FIRST ONE MILLION CORPORA FOR BELARUSIAN NOOJ MODULE

Reentovich, Yu. Hetsevich, V. Voronovich, E. Kachan

In this report first 1 million corpus for Belarusian NooJ module isrepresented. The given corpus has been built up of texts, patched up intosections of different subject lines. From the broad list of possible subject lines inthe sections the corpus focuses on fiction, historic, medical, scientific,sociological literature and etc. And if being of the view that there is a great manyof analogous subject lines, then this first 1 million corpus can be considered asthe first subject collection of texts for Belarusian NooJ module.

The text corpus that is used in NooJ will be effective for the researchactivity development on the following respects:

1) the words polysemy processing in texts of different subjects;

2) the polysemic punctuation marks processing;

3) the new lexical items search.

Besides, the 1 million corpus will be for all intents and purposes applicablefor solving many crucial tasks:

in general

– use this corpus in a linguistic development environment called NooJ [1] tooptimize and expand the development of high-quality linguistic algorithms for the electronic texts pre-processing TTS block;

in particular

– conduct several experiments in order to specify at the minimum and, possibly, maximum level of various syntactic and morphological grammarsusing effectiveness for texts of each subject section;

– take thorough measures in order to create the subject domain generator(which will be then very useful for the formation of special subject-oriented NooJ dictionaries);

– in the most extent use the given corpus in the process of text-to-speech synthesis with the help of available programs [2], required for such process, and also when testing newly created applications;

– make comparative analysis of this corpus with the same corpora in other languages (taking into account all necessary rules, language features in texts of each current corpus, various possible emerging issues, while building syntactic and morphological grammars, etc.).

It is very essential that the first 1 million corpus for Belarusian NooJ module can be completely applicable in any line of linguistic research. And in the near future the corpus is planned to be expanded up to approximately 5–10 million words.

References

1. NooJ: A Linguistic Development Environment [Electronic resource]. – 2015. – Mode of access: http://www.nooj4nlp.net/. – Date of access: 08.02.2015.

2. Corpus.by // Corpus.by [Electronic resource]. – 2015. – Mode of access: http://www.corpus.by/. – Date of access: 08.02.2015.