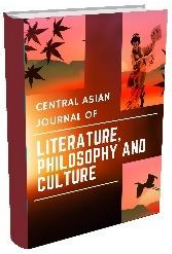

CENTRAL ASIAN JOURNAL OF LITERATURE, PHILOSOPHY AND CULTURE

eISSN: 2660-6828 | Volume: 04 Issue: 06 June 2023
https://cajlpc.centralasianstudies.org

Uzbeki-English Parallel Corpus Algorithm and Alignment Problem

Botir Boltayevich Elov
Doct. of phil. in technical sciences (PhD), associate professor, Computational Linguistics and Digital Technology, Tashkent state Uzbek language and University of Literature

Ma'rufjon Amirkulov
Compuatational Linguistics II stage graduate student, Tashkent state Uzbek language and University of Literature

Received 4th Apr 2023, Accepted 5th May 2023, Online 14th June 2023

ANNOTATION

This article discusses the stages of creating an Uzbek-English parallel corpus and proposes its model algorithm, relying on the achievements of world corpus studies and the latest technologies for creating parallel corpora.

KEYWORDS: corpora, development, corpus, linguistics, algorithm, symbols.

Corpus studies has become one of the most pressing issues in computational linguistics today. In particular, attention is being paid to multilingual parallel corpora, and various studies are being conducted on its creation. In world corpus studies, monolingual corpora were initially created, and the initial research and development processes of corpus and corpus linguistics were mainly carried out within the framework of English linguistics. The Brown corpus created by scientists N. Francis and G. Kucher in the 1960s left a mark in history as the first machine-readable corpus, and the same can be said about the Lancaster-Oslo/Bergen (LOB) corpus created after itIn turn, the creation of corpora that improves along with the development of corpus linguistics is one of the urgent problems of modern linguistics. Nowadays, the use of corpora plays a leading role in many linguistic studies.

In the process of researching many parallel corpora, it can be observed that different approaches were used in their creation. In the early years, when matching the texts in the corpus, aligning the text on the sentence level was preferred, while in the corpora of recent years, aligning the text on the word level is also used. Different Aligners, i.e. (sentence or word or both)matching programs, can be used depending on the cross-section of the cases. Hunalign, FastAlign, E-Align, InterText programs are mainly designed for sentence alignment, while Giza++ program is used for word alignment in modern corpora. When parallel corpora between certain languages are in the stage of creation for the first time, it has been approved by many representatives of the field that the sentence level alignment method should be used in the processing of corpus data. As an example,

we can see the parallel corpora created in the Arabic-English language pair so far. Most of these corpora were completed with a sentence level matching:

	Egypt Corpus	**The English-Arabic Parallel Corpus**	**OPUS**	**Arabic-Spanish-English Parallel Corpus**	**Quranic Arabic Corpus**	**EAPCOUNT**	**LDC**
Size (words)	77,430	3 million	352 million	3 million	77,430	5,5 million	40 million
Availability	x	x	✔	✔	✔	x	x
Medium	Written	Written	Written	Written	Written	Written	Written and spoken
Source	The Holy Quran	The World of Knowledge (a series of translated texts)	OpenOffice.org documentation KDE manuals including KDE system messages PHP manuals	UN documents	The Holy Quran	UN documents	Broadcast conversation, traditional broadcast news and newswires
Alignment level	Sentence	Sentence	Sentence	Sentence	Sentence	Paragraph	Sentence and word
POS tagging	x	x	x	✔	✔	x	x
Stemming	x	✔	x	x	✔	x	x

Figure 1. *Examples of English-Arabic parallel corpora.*

In addition, based on the table, it can be noted that in order to increase the coefficient of perfection and usefulness of some cases, operations such as stemming and POS tagging were also used. Therefore, a perfect and up-to-date parallel corpus algorithm, which can be useful for users in every way, should include the following processes: text collection, "cleaning", extralinguistic annotation, sentence level alignment , validation, tokenization, lemmatization, POS tagging, word level alignment.

If we classify these processes as an algorithm, we may encounter the following sequence when creating a parallel corpus of Uzbek-English languages:

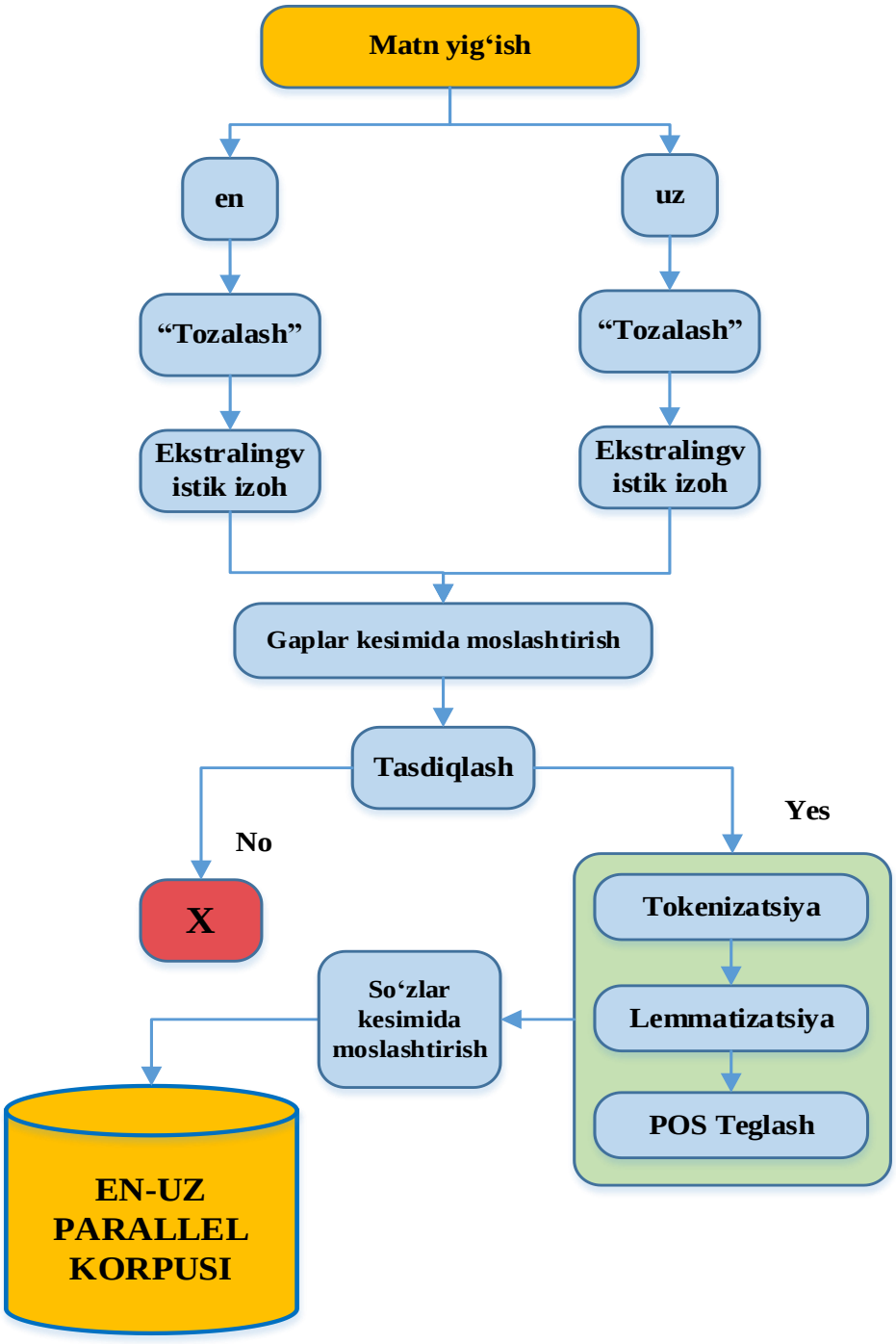

Figure 2. *The proposed algorithm for the Uzbek-English parallel corpus*

PROCESS OF SOURCE/TEXT COLLECTION

If the corpus is a mirror showing all the possibilities of the language, representativeness is the transparency of this mirror. A linguist should keep in mind the practical significance of the phenomenon of representativeness

in the process of using the corpus, as well as in the process of creating it, that it is the main factor in creating a corpus. When choosing a source for a parallel corpus, it is necessary to refer to quality texts. As much as possible, the information covered by the corpus should be diverse in terms of genre and type. However, this process is not so easy to implement. Because it is a difficult task to find translations of different genres and types from one language to another. For example, if the translation of journalistic and official texts from English to Uzbek or from Uzbek to English is carried out on a large scale, the same cannot be said about the number of artistic translations. However, when creating the first parallel corpus in a language, we found it advisable to choose high-quality literary translations as a source in order to show all the charm of that language. For this reason, for the Uzbek-English parallel corpus, the works of Abdulla Qadiri's "O‘tkan kunlar", Cholpon's "Kecha va Kunduz", and O'tkir Hashimov's "Dunyoning ishlari" are acceptable choices.

"CLEANING" STAGE

In this process, texts are formatted as UTF-8. After that, the texts are cleaned of unnecessary symbols and tags. This process can be done automatically by Python codes.

EXTRALINGUISTIC TAGGING

Extralinguistic data can include information about the name of the text, author, year, style, region, source language, translator. Such information is implemented through XML encoding and is represented by special tags. Each corpus can specify the number of these extralinguistic data according to the purpose of its creation. Thus every corpus may be unique in terms of these factors. Below we can see an example of extralinguistic and linguistic information attached to the text in XML format:

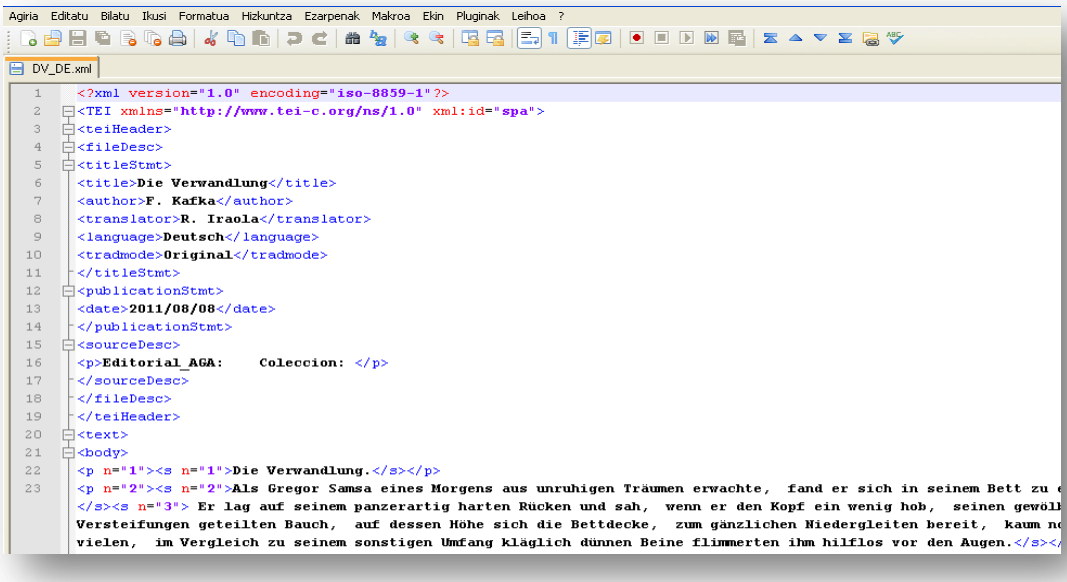

Figure 3. *Extralinguistic and linguistic tagging of data in the corpus.*

SENTENCE LEVEL ALIGNMENT

For many created parallel corpora, we can observe that the alignment of texts on sentence level is taken as a basis. Various programs and software tools can be used for this process. Mainly 3 different methods are used

for aligning sentences. These include heuristic (rule-based), statistical and hybrid (statistical and rule-based methods together) methods. Various programs have been created based on these methods:

Heuristic or rule-based (LF Aligner, YASA, FASTALIGN).

Statistical (Giza++ (mainly used for word alignment), Hunalign).

Hybrid (LASER, SALT).

Heuristic or rule-based sentence matching algorithms are algorithms that use linguistic rules of particular languages to match sentences in different languages. These methods do not require large parallel corpora for training, but rely on linguistic and structural similarities between the languages.

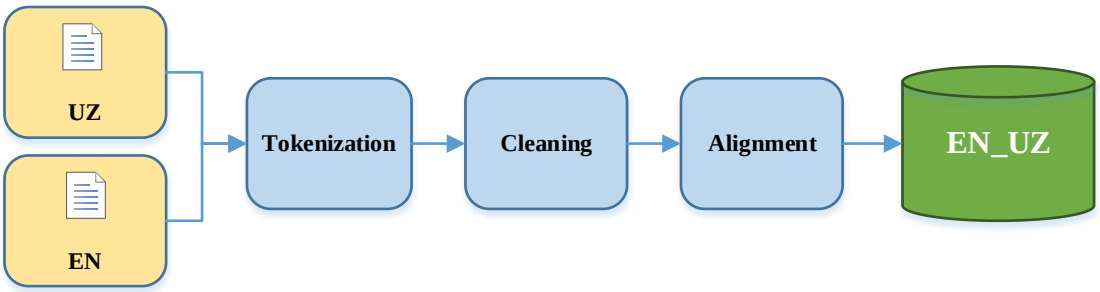

Figure 4. *Sentence matching sequence based on heuristic algorithm.*

Statistical sentence matching algorithms are algorithms that use statistical methods to calculate and match similarities between sentences in different languages. These methods are often based on word-level matching models. In fact, programs created in this way can rely not only on statistical methods, but also on heuristic or rule-based methods, depending on the level of adaptation. As an example, we can cite the Giza++ program. Giza++ first relies on heuristic methods to match the data in the corpus on a sentence level, while this program then uses statistical methods to match the data on a word level. It is not wrong to say that the data of the parallel corpus contained in the national corpus of the Russian language has been implemented an alignment of sentences and words based on the Giza++ program.

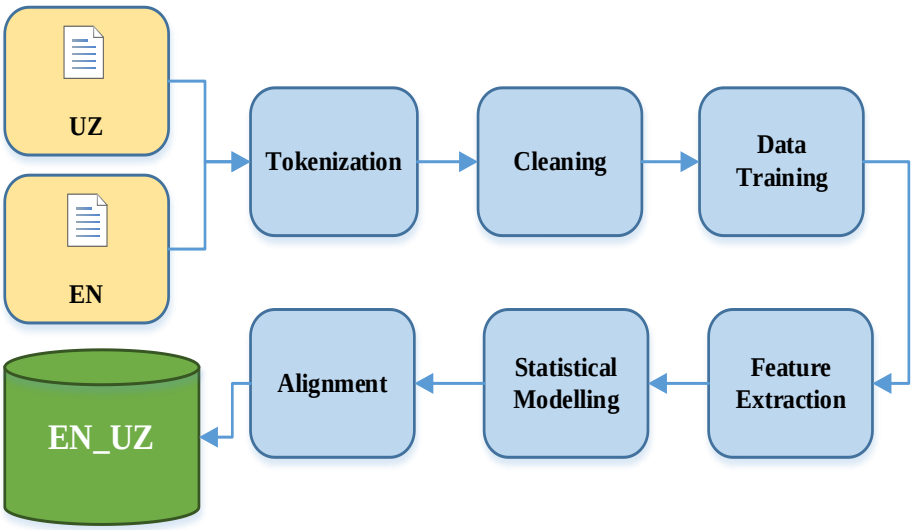

Figure 3. *A sequence of sentence alignment based on a statistical algorithm.*

Most of the programs created on the basis of heuristic methods are designed for matching on a sentence level, and we can find free and open source versions of such programs through various Internet networks.:

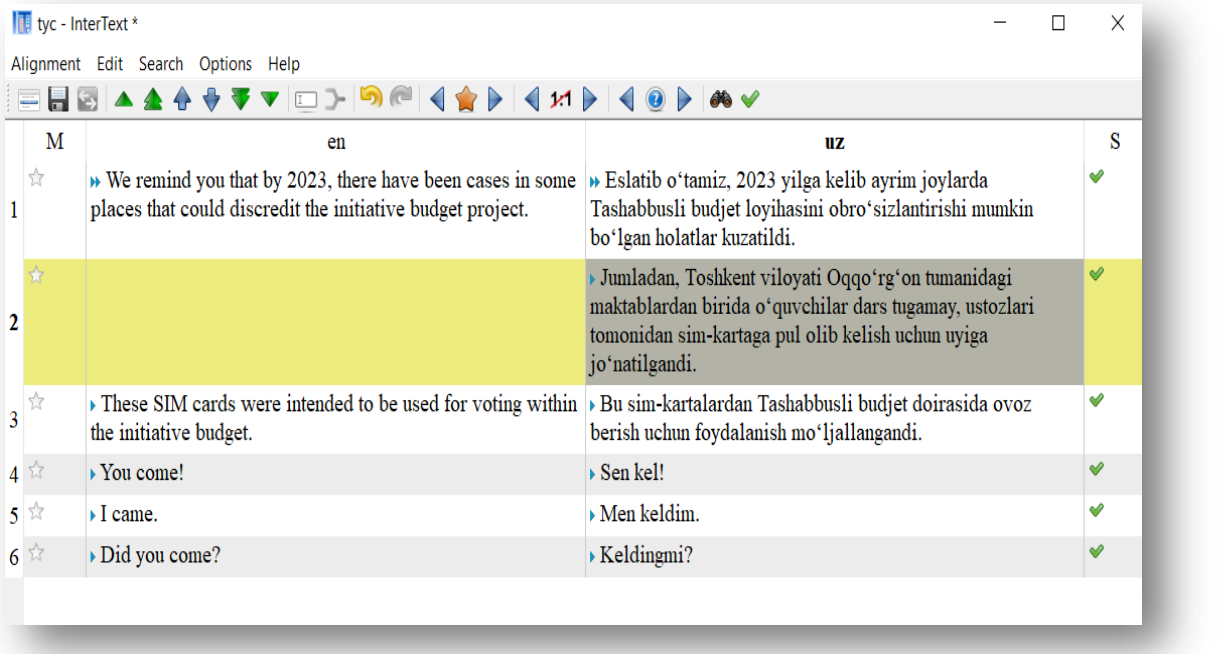

Figure 3. *Alignment (InterText) performed on the basis of a heuristic algorithm*

VALIDATION

In this stage, the quality of automatically aligned sentences in two languages is checked using the human factor. In particular, untranslated sentences and disproportionate alignments are removed. Only satisfactory translations will be moved to the next stage to take up space in the corpus.

TOKENIZATION, LEMMATIZATION, POS TAGGING STAGE

This process can be done automatically. It is possible to perform the mentioned actions through the site https://uznatcorpara.uz, created by the team of the Department of Computer Linguistics and Digital Technologies of Alisher Navoi Tashkent State University of Uzbek Language and Literature, and these steps perfectly serve as an important and necessary element in creating a parallel corpus of Uzbek-English languages. It is worth noting that the existence of these automatic processes greatly speeds up the creation of the Uzbek-English parallel corpus, and as a result, it becomes possible to tag large volumes of data in a short period of time.

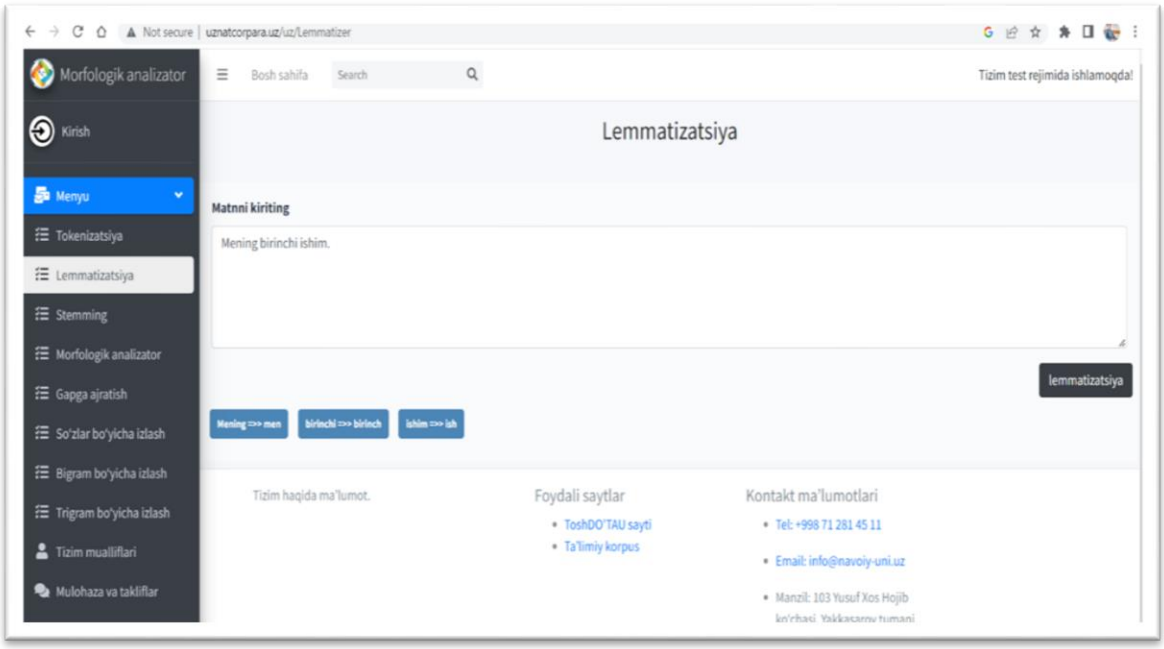

Figure 7. *uznatcorpara.uz web page interface.*

WORD LEVEL ALIGNMENT

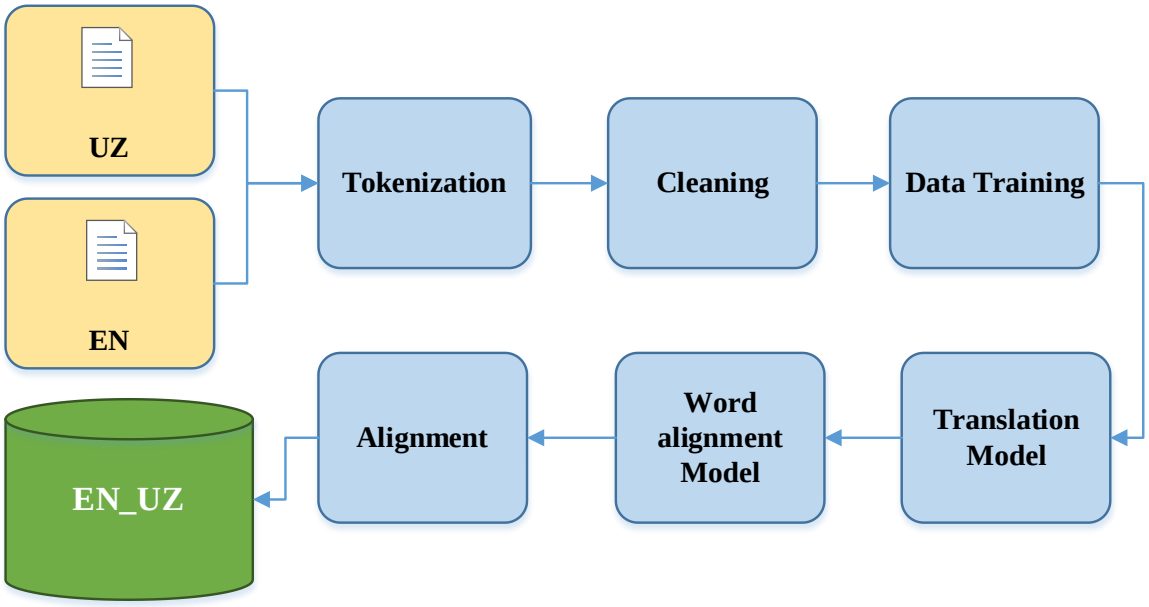

Figure 8. *Word alignment sequence based on statistical algorithm.*

Word alignment is a relatively new concept for parallel corpora, which means that we can see the active participation of this element in recently created corpora or updated versions of older corpora. A number of programs have been created to match words, and you can find versions specific to specific languages or designed for any language. The Giza++ program is also designed for most languages in the world, and based

on many studies and the opinion of famous scientists, we found it to be the most appropriate choice for the Uzbek-English parallel corpus.

It is interesting to say that the Giza++ program based on the above algorithm is used in many large, well-known corpora, including the parallel corpora of the national corpus of the Russian language. The listed steps are considered very important steps in the process of creating a perfect parallel corpora. The number, sequence and composition of these stages may change depending on the purpose of the corpus. We hope that the Uzbek-English parallel corpus, which is expected to be created on the basis of an abovementioned algorithm created as a result of many scientific researches, will serve as a perfect and useful resource for users.

LITERATURES

1. Alia Al-Sayed Ahmad, Bassam Hammo and Sane Yagi. ENGLISH-ARABIC POLITICAL PARALLEL CORPUS: CONSTRUCTION, ANALYSIS AND A CASE STUDY IN TRANSLATION STRATEGIES. Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 3, No. 3, December 2017.
2. I.Shamsudinova. Kelishik-ko‘makchi paradikmatik munosabativa uning o‘zbek milliy korpusida ifodalanishi. Filologik tadqiqotlar: til, adabiyot, ta’lim. 2023; 2
3. Svartvik, J. 1992. Corpus linguistics comes of age. In J. Svartvik (ed.), Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991. Berlin: Mouton de Gruyter, 7-13.
4. Захаров В.П., Богданова С.Ю. Корпусная лингвистика: учебник для студентов гуманитарных вузов. – Иркутск: ИГЛУ, 2011. – С. 18.
5. https://ruscorpora.ru/
6. https://uznatcorpara.uz