SEMI-AUTOMATIC PART-OF-SPEECH ANNOTATING FOR BELARUSIAN DICTIONARIES ENRICHMENT IN NOOJ

Yu. Hetsevich[1], V. Varanovich[2], E. Kachan[1], S. Lysy1, I. Reentovich[1]

[1] United Institute of Informatics Problems, Minsk, Belarus

[2] Belarusian State University, Minsk

e-mails: yury.hetsevich@gmail.com, gamrat.vvv@gmail.com, evgeniakacan@gmail.com, stanislau.lysy@gmail.com, mwshrewd@gmail.com,

Lexicology is the most flexible level of language, which is continuously changing, updating, improving and at the same time it responds to changes in the reality. Recently, in connection with the development of science, culture, technology and industry the vocabulary of the Belarusian language has changed, manifested in the appearance of new words and phrases denoting new objects, phenomena, concepts. New words reflect the current state of the vocabulary.

Our research is aimed to develop the algorithm for the Belarusian main dictionaries enrichment in NooJ on the basis of the first one-million corpus for the Belarusian NooJ module. From the broad list of possible subject categories the corpus focuses on fiction, historic, medical, scientific, sociological literature and etc., which is considered to be the finest source for searching unknown words of different domains [2]. NooJ as a linguistic development environment gives an opportunity to develop and expand high-quality linguistic algorithms for text processing at the level of lexicology [1].

The task is to work out a mechanism or algorithm for further processing (annotating different categories and paradigms according to flexion classes) of all extracted unknown words (neologisms, dialectal words, subneutral words, colloquial words, jargon words and various foreign words), and adding to the present dictionary on the basis of the Belarusian NooJ module.

So, for this purpose two specific algorithms for automatic generation of word paradigms based on the given grammatical information and automatic stress arrangement for all forms of entire word is planned to develop.

References

1. Max Silberztein. 2014. Nooj Manual. http://www.nooj4nlp.net. Date of access : 31.12.2014.

2. Reentovich, I. First One Million Corpora for Belarusian NooJ Module / I. Reentovich, Yu. Hetsevich, V. Voronovich, E. Kachan, H. Kozlovskaya // International Scientific Conference on the Automatic Processing of Natural-Language Electronic Texts "NooJ'2015" : Abstracts (11–13 June, 2015, Minsk, Belarus) / ed. B.M. Lobanov, Yu.S.Hetsevich. — Minsk : UIIP NASB, 2015. — C. 54-55.