UNITED INSTITUTE OF INFORMATICS PROBLEMS
OF THE NATIONAL ACADEMY OF SCIENCES OF BELARUS

International Scientific Conference on the Automatic Processing of Natural-Language Electronic Texts “NooJ’2015”

NOOJ 2015

Abstracts

June 11–13, 2015, Minsk, Belarus

Minsk
UIIP NASB
2015

УДК 004.91

International Scientific Conference on the Automatic Processing of Natural-Language Electronic Texts "NooJ'2015" : Abstracts (11–13 June, 2015, Minsk, Belarus). – Minsk : UIIP NASB, 2015. – 80 p.
ISBN 978-985-6744-89-4.

This volume contains the abstracts of the International conference "NooJ 2015". The research presented covers different aspects of natural language processing using NooJ, including formalizing such levels of linguistic phenomena as syllabification, phonemic and prosodic transcription, multiword units and discontinuous expressions, local and structural syntax; transformational syntax and paraphrase generation, semantic analysis and machine translation, etc.

Abstracts are published in the form presented by authors.

У дадзеным зборніку прадстаўлены тэзісы дакладаў Міжнароднай канферэнцыі "NooJ 2015". Разглядаюцца розныя аспекты апрацоўкі натуральнай мовы з выкарыстаннем лінгвістычнага асяроддзя распрацоўкі NooJ, улічваючы фармалізаванне такіх напрамкаў лінгвістычнага аналізу як склададзяленне, фанетычная і прасадычная транскрыпцыі, устойлівыя выразы і дыскрэтныя слоўныя канструкцыі, лакальны і структурны сінтаксісы, трансфармацыйны сінтаксіс і перафразаванне, семантычны аналіз і машынны пераклад і г. д.

Тэзісы друкуюцца ў выглядзе, пададзеным аўтарамі.

Scientific Editors:

DSc in Engineering B.M. Lobanov,
PhD in Engineering Yu.S. Hetsevich

ISBN 978-985-6744-89-4

FIRST ONE MILLION CORPORA FOR BELARUSIAN NOOJ MODULE

I. Reentovich[1], Yu. Hetsevich[1], V. Voronovich[2], E. Kachan[2], H. Kozlovskaya[2]
[1] United Institute of Informatics Problems of the NAS of Belarus, Minsk;
[2] Belarusian State University, Minsk
e-mail: mwshrewd@gmail.com

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

The text corpus that is used in NooJ will be effective for the research activity 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 applicable for solving many crucial tasks:

in general

- use this corpus in a linguistic development environment called NooJ [1] to optimize 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 grammars using 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.

CONTENTS