Slovenská akadémia vied
Jazykovedný ústav Ľudovíta Štúra

Natural Language Processing, Corpus Linguistics, E-learning

Seventh International Conference
Bratislava, Slovakia, 13–15 November 2013
Proceedings

Editors
Katarína Gajdošová
Adriána Žáková

RAM-Verlag
2013

The articles have been reviewed by members of the Program Committee.

Slovak National Corpus
Ľ. Štúr Institute of Linguistics
Slovak Academy of Sciences
Bratislava, Slovakia 2013
`http://korpus.juls.savba.sk/~slovko/`

Table of Contents

Foreword
Mária Šimková 7

Úvod
Mária Šimková 8

A New Path to a Modern Monolingual Dictionary of Contemporary Czech: the Structure of Data in the New Dictionary Writing System
Kamil Barbierik, Martina Holcová Habrová, Pavla Kochová, Tomáš Liška, Zdeňka Opavská, and Miroslav Virius 9

Data Deduplication in Slovak Corpora
Vladimír Benko 27

Software System for Processing Bulgarian Digital Resources: Parallel Corpora and Bilingual Dictionaries
Ralitsa Dutsova and Ludmila Dimitrova 40

Slovene Corpora for Corpus Linguistics and Language Technologies
Tomaž Erjavec 51

Obstacles and Solution to Recognizing Compound Nouns in Greek: A Corpus Study
Vasiliki Foufi, Kyriaki Ioannidou, and Olympia Tsaknaki 63

From Multilingual Dictionary to Lithuanian WordNet
Radovan Garabík and Indrė Pileckytė 74

Corpora of Private Correspondence as a Source of Material Focused on a Research of Diminutives
Zdeňka Hladká 81

Identification of Idioms in Spoken Corpora
Milena Hnátková and Marie Kopřivová 92

The Corpus CzeSL in the Service of Teaching Czech for Foreigners – Errors in the Use of the Pronoun *který*
Andrea Hudousková 100

Delimitation of Participles in the Manual Morphological Annotation
Agáta Karčová 108

Corpus Based Identification of Czech Light Verbs
Václava Kettnerová, Markéta Lopatková, Eduard Bejček, Anna Vernerová, and Marie Podobová 118

Agents Expressed by Prepositionless Instrumental Modifying Czech Nouns Derived from Intransitive Verbs
Veronika Kolářová 129

Corpus-based Online Word Formation Exercises for Advanced Learners of English – Challenges and Solutions
Grzegorz Krynicki 148

Experimenting with Slovak Wikipedia as a Source for Language Technologies
Michal Laclavík, Štefan Dlugolinský, and Michal Blanárik 160

Query Interface for Diverse Corpus Types
Tomáš Machálek and Michal Křen 166

The Effect of Stop Words elimination on Sequence Patterns Extraction in Comparable Corpora
Dáša Munková, Michal Munk, and Martin Vozár 174

Valency of Selected Primary Adjectives in the SYN2010 Corpus
Kateřina Najbrtová 183

Event Extractor: Email Events Detection and Calendar Integration
Filip Ogurčák and Michal Laclavík 197

Formal (Morpho)Syntax Properties of Reflexive Particles *se*, *si* as Free Morphemes in Contemporary Czech
Vladimír Petkevič 206

Introduction to Online Learning
Katarína Pišútová 217

Automatic Extraction of Multiword Units from Slovak Text Corpora
Ján Staš, Daniel Hládek, Jozef Juhár, and Martin Ološtiak 228

Verb Valency and Argument Non-correspondence in a Bilingual Treebank
Jana Šindlerová, Zdeňka Urešová, and Eva Fučíková 238

Determination of Czech BCT Prototypes on the Basis of Corpus Data
Tatiana Timoshchenko 248

Veni, Vidi, Vici: The Language Technology Infrastructure Landscape after CESAR
Tamás Váradi 261

Czech Subjectivity Lexicon: A Lexical Resource for Czech Polarity Classification
Kateřina Veselovská 279

A Corpus-based Analysis of the Functionality and the Meaning of Infinitive “Frustrative Construction” in Czech and Slovak
Uliana Yazhinova 285

Foreword

Slovko 2013 – Natural Language Processing, Corpus Linguistics, E-learning will be again held in Bratislava. The organizers – Slovak National Corpus Department of Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences are honoured to host participants from eight countries: Bulgaria, Czech Republic, Germany, Greece, Hungary, Poland, Slovakia and Slovenia.

Over three days participants will be able to benefit from 29 presentations, including 3 plenary talks. Unfortunately, one third of submitted papers on given topics has not been recommended by the Programme Committee members. We thank to all reviewers for their constructive suggestions and their help to make the conference even more successful.

The 7th edition of the biennial conference increased the presence of the linguistically-oriented (corpus-based and corpus-driven) studies. The more technically oriented papers provide information on effectiveness of the approaches applied, experimenting and innovative methods. Latest trends and tendencies in enhancing the corpus data can be found also in the papers written by Slovak authors.

We wish all participants of the conference Slovko 2013 profitable time and positive inspiration for further cooperation in the field of natural language processing, corpus linguistics and similar research.

Mária Šimková
Translated by Adriána Žáková