



Filozofická
fakulta
Faculty
of Philosophy

Jihočeská univerzita
v Českých Budějovicích
University of South Bohemia
in České Budějovice



Université de Franche-Comté, Besançon



The NOOJ Association

NOOJ 2016

International Conference

Book of Abstracts

České Budějovice (Czech Republic)

9-11 June 2016

Scientific Committee

Xavier Blanco (Autonomous University of Barcelona, Spain)

Yuras Hetsevich (United Institute of Informatics Problems, Belarus)

Svetla Koeva (University of Sofia, Bulgaria)

Peter Machonis (Florida International University, USA)

Slim Mesfar (University of Manouba, Tunisia)

Mario Monteleone (University of Salerno, Italy)

Johanna Monti (University of Sassari, Italy)

Karel Pala (Masaryk University, Czech republic)

Vladimír Petkevič (Charles University, Prague)

Jan Radimský (University of South Bohemia, Czech republic)

Max Silberztein (Université de Franche-Comté, France)

Marko Tadić (University of Zagreb, Croatia)

François Trouilleux (Université Blaise-Pascal, France)

Organization Committee

Jan Radimský (University of South Bohemia, Czech Republic)

Max Silberztein (Université de Franche-Comté, France)

Zuzana Nevěřilová (Masaryk University, Czech Republic)

Petr Kos (University of South Bohemia, Czech Republic)

Mario Monteleone (Università degli studi di Salerno, Italy)

NOOJ 2016 International Conference – Book of Abstracts

Edited by Jan Radimský

Published by the University of South Bohemia in České Budějovice (2016)

Table of Contents

A decision-support tool of <u>medicinal plants</u> using a NooJ platform	5
A pedagogical application of NooJ for teaching Spanish as a foreign language	6
✓Addition of IPA transcription to the belarusian NooJ module	8
Automatic recognition of enumerative series with NooJ. First results	9
Avancement des recherches linguistiques au Maroc dans le cadre de l'environnement Nooj.....	10
Clinical term recognition: from local to LOINC terminology. An application for Italian language.	12
Detection of Verb Frames with NooJ	14
Endpoint for Semantic Knowledge (ESK)	15
eSPERTo's Paraphrastic Knowledge applied to Question-Answering and Summarization	17
Formalising Natural Languages: The NooJ Approach.....	19
Generating alerts from automatically extracted tweets in Standard Arabic.....	20
Generation of sentences of the Japanese honorific language with NooJ.....	21
Inflectional and morphological variation of Arabic Multi-Word Expressions.....	23
Integration of a segmentation tool for Arabic corpora in NooJ Platform to build an automatic annotation tool	24
✓Knowledge-based system for the solution of the phoneme-to-grapheme problem of belarusian speech recognition using NOOJ	26
Latin Name of Plants: A morphological grammar based approach for Recognition and Extraction	28
Morphological treatment of the verbal forms in kabylian language using Nooj software.	29
Morpho-syntactic treatment of Arabic Adverbial degree expressions using NooJ.....	31
NooJ Local Grammars for Endophora Resolution.....	33
NooJ Local Grammars for Innovative Startup Language	35
Paraphrases for communication predicates.....	37
Phrasal Verb Disambiguating Grammars: Cutting out noise automatically	39
Quechua module for Nooj: Multilingual linguistic resources for MT	41
Recognizing Diminutive and Augmentative Croatian Nouns	43
Research for Chinese electronic dictionary	44

Knowledge-based system for the solution of the phoneme-to-grapheme problem of belarusian speech recognition using NOOJ

Lesia Kaigorodova

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

lesia.piatrouskaya@gmail.com

Anna Karpenko

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

rfe.karpenko@gmail.com

Yury Hetsevich

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

yury.hetsevich@gmail.com

Speech recognition for some groups of languages, i.e. Slavic, is still of poor performance compared to some other widely used languages, i.e. English, Spanish, French, etc. Some new methods for language modeling, phone-to-word task and phoneme recognition should be proposed for speech recognition of Belarusian, Ukrainian and even for Russian language. For Russian language this problem is being actively solved by some academia and speech technology companies, but for Belarusian language (for which this work is addressed) it is still a challenging task.

The classic design of the decoder that is the core of any speech recognition system is represented by a chain of finite state transducers that solve complex phoneme-to-grapheme problem (P2G) as well as language modeling task.

The aim of the proposed discussion is to describe how the transcribed data of speech corpus are collected and organized using NooJ [1] in order to solve phoneme-to-grapheme problem for speech recognition task for Belarusian language.

The overall problem of creation of transcribed data is split into two simpler problems, i.e. generation of lexicon for the speech corpus and generation of triphone database for the same corpus based on the obtained lexicon and corresponding text corpus.

Lexicon is the dictionary that is based on unique words of the text corpus that corresponds to the speech corpus and unique phonemic transcriptions of these words. Phonemic transcription of units for the lexicon is generated using the text corpus and Text-to-Speech Synthesizer [2]. NooJ Syntactic and Morphological grammars and some Python script are used to process text corpus and phonemic transcriptions in order to obtain lexicon.

Generation of triphone database is based on the transcription of words from obtained lexicon and 2-word and 3-word sequences from the text corpus. Here NooJ Syntactic grammars are also used to process the data in order to generate the database. This database is the source for creating weighted finite state transducer that is supposed to solve phoneme-to-grapheme problem.

Given that all the components, i.e. speech and text corpus of Belarusian language, lexicon and triphone database, are collected and organized in a specific format, it can be considered that the basis for the Nooj knowledge-based system for the P2G problem of Belarusian speech recognition task is created.

References

Silberztein, M. Nooj Manual / M. Silberztein [Electronic resource]. – 2014. – Mode of access : <http://www.nooj4nlp.net> – Date of access: 01.02.2016.

Text-to-Speech PHP-Based Synthesizer [Electronic resource]. – 2013. – Mode of access: <http://corpus.by/tts3>. – Date of access: 01.02.2016.