SEMI-AUTOMATIC PROOFREADING OF BELARUSIAN AND ENGLISH TEXTS USING NOOJ

N. Drahun, Y. Shynkievich, Dz. Dzenisiuk, A. Bakunovich, Yu. Hetsevich,
United Institute of Informatics Problems of the NAS of Belarus, Minsk
e-mail: ndrahun@gmail.com, silenoschestra@gmai.com, d.dzenisiuk@gmail.com,
bakunovich.andrei@gmail.com, yuras.hetsevich@gmail.com

It is very difficult to proofread big electronic texts because it takes a lot of time and efforts to proofread text carefully. Such necessary conditions as rewriting unclear sentences, correcting grammatical, punctuation, spelling mistakes, formatting citations, footnotes, references etc. may be an obstacle on the way to correctly outputted texts in the short terms.

To proofread Belarusian texts there are several on-line services, for example on *corpus.by* web-portal. For Belarusian, we could use *Character Frequency Counter*, which allows finding mistakes in usage of Cyrillic and Latin letters and punctuation. Then, we could check texts with *Short U Spell Checker*, which shows mistakes in usage of letter *ў*. At the end, we could use *Spell Checker*, which find unknown words. However, the Belarusian language need special tool and technics that will cover all language formalism in the point of view proofreading.

To find the way and tools for semi-automatic search mistakes in electronic Belarusian big texts, we collected the special corpus of medical domain texts. The text corpus was created by getting news from the website of medical organization. The corpus contains around 270 texts regarding medical sphere with correctly marked foreign characters and punctuation marks, cites and contextual disambiguation.

In current project, we proposed exactly number of steps using NooJ approch for text proofreading that saves time and energy of editors. Steps cover alphabet checking, foreign characters marking, words location in different dictionaries and other linguistic phenomena, which are not solved in usual Belarusian and other proofreading systems.

References

1. Max Silberztein. 2018. Nooj Manual. http://www.nooj4nlp.net . Date of access : 30.12.2014.
2. Computational platform for electronic text & speech processing // http://corpus.by/ [Electronic resource]. — 2019. Mode of access : http://corpus.by/. — Date of access : 31.01.2019.
3. Language Tool . — 2017. https://www.languagetool.org/ 02.04.2017.