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

GRAMMARS FOR THE SENTENCE INTO PHRASE SEGMENTATION: PUNCTUATION LEVEL

Yu. Hetsevich, T. Okrut, B. Lobanov

United Institute of Informatics Problems of the NAS of Belarus, Minsk

e-mail: yury.hetsevich@gmail.com

This paper deals with, so called, punctuational phrases in the sentences and their intonation type marking in Belarusian electronic texts. Such marking allows to implement the algorithm of intonationally coloured (expressive) synthetic speech and to avoid its monotony.

There are 4 main categories of intonation types: finality (P), non-finality (C), interrogation (Q) and exclamation (E). The intonation types, constituting these categories, are designated by the category symbol (P, C, Q or E) and their subtypes marked the proper index (1-n).

The indicating model works as follows:

<PHRASE TYPE="C1">Амаль забылася</PHRASE> і <PHRASE TYPE="C7">здарэнне з лазняй</PHRASE>, а <PHRASE TYPE="C1"> потым </PHRASE> і <PHRASE TYPE="P4">лазню злізаў Дняпро</PHRASE>. <PHRASE TYPE="C7">Быццам і не было ні людзей</PHRASE>, ні <PHRASE TYPE="P4">закуранай нізкай будыніны на беразе</PHRASE>.

Each phrase in these sentences is tagged with an intonation index, namely: "C1" – non-final "і"-intonation with "і" (a Belarusian co-ordinating conjunction meaning "and") stating the end of the phrase, "C7" – non-final intonation of "comma staying before a co-ordinating conjunction" with a comma stating the end of phrase, "P4" – "full stop"– intonation with a full stop stating the end of phrase.

When marking intonation type of a phrase, not only punctuation marks are taken into account but also the nearest context in a text. For example, the intonation type "Q12" corresponds to each even phrase in a line of consecutive "Q1"– phrases – interrogative phrases containing an interrogative word:

<PHRASE TYPE="Q11">Як так атрымалася</PHRASE>? <PHRASE TYPE="Q12">Чаму ты мне нічога не сказаў</PHRASE>?

The algorithm developed represents the initial stage of prosodic processing in a speech synthesizer, the following step will be to provide segmentation of punctuational phrases into syntactic phrases.

CONTENTS