

Table of Contents

I NLP Applications

Parallel, or Comparable? That Is the Question	3
<i>Michaela Denisová</i>	
Automatic Identification of Speakers and Parties in Steno Protocols of the Czech Parliament	15
<i>Ota Mikušek</i>	
Keyness Analysis and Its Representation in Russian Academic Papers on Computational Linguistics: Evaluation of Algorithms	25
<i>Maria Khokhlova and Mikhail Koryshev</i>	
Information Extraction from Business Documents: A Case Study	35
<i>Martin Geletka, Mikuláš Bankovič, Dávid Meluš, Šárka Ščavnická, Michal Štefánik, and Petr Sojka</i>	
Keyword Extraction for Automatic Evaluation of Machine Translation	47
<i>Livia Kelebercová and František Forgáč</i>	
Using NVH as a Backbone Format in the Lexonomy Dictionary Editor ...	55
<i>Miloš Jakubíček, Vojtěch Kovář, Michal Měchura, and Adam Rambousek</i>	

II Evaluation Methods

Are Dictionary Definitions of Verbs in Corpora?	65
<i>Marie Stará</i>	
Evaluation of Various Approaches to Compute BLEU Metrics	71
<i>Lucia Benkova, and Lubomír Benko</i>	
Compressed FastText Models for Czech Tagger	79
<i>Zuzana Nevěřilová</i>	
Blooming Onion: Efficient Deduplication through Approximate Membership Testing	91
<i>Ondřej Herman</i>	
CompAn – A Tool for Quantitative Comparison of Corpus Annotation ...	97
<i>Vlasta Ohlídalo</i>	

III Text Corpora

Piötòst Ché Niènt, Mèi Piötòst - A Manually Revised Lombard-Italian Parallel Corpus	105
<i>Edoardo Signoroni</i>	
Aranea Go Middle East: Persicum	113
<i>Vladimír Benko</i>	
Pipeline Effectiveness in the Sketch Engine	123
<i>Matúš Kostka</i>	
Constructing Datasets from Dialogue Data	131
<i>Ondřej Sotolář, Jaromír Plhák, Michaela Lebedíková, Michał Tkaczyk, and David Šmahel</i>	
Semi-Manual Annotation of Topics and Genres in Web Corpora, The Cheap and Fast Way	141
<i>Vít Suchomel, Jan Kraus</i>	
Utok: The Fast Rule-based Tokenizer	149
<i>Pavel Rychlý and Samuel Špalek</i>	

IV Semantics and Language Modelling

When Tesseract Meets PERO: Open-Source Optical Character Recognition of Medieval Texts	157
<i>Vít Novotný and Aleš Horák</i>	
Medical Knowledge Resources for Text-Mining of Health Records in Czech, Polish, and Slovak	161
<i>Krištof Anetta</i>	
Secondary Prepositions with Causal Meaning in Russian	171
<i>Victor Zakharov, Kirill Boyarsky, Eugeny Kanevsky, and Anastasia Kozlova</i>	
Towards General Document Understanding through Question Answering	181
<i>Šárka Ščavnická, Michal Štefánik, Marek Kadlčík, Martin Geletka, and Petr Sojka</i>	
Manipulative Style Recognition of Czech News Texts using Stylometric Text Analysis	189
<i>Radoslav Sabol and Aleš Horák</i>	

The Influence of a Machine Translation System on Sentiment Levels	199
<i>Jaroslav Reichel, and Lubomír Benko</i>	
Subject Index	209
Author Index	211