USING TOOLS FOR ESTABLISH FORMAL RELATIONS IN A PARALLEL CORPUS BETWEEN UZBEK AND GERMAN LANGUAGES

Rakhimova Sh.S.

Termiz State University, Teacher of the Department of Romano-Germanic Languages

https://doi.org/10.5281/zenodo.11521281

Abstract. *This article is about using tools for establish formal relations in a parallel corpus between Uzbek and German languages. Evaluating the quality of the formal relations established in the parallel corpus. This could involve manual validation by linguists or using automatic evaluation metrics to assess the quality of the alignment and formal relations. A collection of syntactic and morphological annotations for many languages, including German and Uzbek.*

Keywords: *corpora, parallel corpus, formal relations, subcorpora, linguistic annotations.*

Introduction. A parallel corpus is a collection of texts, each of which is translated into one or more other languages than the original. The simplest case is where two languages only are involved: one of the corpora is an exact translation of the other. Some parallel corpora, however, exist in several languages.

Then we can say a corpus is parallel if the corpus contains source texts and translations in parallel, or it is a comparable corpus if its subcorpora are comparable by applying the same sampling frame.

Research Methodology. Establishing formal relations in a parallel corpus between Uzbek and German languages involves several steps:

Collecting Parallel Data
Cleaning and Preprocessing
Aligning Sentences
Building a Parallel Corpus
Establishing Formal Relations
Annotating Relations
Evaluation and Validation
Creating Resources

Collecting Parallel Data: Gathering a collection of texts that are translations of each other in Uzbek and German. This could be done by searching for translated texts, such as official documents, books, articles, or websites that have versions in both languages.

Cleaning and Preprocessing: Cleaning the data to remove any irrelevant information, formatting, or errors. Preprocess the text by tokenizing, removing punctuation, and converting the text to lowercase.

Aligning Sentences: Aligning sentences in the Uzbek text with their corresponding translations in the German text. This can be done manually or using alignment tools such as GIZA++, fast_align, or other alignment software.

Building a Parallel Corpus: Once the sentences are aligned, compile them into a parallel corpus, where each sentence in one language is paired with its translation in the other language.

Establishing Formal Relations: Identifying formal relations between the sentences in the parallel corpus. This could involve identifying sentence-level correspondences, word alignments, and syntactic correspondences between the two languages.

Annotating Relations: Annotating the identified formal relations in the parallel corpus. This could include marking word alignments, syntactic dependencies, and other linguistic features that establish formal relations between the two languages.

Evaluation and Validation: Evaluating the quality of the formal relations established in the parallel corpus. This could involve manual validation by linguists or using automatic evaluation metrics to assess the quality of the alignment and formal relations.

Creating Resources: Finally, creating resources such as aligned parallel corpora, dictionaries, and linguistic annotations that can be used for various natural language processing tasks such as machine translation, cross-lingual information retrieval, and language understanding.

Literature review. Corpus is electron version of collected texts used for teaching over 30 years. This methodological approach has effective influence on the process of education. Amount of investigations conducted on corpus linguistics according to the sphere of domain and purposes has developed for a few years [6, 132].

The technology of creating a corpus is covered in scientific studies. A. Baranov provided information on the basic concepts of corpus linguistics, text corpus, problem area, database, corpus data storage units, research corpus, illustrative corpora, methods of displaying and storing dynamic and static text corpus [1, 34]. The following are the parallel corpora available in the world computer linguistics system:

1. English-German translation corpus

2. English-Norwegian Parallel Corpus (ENPC)

3. English-Swedish parallel corpus (ESPC). It was created in 1993. It has now become an important resource for learning English and Swedish. The corpus database contains 64 English texts and translations, 72 Swedish texts and translations. The corpus contains 2.8 million words. The texts are adapted as much as possible in terms of text type, topic, and style, so they can be used as a two-way parallel corpus and as a comporative corpus. Served as a resource for research on epistemic modality and adverbial conjunctions in English and Swedish.

4. International Telecommunication Union Corpus (English-Spanish)

5. The Intersect Parallel Corpus (English-French)

6. Multilingual parallel corpus (Danish, English, French, German, Greek, Italian, Finnish, Portuguese, Spanish, Swedish texts) [2, 122].

Analysis and results. Throughout this process, it's important to ensure the quality and consistency of the aligned parallel corpus and the formal relations established between the languages.

Results. Here are some tools and resources that might be helpful:

GIZA++: A popular tool for word alignment in parallel corpora.

fast_align: Another tool for word alignment.

UDPipe: A tool for training and applying neural network models for tokenization, tagging, lemmatization, and parsing.

Universal Dependencies: A collection of syntactic and morphological annotations for many languages, including German and Uzbek.

By following the steps outlined and utilizing the appropriate tools, you can establish formal relations in a parallel corpus between Uzbek and German languages. Here's a simplified version of those steps:

Collect Parallel Data: Gathering texts translated between Uzbek and German.

Clean and Preprocess Data: Removing irrelevant information and format the text uniformly.

Align Sentences: Aligning corresponding sentences in the parallel texts.

Build Parallel Corpus: Compile aligned sentences into a parallel corpus.

Establish Formal Relations:

Word Alignment: Aligning words in parallel sentences.

Syntactic Annotation: Annotating syntactic structures.

Semantic Annotation: Annotating semantic relations.

Annotate and Analyze: Annotating the corpus with linguistic information and analyze formal relations.

Format Corpus: Choosing a standard format like TMX or parallel text format.

Store and Distribute Corpus: Storing the corpus and consider making it publicly available.

Tools such as GIZA++, fast_align, UDPipe, and Universal Dependencies can assist in this process.

Conclusion. By following these steps and using these tools, you can establish formal relations in a parallel corpus between Uzbek and German languages.

The use of parallel corpora in instructional activities is common. They are able to create educational materials, analysing the original source during the teaching process, and recognising common linguistic phenomena in the language with the use of the parallel corpus. Parallel corpora are regularly employed, particularly throughout the language learning process.

Parallel corpora are useful for studying foreign languages, translating literary works from other languages into Uzbek, and teaching Uzbek as a foreign language. This corpus provides linguistic and material assistance for the development of educational corpuses and the national corpus. Using parallel corpora is crucial for understanding and identifying nuances in linguistic meaning.

REFERENCES

1. Баранов А.Н. Введение в прикладную лингвистику. – М.: Эдиториал УРСС, 2001. С-176.
2. Musurmankulova Sh. The Significance of Parallel Corpuses in Language Teaching. CENTRAL ASIAN JOURNAL OF THEORETICAL AND APPLIED SCIENCES Volume: 04 Issue: 05 | May 2023 ISSN: 2660-5317. 121-123. https://cajotas.centralasianstudies.org
3. Karshieva, B. F. "METHODOLOGICAL MODEL OF TEACHING TECHNICAL STUDENTS IN ENGLISH ON THE BASIS OF INTEGRATED BILINGUAL EDUCATION." interdiscipline innovation and scientific research conference. Vol. 1. No. 9. 2023.
4. Fakhriddinovna, Karshieva Bogdagul. "The current state of teaching english to technical students, methodological approaches: the current state of teaching english to technical students, methodological approaches." (2023): 368-374.

5. Қаршиева, Боғдагул. "Муҳандислик йўналиши битирувчисининг чет тили коммуникатив компетентлиги модели." (2022).
6. Abdurakhmonova, Nilufar. (2019). Corpus Based Teaching Uzbek As A Foreign Language. IRAL - International Review of Applied Linguistics in Language Teaching. 6. 131-137.