ISSN NO: 2770-2367
Date of Publication: 25-04-2022

Challenges of compiling a parallel corpus

Gulhida Hakimova Abdulaziz qizi
Uzbek State University of World Languages
Comparative Linguistics and Linguistic Translation
Scientific adviser: Anorkhon Akhmedova Nasibaliyevna
Lecturer at the University of World Languages of Uzbekistan

Annotation. This article is aimed to feature parallel corpus and the process of compiling it. In particular, the main focus of interest is the challenges that a compiler can come across during the creation of a parallel corpus.

Keywords: Corpus Linguistics, Parallel Corpus, Copyright, Custom Interface, Digitization

Today, there are more than 6,500 languages, and the growing demand for interaction among these languages is increasing the need for a computerized database of each language, or at least one of the most widely used languages in the world. Therefore, in recent years, a new branch of linguistics - corpus linguistics has developed rapidly. Corpus linguistics has undergone a series of changes over the past few decades, from a “little donkey cart” to a “railroad car”. (Leech 1991: 25).

The study and practice of parallel corpus is one of the branches of the computer based linguistics or corpus linguistics. A parallel corpus is a corpus that contains a collection of original texts in one language and their translations into one or more other languages. In most cases, the parallel corpus contains data from only two languages. Although it is a new field, many scientists have conducted numerous research in this field, including, the Scandinavian scholars Sagvall Hein and Axelsson Berglund, who conducted their research mainly on a comparative analysis of the original text with several foreign languages.[1]

The parallel corpus can be bilingual or multilingual, i.e. they consist of texts in two or more languages. They can be one-sided (e.g., English text translated into German), two-sided (e.g., English text translated into German, and vice versa), or multi-directional (e.g., text translated into German, Spanish, French, etc).

“The process of creating parallel corpus begins with the selection of corpus texts according to specific criteria that depend on the purpose for which they were created. In particular, corpus developers need to be able to define a static or dynamic set of texts and include whole texts or text samples. ”[2] Issues of authorship, size, subject matter, genre, medium, and style should be considered consistently. In any case, the corpus must meet the following requirements:

a) it must contain real (naturally occurring) language information;

b) it should be representative, i.e. it should contain information from different types of speech.

One of the biggest issues once these above mentioned criteria are met is the use of copyrighted texts in this corpus. At first glance, this may seem like an easy task, but it will take a lot of time and effort. The most difficult task for the developer of a parallel corps is to explain to the owners of the text the content and the objective of their scientific work, as well as for whom the corpus is expected to be built, giving full information about the parallel corpus. Without going

[1] Lars Borin.Parallel Corpora,parallel words: Selected papers from simposium on parallel corpora and comparable corpora at Upsala University, Sweden. 1999. B.12.

[2] Fayez Gebali .Algorithms and parallel computing wiley series on parallel and distributed computing. A John Wiley & Sons, Inc., Publication . 2011.B. 25

through this part, the assembly of the corps is likely to lead to major problems later on as the authors of the original texts can sue the corpora creators for copyright infringement.

The next most important issue in this process is the ordering of the corpus texts by text, paragraph, and sentence. Although this is not a problem for placing and arranging the original text, it does pose a daunting task for the corpus developer in translating these texts. Because the clear boundary between sentences is broken during the translation process, sometimes two simple sentences can be translated by a translator into a single sentence, and the corpus linguist cannot change the original text or the translated text at will. Even though this can be handled somehow, in some parallel corpuses some sets of phrases should be identified in both languages, such as adjectives, verbs and some idiomatic phrases, which really challenge the creators seeing that while translation some words are just omitted or the meaning of the target word is partially or completely lost. So, the process of tagging the texts for specific types of parallel corpuses , such as parallel corpus of synonyms, antonyms can be time consuming and less accurate.

Then the process of digitization of texts begins, and at the same time any characters, page numbers that are not important to the reader and translators are deleted, and then special characters are added to the text. These can include nicknames in the text, foreign words, translators, or author's notes. These characters determine exactly what language unit will be the main research area of the compiler. It should be noted that this process is often performed by specialized programmers and requires direct and indirect intervention of the developer. Given that the industry is new, finding a good specialist will play a big role in demonstrating the full effectiveness of the work done so far. This process is continued by sorting the text units, such as paragraphs and sentences. The text units are selected according to the compiler's goals, and in the final stage, all materials are placed on special corpus interfaces in a special IMS Corpus Workbench format. One of the most common of these is www.linguateca.pt/COMPARA.

To sum it up, corpus linguistics, which is gaining popularity among linguists today, and especially their compilation, is a work that requires great effort. This process consists of the organizational and practical parts of the process prior to the initial technical process, and both of these interrelated processes require in-depth knowledge, patience, and time in the field from the developer. But the ready-made parallel corpus serves as an original resource not only for researchers but also for learners of the languages used in the corps.

List of used literature

1. Lars Borin.Parallel Corpora,parallel words: Selected papers from simposium on parallel corpora and comparable corpora at Upsala University, Sweden. 1999.
2. Fayez Gebali .Algorithms and parallel computing Wiley series on parallel and distributed computing. A John Wiley & Sons, Inc., Publication . 2011.
3. Libo Huang. Style in Translation: A Corpus-Based Perspective. China. 2007.
4. https://www.researchgate.net/publication/47397240_Compilation_and_Exploitation_of_Parallel_Corpora/link/00b7d51876da9c0ce6000000/download
   http://www.glottopedia.org/index.php/Parallel_corpus
5. https://benjamins.com/catalog/persons/196111643