academic publishers
INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE (ISSN: 2692-5206)
Volume 04, Issue 05, 2024
Published Date: 26-07-2024

CORPUS STUDY IN LEXICOGRAPHIC RESEARCH

Latipova Gulasal Bahrom qizi
Doctoral student of Alisher Navoi Tashkent State University of Uzbek Language and Literature

Abstract: Technologies, especially the results obtained in artificial intelligence and machine learning or deep learning prove that the language model and its knowledge based data can be created through a corpus. This article contains analytical materials related to scientific theories conducted in this field. The theoretical ideas included in this analysis will serve as a source for practical work to be carried out for the Uzbek language in the future.
Key words: corpus lexicography, artificial intelligence, thesaurus, WordNet.

A corpus-based approach to terminographic research has been expressed in a number of scientific studies. In this regard, it is worth noting the work done by N. Abdurakhmonova on the Uzbek language. In the scientist's work, "terminography (terminological lexicography) is important in the study of scientific and technical terms in the automatic translation of texts and in the translation of their alternative versions. notes that the relevance of multilingual terminology is explained by the emergence of rapidly growing automated dictionaries and terminological data banks. Also, it is important to structure terminological dictionaries in the above-mentioned source and design them in the database. Terminography in machine translation is studied separately because it is the main part of linguistic support, in machine translation, separate sectoral dictionaries of terms are created, in improving the quality of translation, because terminology is an ever-changing field, this direction is seriously researched in developed countries. , but states that computer terminology is hardly studied in Uzbek linguistics.

N. Abdurakhmonova showed that the following issues should be studied in terminography: 1) non-existence of this or that term in a second language; 2) expression of terms in the form of word combinations; 3) determining the style of the text and changing the meanings of scientific terms according to the style; 4) the use of the same lexeme in different meanings within a specific field. 5) expression of a certain term in different forms in several areas

The field of terminology has a great role in the study of technical terms in texts and automatic translation of their alternatives. Creating dictionaries of different forms of terms is studied in the field of terminography (terminological lexicography). "The relevance of multilingual terminology is explained by the emergence of rapidly growing automated dictionaries and terminological data banks." According to information, the amount of all scientific and technical terms of the German language in the 20th century was 1 million. By the middle of the 20th century, it was determined that the terms related to electrical engineering reached 4 million.

Structuring terminological dictionaries and designing them in the database is one of the important issues. Terminology has a special place in machine translation. The creation of field dictionaries of terms in machine translation increases the quality of the translation. Terminology is always changing. The following issues are studied in terminology:

1) non-existence of this or that term in a second language;
2) expression of terms in the form of word combinations;
3) determining the style of the text and changing the meanings of scientific terms according to the

style;

4) use of the same lexeme in different meanings within a specific field;

5) expression of a certain term in different forms in several areas.

For example, the English term pin used in the technical field has the following meanings as a noun phrase: finger, pin, pin, splint, etc., and this lexeme also has different meanings as another word phrase: to spoil his life, hang on his word, etc.

Scientists express their views on the issue of standardization of terms related to science and technology. It cannot be denied that a certain term is used instead of some of them under the influence of its creator or social events. Providing a full glossary of terms in the linguistic support serves to increase the quality of the translation.

Lexical units of scientific texts can be divided according to the following categories:

- words accepted in the literary language: prepositions, auxiliary words, pronouns
- words accepted in the general literary language, as a rule, have a specific meaning in a scientific text: part, condition, ability
- words that are very rare in non-scientific texts and show the character specific to a scientific text: analyze, classify, method, neuro.
- phraseological combinations: to satisfy a need, influence of components...
- special terminological units within the field of science: probability, set, formal language.
- units that are not found in the scientific text and are accepted in the general literary language as the subject of a specific science: aboriginal, absurd.

Many people are interested in how much lexical richness of the modern language is included in the creation of machine translation systems. According to the information provided by the international organization INFOTERM (International Terminology Center in Vienna, Austria), the number of terms related to European languages is 50 million, and the number of product (goods) names reaches 100 million. This means that any product or innovation created in society will have a specific name and will enrich the lexicography as a term.

Logos (LOGOS), Engspan (ENGSPAN) and Systran (SYSTRAN) machine translation systems use the transfer method.

Founded in California in 1964, Sistran's translation system has 27 language combinations. Words have a general and technical meaning (as a term) and are created on the basis of syntactic and semantic analysis of word combinations. The semantic classification system consists of a hierarchical structure of 500 categories. One of the important tasks of terminology is the correct translation of ambiguous terms. In the Sistran system, special programs have been developed for the transfer stage of the translation process. According to it, it is first determined in the text analysis whether the word is alive or inanimate. A word in the dictionary is coded as [+animate / inanimate]: (for example, seal (1) seal; 2) seal, stamp), acquaintance (1) familiar; 2) like dating). In thematic glossaries, the word being translated can have several meanings in the second language. For example, in English, the term stem is used in several fields and is translated differently in the second language: in botany, plant stem - "plant root", in anatomy, brain stem - "cerebral vein", in linguistics, word stem - "word stem". . In this system, after the user selects one of the ten fields attached before the translation, the translation process is carried out. This gives an opportunity to classify words reserved for a specific field. In this case, corpora are important for the analysis of word combinations with nouns. In this regard, significant results have been achieved in the LOGOS machine translation system. In this system, more than five hundred special internal dictionaries for various fields are included to identify semantic ambiguities in words. It contains more than 130 hierarchically structured semantic categories. Also VISTA (English-Russian, Russian-English) machine translation system 3 mln. It contains 300 thousand lexical units. 80% of them are between two and seventeen words. In addition, a thematic dictionary of more than 400,000 lexical units is included. The possibility of translating English texts into Uzbek is expanded by including the terminological dictionary as a linguistic resource for machine translation, by encoding their general and specific meanings. For this, it is necessary to include the terminological dictionaries of two languages in the software, provide their semantic analysis and standardize the terms within the selected fields. For example, it is desirable to ensure uniformity of terms in social and humanitarian sciences. Regarding this issue, issues of standardization of terms are being studied by the International Organization

for Standardization (ISO-International Standardization Organization). At this point, we acknowledge that the theoretical foundations of such a separation have been confirmed in a number of other works.

According to A. A. Reformatsky, terminology is a systematized field of a specific science, appropriately reinforced in verbal expression. In this respect, it is necessary to study terminology as a separate field in machine translation. The formation of terms is carried out by three methods: semantic, morphological, syntactic. The formation of terms by the semantic method is characterized by the specialization of metaphoric and metonymic meanings in connection with other terminology. This includes the phenomena of transposition and transmutation. Suffixation, prefixation, conversion, and base assimilation are the leading phenomena in forming a term using the morphological method. The role of the transformational method is significant in creating a term using the syntactic method. In addition, concretization, generalization, compensation, transliteration, explanation, amplification, differentiation of meaning, etc. events should not be overlooked.

References:

1. Atkins B. T. S. and A. Zampolli Computational Approaches to the Lexicon, Oxford University Press, 1994. — 496 p.
2. Vincent B.Y. Computer corpus lexicography. Edingurg university press, 1998. – P. 37.
3. Krishnamurthy, Ramesh. 2006. “Corpus Lexicography,” January. https://doi.org/10.1016/b0-08-044854-2/00416-8.
4. 4. N. Abdurakhmonova, I. Alisher and R. Sayfulleyeva, "MorphUz: Morphological Analyzer for the Uzbek Language," 2022 7th International Conference on Computer Science and Engineering (UBMK), Diyarbakir, Turkey, 2022, pp. 61-66, doi: 10.1109/UBMK55850.2022.9919579.
5. N. Z. Abdurakhmonova, A. S. Ismailov and D. Mengliev, "Developing NLP Tool for Linguistic Analysis of Turkic Languages," 2022 IEEE International Multi-Conference on Engineering,Computer and Information Sciences (SIBIRCON), Yekaterinburg, Russian Federation, 2022,pp.1790-1793,doi:10.1109/SIBIRCON56155.2022.1