentences.

Although parallel texts are considered a new branch of science within corpus linguistics, its history dates back to BC. In 196 BC, Egyptian churches inscribed texts in honor of the king on a stone dedicated to King Ptolemy V in two languages (Greek and Egyptian). This is a perfect example of what we now call "parallel texts," that is, texts in which one text is translated into a second or more languages. This stone is included in science as the Rosetta stone. Jean-François Champollion's study of the Rosetta Stone in 1822 showed that the hieroglyph is the key to deciphering the text code, and that the study of this inscription has brought about great changes in science and put an end to many debates and myths. Although the Rosetta stone is relatively new to the stone inscriptions, it is more perfect than other stone monuments due to the completeness of the text and the fact that it is presented in two languages in parallel.

It was not until the eighties of the nineteenth century that parallel texts began to be used in the system of automatic language processing. Several attempts to use them were tested in machine translation in the late 50s. However, the limited data warehouse and computer capacity of these years, the storage and computing capabilities of computers, the difficulty of entering large amounts of textual data at that time limited the use of the case.

Later, ideas emerged, such as conducting various studies, such as storing, processing, and editing translation samples into the corps structure. In the late 1970s, the first automatic method of parallel textual alternation of text was developed in 1987 by Martin Kay and Martin Roscheisen[1]. They proposed a number of methods for alternating large volumes and different levels of text, i.e., they developed units that translated each other. These units included *paragraphs, sentences, words,* and *phrases*.

The use of modified parallel texts was suggested by Harris[2]. The theory he proposed involved compiling translation memories, compiling dictionaries, and a list of bilingual terms. Harris theorized that parallel corpses could serve as a source of research for interlingual learning, computer-assisted learning, or comparative linguistics and translation studies.

In our view, parallel corpus is a type of corpus that is used in corpus database research, corpus annotation, use of metadata, and bilingual concordance to facilitate the researcher's work for modern translation studies.

[1] Kay M. &RoscheisenM. (1993). Text-translation alignment. Computational Linguistics, 19 (1), 121-142.
[2] Harris B. (1988b). Bitexts: A new concept in translation theory. Language Monthly, 54, 8-10.

Corresponding Author: Anorkhon Nasivali qizi, Akhmedova

The first step in any parallel corpus creation project is to select the appropriate parallel or comparative text material. The first phase of this project often requires a great deal of patience and time, as it involves copyright clearance, typing or scanning, and error correction. Sometimes data is automatically collected to create parallel and comparative corpora. In this case, mainly WWW (World Wide Web) materials are considered as a real alternative. The *URL*[3] and the parallel *HTML*[4] must be specified. *Resnik*[5] then added an identifier filter to the parallel case.

The importance of being multilingual in the field of language has increased, global markets and the exchange of information around the world, the use of parallel text has required the study of parallel corpus and the improvement of the results obtained from them. Parallel texts are sometimes called bitexts - bilingual or multitext - if the number of languages is more than two languages. Today, the number and coverage of parallel texts is increasing, and due to improved means of storing and archiving electronic documents, various firms and www are using parallel corpuses containing multilingual documents.

Many scientists have conducted scientific research to improve the operations performed in the parallel corps. Among them, Martin Kay and Martin Roscheisen[6], Gale and Church[7], Resnik[8], Brown, Lai and Mercer[9] expressed their views on the stages and ways of creating a parallel corps. According to Key and Roscheisen, in order for words to be compatible in translation, first of all, word equivalence must be achieved, i.e., for lexical mapping, the degree of textuality of the texts must be adjusted in the two languages. In their view, it is not easy to achieve an alternative to text words in two languages, and this alternative may not be adequate, but the study process of the research is facilitated by giving multiple text meanings of the word sought in the parallel corpus.

Another problem in creating a parallel corpus is that if one concept is represented by one word in one language, it can be represented by two or more words in another language. If in the original text the idea is expressed in a short sentence, in the translated text it can be expressed in a long sentence.

The parallel corpus consists of the same text translated into one or more languages. The texts are alternated, i.e. the corresponding segments are connected by sentences. The corpus allows you to search in one or both languages to search or compare translations.

A parallel corpus is a collection of texts, each of which is presented with a set of options translated from the original into one or more other languages. There are 2 types of parallel buildings according to their structure: simple and multilingual parallel buildings.

1. In simple parallel corpus, only two languages are involved: one corpus is an exact translation of the other.

2. In multilingual parallel corpus, texts are provided with translations in several languages.

When creating a parallel corpus, words and sentences are alternated. This process is given by the term alignment in English. To create a parallel corpus, you need to divide the corpus texts into segment units. Segment units are alternate words or sentences that are separable parts of a parallel (multilingual) corpus. An altered word or sentence in the corpus represents information about which segment in one language is the translation of which segment in another language. The shorter the segments, the easier it is to find the translated word or phrase from the segment. Segments are usually given in the form of sentences, but some corpus can be alternated at the paragraph or document level. The easiest way to present alternative data for a parallel corpus is to load the data in tabular form. The spreadsheet is usually prepared using Excel.

[3]*URL* – the resource is the source's Internet address.
[4]HTML, or hypertext markup language, is a programming language used to create the front end of websites. It is written to include the structure of a web page and it is understood that web browsers place it on the websites we see. HTML includes visual elements such as color, font size, and text layout.
[5]Resnik Philip "Mining the Web for bilingual text", in: Proceedings of the 37th annual meeting of the Association for computational linguistics.1999.
[6] Kay M.,Roscheisen M. Text-translation alignment. Computational Linguistics: 1993. 19 (1),– P 121-142.
[7] Gale W.A., Church K.W. A program for aligning sentences in bilingual corpora. Computational Linguistics, 19 (3), 1993. – P 75-102.
[8]Resnik P. Parallel Strands: A Preliminary Investigation into Mining the Web for Bilingual Text, Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA-98), Langhorne, PA, October, 1998.
[9] Brown P.F., Lai J.C., Mercer R.L. Aligning Sentences in Parallel Corpora, Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics. Berkeley, 1991. – P 169-176.

For proper use of the parallel body, it is necessary to adjust the source text and its translation. This means that it is necessary to identify a pair or set of sentences, phrases and words in the original text and their translations into other languages. It is very important to alternate the text in parallel, because in the translation process the translator can *separate, combine, delete, add or change the order of the sentences* to create a natural translation in the target language. The level of correspondence between parallel corpus texts varies depending on the type of text. For example, a literary text can give the translator more freedom.

When compiling parallel cases, the texts of the case are selected according to specific criteria, depending on the purpose of its construction. In particular, you will need to decide whether to include a static or dynamic set of texts, full texts, or text samples. The author should consider size, subject, genre, and style requirements. Any type of housing must meet the following requirements:

1) texts should contain naturally occurring linguistic information;

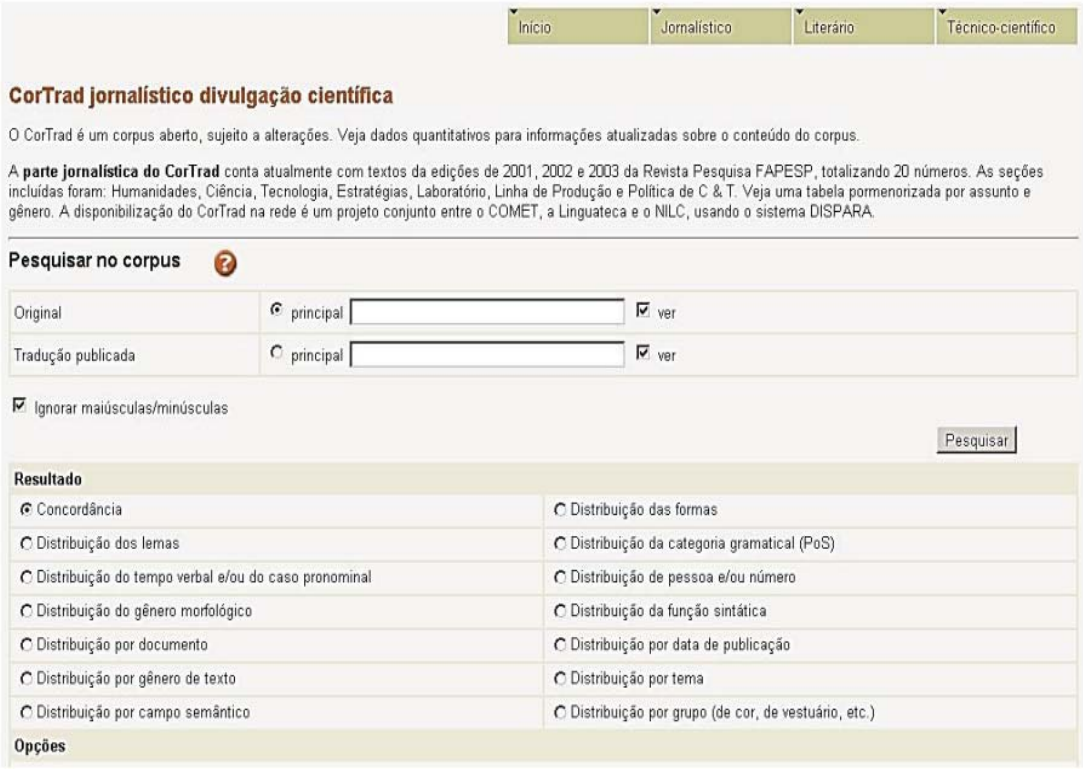
2) it should contain information from different types of speech.

We will focus on some examples of parallel corpus. Created in 2009, *CORTRAD*[10] Translation is a necessary guide for parallel corpus translators and educators. In it, the texts are presented in linguistic pairs in Portuguese and English. *COMPARA*[11] is a two-way Portuguese English parallel corps, created in 2011. Opus Corpus (2012) is an open source parallel corpus.

Among the parallel bodies, the CORTRAD parallel body is distinguished by its multi-functionality. The building is based on the COMET project of the University of São Paulo and has two innovative functions:

1) the ability to compare different variants of the same text. It consists of the original text, a reworked version, and a published translation form;

2) There is a search engine for each text you search. This function provides texts sorted by type.

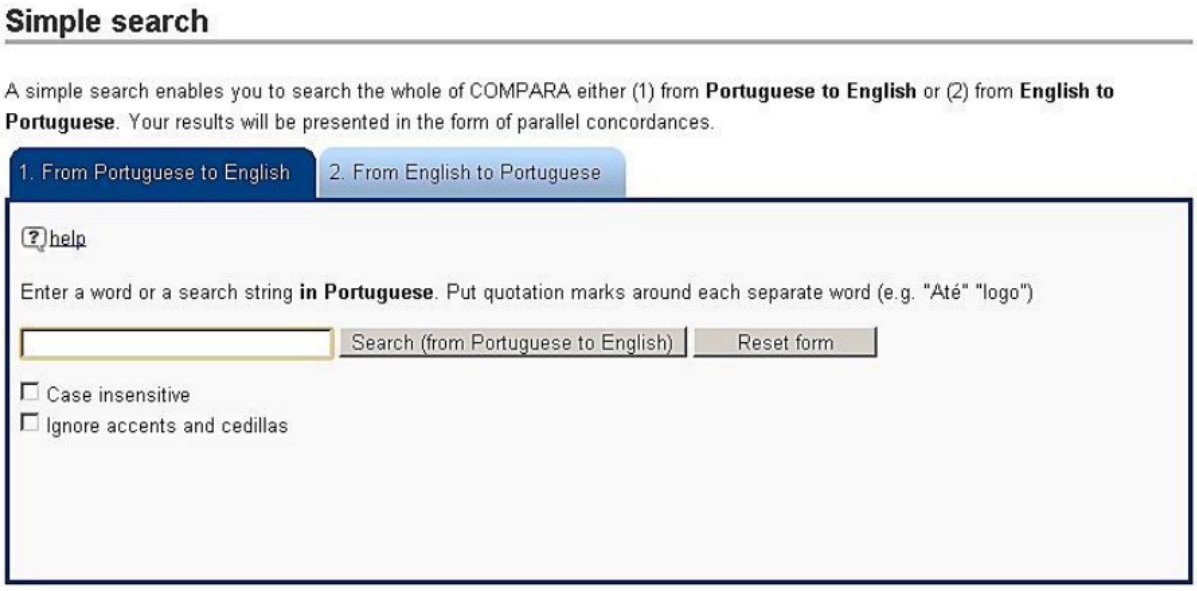


1. picture. The search engine of the *CORTRAD*

[10]Available at: http://www.fflch.usp.br/dlm/comet/consulta_cortrad.html
[11]Available at:http://www.193.136.2.104/COMPARA/index.php

CORTRAD consists of 3 different subcorpora: journalistic, scientific and artistic texts. Uses *CORTRAD DISPARA* software and provides easy access to the system interface.

The *COMPARA* case also uses the *DISPARA* program. This case is an open corpus[12]. *COMPARA* is based on the Linguateca project. This corpus consists of Portuguese-English literary texts. All texts are presented in two language alternatives at the sentence level. Experienced corps researchers and a user with no experience working with the corps can use it equally easily.



1. picture. The search engine of the *COMPARA*

The COMPARA case has an inconvenience for the user, i.e. every word searched must be enclosed in quotation marks. For example: *"help", "yourself".* The next parallel corpus is the Opus[13] corpus, which is comprehensive in terms of areas and languages and includes several subcorporates. In this case, there are many subcorporations and a large database, the searched texts are automatically presented in the interface with parallel text.

[12]An open corpus is a corpus where both the user and the researcher can add text and use their own text analysis.
[13]Available at: http://opus.lingfil.uu.se/

Sub-corpora (downloads & infos):

- ada83 - Ada 83 manuals
- Bianet - Translated Turkish articles
- Bible (uedin) - Collection of Bible translations
- Books - A collection of translated literature
- CAPES - Thesis and dissertation abstracts
- CCAligned - Parallel documents from Common Crawl
- CCMatrix - Parallel sentences from Common Crawl
- DGT - A collection of EU TMs provided by the JRC
- DOGC - Documents from the Catalan Goverment
- ECB - European Central Bank corpus
- EhuHac - Hizkuntzen Arteko Corpusa
- EiTB-ParCC - Parallel Corpus of Comparable News
- Elhuyar corpus
- ELRC public corpora
- EMEA - European Medicines Agency documents
- The EU bookshop corpus
- EUconst - The European constitution
- EUROPARL - European Parliament Proceedings
- EUROPAT - Parallel corpus of patents
- Finlex - Legislative and other judicial information of Finland
- fiskmo - Data from the fiskmö project
- giga-fren - French-English Gigal-Word Corpus
- Global Voices - News stories in various languages
- GNOME - GNOME localization files
- GoURMET - Parallel data from web crawls
- The Croatian - English WaC corpus
- Infopankki
- JRC-Acquis- legislative EU texts
- JW300 multilingual corpus
- KDE4 - KDE4 localization files (v.2)
- KDEdoc - the KDE manual corpus
- MBS - Belgisch Staatsblad corpus
- memat - Xhosa/English parallel data
- MIZAN - A large Persian-English corpus
- MontenegrinSubs - Montenegrin movie subtitles
- MultiCCAligned - Pivot-based Bitexts from CCAligned
- MultiParaCrawl - Non-English Bitexts from ParaCrawl
- MultiUN - Translated UN documents
- MT560 - A large MT dataset for >500 languages
- News Commentary (v11, v9.1,v9)
- OfisPublik - Breton - French parallel texts
- OO - the OpenOffice.org corpus (v2)
- OpenSubtitles - translated subtitles (v1, v2011, v2012, v2013, v2016)
- OPUS-100 corpus
- ParaCrawl corpus
- ParCor - A Parallel Pronoun-Coreference Corpus
- PHP - the PHP manual corpus
- QED - The QCRI Educational Domain Corpus
- Regeringsförklaringen - a tiny example corpus
- The sardware corpus
- SciELO - Artciles from SciELO
- SETIMES - A parallel corpus of the Balkan languages
- SPC - Stockholm Parallel Corpora
- Tanzil - A collection of Quran translations
- Tatoeba - A DB of translated sentences
- TedTalks hr-en
- TED Talks 2013
- TED Talks 2020
- The tico-19 corpus
- The Tilde MODEL corpus
- TEP - The Tehran English-Persian subtitle corpus
- Ubuntu - Ubuntu localization files
- UN - Translated UN documents
- UNPC - The United Nations Parallel Corpus
- WikiMatrix - Parallel sentences extracted from Wikipedia
- Wikimedia - Documents from the Wikimedia Translation project
- Wikipedia - translated sentences from Wikipedia
- WikiSource - (small en-sv sample only
- WMT News Test Sets
- The Xhosa - English Navy corpus

3 picture. Subcorpora of the Opus corpus.

In addition to the files displayed on this webpage, *Opus* also provides pre-configured word matching and phrase tables, bilingual dictionaries, and frequency counts, making it possible to find these files through a resource search form on the *Opus* top-level website. *Opus* corpus results can be downloaded to any computer provider.

The creation of machine translation systems, automated and automated dictionaries, terminological databases requires the study and analysis of parallel texts, i.e. a set of texts translated from one language to one or more other languages. When selecting a series of parallel texts, it is necessary to solve the following tasks:

1) to study the linguistic machine and to determine the exact goals and objectives of the creation of a parallel corpus;

2) create samples separated from the total number of texts. In this case, to approximate the stylistic homogeneity / heterogeneity of the general text and its limitation / infinity;

3) auxiliary dictionaries - determination of the possibility and necessity of compiling frequency dictionaries and alternative dictionaries;

After solving these problems, we can move on to the direct analysis of parallel texts.

Creating effective dictionaries of input and output languages is the main source for accurate translation of vocabulary when creating a parallel corpus.

The parallel corpus of analogies we are exploring offers a number of possibilities for corpus users. In this case, the translation options of analogies in two languages on the basis of parallel texts are given. In this case, the user will be able to quickly find the simile he needs, and will be able to combine several equivalent translations, as well as synonyms and antonyms.

In short, the creation of dictionaries of literary translation is a requirement of the time, in which the translation of the masterpieces of our literature into foreign languages, the study and analysis of translation techniques based on them will give great results. Of course, the user is required to have the skills to use dictionaries of different views at different stages of translation.

References

[1] Harris B. (1988b). Bitexts: A new concept in translation theory. Language Monthly, 54, 8-10.

[2] Resnik Philip “Mining the Web for bilingual text”, in: Proceedings of the 37th annual meeting of the Association for computational linguistics. 1999.

[3] Kay M.,Roscheisen M. Text-translation alignment. Computational Linguistics: 1993. 19 (1), – P 121-142.

[4] Gale W.A., Church K.W. A program for aligning sentences in bilingual corpora. Computational Linguistics, 19 (3), 1993. – P 75-102.

[5] Resnik P. Parallel Strands: A Preliminary Investigation into Mining the Web for Bilingual Text, Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA-98), Langhorne, PA, October, 1998.

[6] Brown P.F., Lai J.C., Mercer R.L. Aligning Sentences in Parallel Corpora, Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics. Berkeley, 1991. – P 169-176.