HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

FORMATION OF THE LINGUISTIC BASE OF THE PARSING PROGRAM OF THE ELECTRONIC CORPUS OF THE UZBEK LANGUAGE THEORETICAL AND PRACTICAL FOUNDATIONS

Choriyeva Hilola Murodovna
Tashkent State University of Oriental Studies,
Master's degree in Computational Linguistics (70230401)
hilolachoriyeva@gmail.com
Doi: https://doi.org/10.5281/zenodo.23369698

Annotation. The article analyzes the development of a syntactic analysis (parsing) program for the electronic corpus of the Uzbek language and the formation of its linguistic base. The agglutinative properties of the Uzbek language, the system of affixes and the linguistic problems that arise in the process of parsing are studied. Also, the interdependence of the stages of text tokenization, morphological analysis and automatic tagging (POS-tagging) is revealed. The results of the research have an important practical value in the integration of the Uzbek language into the digital environment, improving machine translation and natural language processing (NLP) systems.

Keywords: computer linguistics, corpus linguistics, parsing, syntactic analysis, morphological analysis, automatic tagging, Uzbek electronic corpus, natural language processing (NLP).

On a global scale, the rapid development of digital technologies and artificial intelligence systems sharply increases the importance of the field of Natural Language Processing (NLP). Computer analysis of the language, automatic determination of the grammatical and syntactic structure of texts has become one of the priorities of modern linguistics. In the Republic of Uzbekistan, special attention is paid to preserving the purity of the state language, enriching it and ensuring its active integration into modern information technologies and communications.

From this point of view, the adaptation of the Uzbek language to the digital environment, the creation of electronic corpora and the development of systems of automatic morphological and syntactic analysis (parsing) are among the urgent scientific tasks. Syntactic analysis is the central component of natural language processing systems, which provides the necessary basis for determining sentence structure, modeling grammatical relationships and semantic interpretation. However, the morphological richness, agglutinative nature and word-formation capabilities of the Uzbek language require special approaches in the parsing process. The purpose of this article is to create a parsing program for the electronic corpus of the Uzbek language and to scientifically substantiate the linguistic basis that ensures its effective operation.

BASIC SECTION

Fundamentals of Computer Linguistics and Natural Language Processing show that the process of automatic text analysis involves several stages: text segmentation, morphological analysis, syntactic analysis, semantic analysis, and pragmatic analysis. Since Uzbek belongs to the group of agglutinative languages, words are constructed through many suffixes, and several grammatical meanings can be expressed within a single word. For example, the word "from those who did not come" is made up of

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

several grammatical units (kel + ma + gan + s + you + from), such a complex structure requires a special approach for morphological analyzers.

1. Architecture and linguistic base of the parsing program. Parsing program is a complex linguistic-software system designed for automatic natural language processing, which analyzes the text structurally, morphologically and syntactically. When developing a parsing program for the Uzbek language, the following factors should be taken into account:

• Corpus Base: Systematization of Real Speech and Written TextsAn electronic corpus with a resurrected set forms the basic linguistic base of the parsing system.

• Multilevel analysis: The system should be limited not only to the syntactic level, but also to the phonetic, lexical and morphological levels.

• Standardt formats: Storing analysis results and corpus data in CoNLL-U or JSON formats facilitates automatic processing.

2. Morphological analysis and automatic tagging (POS-tagging). The variety of word forms and the strict sequence of affixes in the Uzbek language ensure the priority of the morphological analysis stage. The process of morphological analysis includes the following stages:

1. Tokenization and parsing of text.
2. Segmentation of the root and affixes of a word.
3. Lemmatization and restoration of the vocabulary.
4. Using Automatic Tagging (POS-tagging) definition of parts of speech and grammatical categories.

3. Syntactic parsing models. In modern computer linguistics, two main models of syntactic analysis are widely used: constituency parsing (structural analysis) and dependency parsing (dependency parsing). For the Uzbek language, the dependency parsing model is more convenient because, although the word order in the language is relatively free, the grammatical relationships between words are clearly defined by affixes and rules of control. This model makes it possible to describe the "head-to-head" relationships between words in a sentence in the form of a clear network.

CONCLUSION

The creation of a parsing program for the Uzbek language electronic corpus and the formation of its linguistic base is one of the most urgent tasks of national computer linguistics. The analysis carried out shows:

1. The agglutinative nature of the Uzbek language requires high precision of morphological analysis and automatic tagging.
2. Electronic Corpus Data and Linguistic Rules Their compatibility ensures the stability and effectiveness of the parsing program.
3. The developed linguistic model and parsing technologies will be used in the futurea machine translation serves as an important scientific and practical base for the creation of automatic editing systems and educational programs.

REFERENCES

1. Abdurakhmonova N. Z. Computer Linguistics. – Tashkent, 2023.

HSR (LONDON),
HOUGHTON STREET REVIEW JOURNAL
ISSN: 2058-5519|Impact factor-15.9
http://houghtonstreetreview.com/index.php/hsr

2. Abdurahmonova N. Z. Computer models of the electronic corpus of the Uzbek language. – Tashkent: Science, 2022.
3. Jurafsky D., Martin J. H. Speech and Language Processing. – Pearson, 2023.
4. Manning C. D., Schütze H. Foundations of Statistical Natural Language Processing. – MIT Press, 1999.
5. Mengliyev B. Issues of Uzbek Linguistics and Computational Linguistics. – Tashkent, 2021.