International Journal of Education, Social Science & Humanities.
Finland Academic Research Science Publishers
ISSN: 2945-4492 (online) | (SJIF) = 7.502 Impact factor
Volume-11| Issue-12| 2023 Published: |22-12-2023|

GENERAL CONSIDERATIONS ON CORPUS LINGUISTICS

https://doi.org/10.5281/zenodo.10207502

Ataboyev Nozimjon Bobojon o'g'li
Associate Professor of Bukhara State University, PhD
Astanova Gulnora Maqsudovna
Master student of Bukhara State University

Annotation

Lexical units of one language do not always have exact analogues in another language. Therefore, to solve this problem, the translator often uses translation transformations (conversions). Their skillful use ensures the adequacy of the translation: the translated text accurately reflects the content of the source text[1].

Key words

text corpus, corpus linguistics, methods and mathematical statistics, machine translation, linguistic phenomena, stress and intonation

INTRODUCTION

In linguistics, *a corpus (pl.: corpora)* or text corpus is a dataset, used as a basis for language research. They are used for statistical analysis and testing statistical hypotheses, confirming linguistic rules in a given language. A corpus of texts is the subject of research in corpus linguistics.

Corpus linguistics is one of the popular areas of linguistics. Her task is a description of language as it appears in written and oral speech, which is represented by a specially selected corpus of texts. “The object of study of corpus linguistics is a corpus of texts, representing the original written speech material. Subject researches are mechanisms for using data sets that intended for research"[2]

LITERATURE ANALYSIS AND METHODOLOGY

Research is based on text corpora. Language is studied from the side of written speech according to characteristics that can be called objective, that is, those that can be measure and operationalize. Probabilistic methods and mathematical statistics are used for processing. The work is carried out with linguistic data in the form in which they are used in the context.

[1] Denina O.O. Orenburg State University E-mail: olga-oren@mail.ru
USING TRANSLATION TRANSFORMATIONS TO ACHIEVE ADEQUATE TRANSLATION
[2] Введение в корпусную лингвистику [Электронный ресурс] // Режим доступа: http://www. myshared. ru/slide/472948/

International Journal of Education, Social Science & Humanities.
Finland Academic Research Science Publishers
ISSN: 2945-4492 (online) | (SJIF) = 7.502 Impact factor
Volume-11| Issue-12| 2023 Published: |22-12-2023|

The following main tasks are identified that are solved using text corpora:

1. Automatic extraction of information about language from texts;

2. Information processing;

3. Verification and interpretation of processed data[3].

Corpus linguistics allows you to check the results and conclusions of certain speech studies and conduct new, more extensive and systematic ones.

DISCUSSION AND RESULTS

A corpora is the main repository of knowledge and information in corpus linguistics. The corpus is also used in the following important areas:

Language technology, natural language processing, computational linguistics

The process of processing and analyzing various corpora also includes many topics in computational linguistics, speech recognition, and machine translation, where it is often used to construct hidden Markov models for speech tagging and other purposes. Corpora and frequency lists derived from them are useful for language teaching. Corpus can be seen as a type of foreign language writing aid, as corpora of languages that are not the user's native language are a source of contextual grammatical knowledge acquired through original texts. They can be effectively used to compose sentences and form texts using them.

Machine translation

Multilingual corpora specially formatted for side-by-side comparison are called structured parallel corpora. There are two main types of parallel corpora containing texts in two languages. A translation corpus is a corpus of texts in which texts from one language have been translated into another language. A comparable corpus is a set of texts that are similar in genre and content, but are not translations of each other. Identification of text segments (phrases or sentences) and matching and coordination of text types and genres are the first conditions for using parallel texts. Between Two Languages Machine translation algorithms for translating between two languages are often implemented using parallel parts consisting of a first language corpus and a second language corpus, which is an elemental translation of the first language corpus.

Philology

Text corpora are also used in the study of historical documents (historical texts), such as deciphering ancient writings or secretarial science. Some archaeological corpora can live for a short period of time, allowing for timely photography. The 15- to 30-year-old Amarna script (1350 BC) may be one of the

[3] Корпусная лингвистика [Электронный ресурс] // Режим доступа: http://corpora.iling. spb. ru

International Journal of Education, Social Science & Humanities.
Finland Academic Research Science Publishers
ISSN: 2945-4492 (online) | (SJIF) = 7.502 Impact factor
Volume-11| Issue-12| 2023 Published: |22-12-2023|

shortest corpora. A corpus of an ancient city (e.g. the "Kültepe Texts" of Turkey) may appear through several corpora sorted by the date of their discovery.

Using statistical research methods using corpora, one can either confirm or refute assumptions about linguistic phenomena. To solve the problems that researchers set for themselves, the presence of a corpus is not enough. The text must have linguistic information. This is how the idea of a marked body came about. Markings help to calculate the frequency of words and the frequency of representatives of different parts of speech. Linguistic markup is used to assigning a code (tag) to a word, which denotes a set of grammatical features that describe the word. Markups are conventionally divided into linguistic and external linguistic [6]. Externally linguistic ones include:

- markup reflecting the text formatting features (headings, paragraphs, indents, etc.);

- markings related to information about the author and the text. The author may indicate his name, age, gender, years of life and others, and the text may indicate the title, language, year and place publications and so on. Such information allows for detailed searches in corpora and provides tools to facilitate the identification of a particular document.

Linguistic types of markup include morphological, syntactic, semantic, anaphoric and prosodic markup. The markup implies the presence of a set of tags, characteristics of their meanings, as well as the principles by which tags are assigned to language units. The larger the set of tags, the more detailed text analysis can be carried out. As the volume of the corpus increased, the number of tags began to decrease. Simplification of the encoding procedure leads to a reduction in the number of errors, inconsistencies, as well as the removal of morphological ambiguity, and the rapid marking of a large array of texts (containing more than a million words).

- Morphological markings. Serves as the basis for subsequent stages of analysis, including syntactic and semantic analysis.

- Syntactic markup. The markup is based on a grammar of component structures.

- Semantic markup. Applies a code that consists of characteristics of the general

semantic group and narrower subcategories[4].

[4] Корпусная лингвистика [Электронный ресурс] // Режим доступа: http://corpora.
iling. spb. Ru

International Journal of Education, Social Science & Humanities.
Finland Academic Research Science Publishers
ISSN: 2945-4492 (online) | (SJIF) = 7.502 Impact factor
Volume-11| Issue-12| 2023 Published: |22-12-2023|

- Anaphoric marking (pronoun). Difficult to automate. Most specialized programs analyze text in sentences, so it gets lost coherence of the text produced as a result of the procedure. Such systems would give more accurate results if they correctly determined the reference (the relationship between a word and a thing that is designated by this word, that is, is its referent) substitute pronouns.

- Prosodic marking. Describe stress and intonation. This marking is accompanied by discourse marking, which serves to mark repetitions, clauses, and so on.

Text markup is done using software, which reduces labor costs. For anaphoric and prosodic markings, creating such software tools is a difficult task, so most of the work is done manually. The software used requires post-editing, which occurs manually (morphological homonymy and syntactic ambiguity), since the programs present a certain number of solution variations, and the researcher himself chooses the desired one. In the future there is a complete automated marking process.

CONCLUSION

To sum up, text corpora provide an opportunity for a researcher to skip the time-consuming step of collecting texts. Corpora help to identify patterns, phenomena based on speech material. The development of computer technology has simplified and accelerated the processes of linguistic processing of large texts.

REFERENCES:

1. Denina O.O. Orenburg State University E-mail: olga-oren@mail.ru
USING TRANSLATION TRANSFORMATIONS TO ACHIEVE ADEQUATE TRANSLATION

2. Введение в корпусную лингвистику [Электронный ресурс] // Режим доступа:
http://www. myshared. ru/slide/472948/

3. Корпусная лингвистика [Электронный ресурс] // Режим доступа: http://corpora.iling. spb. Ru

4. Фрэнсис У.Н. Проблемы формирования и машинного представления большого корпуса текстов / У.Н. Фрэнсис // Новое в зарубежной лингвистике. Выпуск XIV. Проблемы и методы лексикографии. - М.: Прогресс. -1983. - С. 334 – 335

5. Brown Corpus: [электронный ресурс]: http://clu.uni.nO/icame/brown/bcm.html#bc3

International Journal of Education, Social Science & Humanities.
Finland Academic Research Science Publishers
ISSN: 2945-4492 (online) | (SJIF) = 7.502 Impact factor
Volume-11| Issue-12| 2023 Published: |22-12-2023|

6. The Lancaster-Oslo/Bergen Corpus of British English, LOB: [электронный ресурс]: http://clu.uni.no/icame/manuals/LOB/INDEX.HTM

7. British National Corpus, BNC: [электронный ресурс]: http://www.natcorp.ox.ac.uk/