

# Lecture Notes in Artificial Intelligence 8655

## Subseries of Lecture Notes in Computer Science

### LNAI Series Editors

Randy Goebel

*University of Alberta, Edmonton, Canada*

Yuzuru Tanaka

*Hokkaido University, Sapporo, Japan*

Wolfgang Wahlster

*DFKI and Saarland University, Saarbrücken, Germany*

### LNAI Founding Series Editor

Joerg Siekmann

*DFKI and Saarland University, Saarbrücken, Germany*

Petr Sojka Aleš Horák  
Ivan Kopeček Karel Pala (Eds.)

# Text, Speech, and Dialogue

17th International Conference, TSD 2014  
Brno, Czech Republic, September 8-12, 2014  
Proceedings

## Volume Editors

Petr Sojka  
Masaryk University  
Faculty of Informatics  
Department of Computer Graphics and Design  
Brno, Czech Republic  
sojka@fi.muni.cz

Aleš Horák  
Ivan Kopeček  
Karel Pala  
Masaryk University  
Faculty of Informatics  
Department of Information Technologies  
Brno, Czech Republic  
E-mail: {hales; kopecek; pala}@fi.muni.cz

ISSN 0302-9743  
ISBN 978-3-319-10815-5  
DOI 10.1007/978-3-319-10816-2

e-ISSN 1611-3349  
e-ISBN 978-3-319-10816-2

Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014946617

LNCS Sublibrary: SL 7 – Artificial Intelligence

© Springer International Publishing Switzerland 2014

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

*Typesetting:* Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

# Preface

The annual Text, Speech and Dialog Conference (TSD), which originated in 1998, is in the middle of its second decade. So far more than 1,000 authors from 45 countries have contributed to the proceedings. TSD constitutes a recognized platform for the presentation and discussion of state-of-the-art technology and recent achievements in the field of natural language processing. It has become an interdisciplinary forum, interweaving the themes of speech technology and language processing. The conference attracts researchers not only from Central and Eastern Europe but also from other parts of the world. Indeed, one of its goals has always been to bring together NLP researchers with different interests from different parts of the world and to promote their mutual cooperation.

One of the ambitions of the conference is, as its title says, not only to deal with dialog systems as such, but also to contribute to improving dialog between researchers in the two areas of NLP, i.e., between text and speech people. In our view, the TSD Conference was successful in this respect in 2014 again.

This volume contains the proceedings of the 17th TSD Conference, held in Brno, Czech Republic, in September 2014. In the review process, 70 papers were accepted out of 143 submitted, an acceptance rate of 49%.

We would like to thank all the authors for the efforts they put into their submissions and the members of the Program Committee and reviewers who did a wonderful job in helping us to select the most appropriate papers. We are also grateful to the invited speakers for their contributions. Their talks provided insight into important current issues, applications, and techniques related to the conference topics.

Special thanks are due to the members of the Local Organizing Committee for their tireless effort in organizing the conference.

The T<sub>E</sub>Xpertise of Petr Sojka resulted in the production of the volume that you are holding in your hands.

We hope that the readers will benefit from the results of this event and disseminate the ideas of the TSD Conference all over the world. Enjoy the proceedings!

July 2014

Aleš Horák  
Ivan Kopeček  
Karel Pala  
Petr Sojka

# Organization

TSD 2014 was organized by the Faculty of Informatics, Masaryk University, in cooperation with the Faculty of Applied Sciences, University of West Bohemia in Plzeň. The conference webpage is located at <http://www.tsdconference.org>

## Program Committee

Nöth, Elmar, Germany, *General Chair*  
Agirre, Eneko, Spain  
Baudoin, Geneviève, France  
Cook, Paul, Australia  
Černocký, Jan, Czech Republic  
Dobrišek, Simon, Slovenia  
Evgrafova, Karina, Russia  
Fiser, Darja, Slovenia  
Garabík, Radovan, Slovakia  
Gelbukh, Alexander, Mexico  
Guthrie, Louise, UK  
Hajič, Jan, Czech Republic  
Hajičová, Eva, Czech Republic  
Haralambous, Yannis, France  
Hermansky, Hynek, USA  
Hitzenberger, Ludwig, Germany  
Hlaváčová, Jaroslava, Czech Republic  
Horák, Aleš, Czech Republic  
Hovy, Eduard, USA  
Khokhlova, Maria, Russia  
Kocharov, Daniil, Russia  
Kopeček, Ivan, Czech Republic  
Kordoni, Valia, Germany  
Krauwier, Steven, The Netherlands  
Kunzmann, Siegfried, Germany  
Loukachevitch, Natalija, Russia  
Matoušek, Václav, Czech Republic  
McCarthy, Diana, UK

Mihelić, France, Slovenia  
Ney, Hermann, Germany  
Oliva, Karel, Czech Republic  
Pala, Karel, Czech Republic  
Pavesić, Nikola, Slovenia  
Pianesi, Fabio, Italy  
Piasecki, Maciej, Poland  
Przepiorkowski, Adam, Poland  
Psutka, Josef, Czech Republic  
Pustejovsky, James, USA  
Rigau, German, Spain  
Rothkrantz, Leon, The Netherlands  
Rumshinsky, Anna, USA  
Rusko, Milan, Slovakia  
Sazhok, Mykola, Ukraine  
Skrelin, Pavel, Russia  
Smrž, Pavel, Czech Republic  
Sojka, Petr, Czech Republic  
Steidl, Stefan, Germany  
Stemmer, Georg, Germany  
Tadić, Marko, Croatia  
Varadi, Tamas, Hungary  
Vetulani, Zygmunt, Poland  
Wiggers, Pascal, The Netherlands  
Wilks, Yorick, UK  
Wolinski, Marcin, Poland  
Zakharov, Victor, Russia

## Additional Referees

Agerr, Rodrigo  
Fedorov, Yevgen  
Gonzalez-Agirre, Aitor  
Grzl, František  
Hana, Jirka  
Hajdinjak, Melita  
Hlaváčková, Dana

Holub, Martin  
Jakubíček, Miloš  
Otegi, Arantxa  
Veselý, Karel  
Veselovská, Kateřina  
Wang, Xinglong  
Waver, Aleksander

## Organizing Committee

Aleš Horák (*Co-chair*), Ivan Kopeček, Karel Pala (*Co-chair*), Adam Rambousek (*Web System*), Pavel Rychlý, Petr Sojka (*Proceedings*)

## Sponsors and Support

The TSD conference is regularly supported by the International Speech Communication Association (ISCA). We would like to express our thanks to the Lexical Computing Ltd. and IBM Česká republika, spol. s r. o. for their kind sponsoring contribution to TSD 2014.

# Table of Contents

## Invited Papers

An Information Extraction Customizer . . . . .	3
<i>Ralph Grishman and Yifan He</i>	
Entailment Graphs for Text Analytics in the Excitement Project . . . . .	11
<i>Bernardo Magnini, Ido Dagan, Günter Neumann, and Sebastian Pado</i>	
Multi-lingual Text Leveling . . . . .	19
<i>Salim Roukos, Jerome Quin, and Todd Ward</i>	

## Text

SuMACC Project's Corpus: A Topic-Based Query Extension Approach to Retrieve Multimedia Documents . . . . .	29
<i>Mohamed Morchid, Richard Dufour, Usman Niaz, Francis Bouvier, Clément de Groc, Claude de Loupy, Georges Linarès, Bernard Merialdo, and Bertrand Peralta</i>	
Empiric Introduction to Light Stochastic Binarization . . . . .	37
<i>Daniel Devatman Hromada</i>	
Comparative Study Concerning the Role of Surface Morphological Features in the Induction of Part-of-Speech Categories . . . . .	46
<i>Daniel Devatman Hromada</i>	
Automatic Adaptation of Author's Stylometric Features to Document Types . . .	53
<i>Jan Rygl</i>	
Detecting Commas in Slovak Legal Texts . . . . .	62
<i>Róbert Sabo and Štefan Beňuš</i>	
Detection and Classification of Events in Hungarian Natural Language Texts . . . . .	68
<i>Zoltán Subecz</i>	
Generating Underspecified Descriptions of Landmark Objects . . . . .	76
<i>Ivandré Paraboni, Alan K. Yamasaki, Adriano S.R. da Silva, and Caio V.M. Teixeira</i>	
A Topic Model Scoring Approach for Personalized QA Systems . . . . .	84
<i>Hamidreza Chinaei, Luc Lamontagne, François Lavolette, and Richard Khoury</i>	

Feature Exploration for Authorship Attribution of Lithuanian Parliamentary Speeches . . . . .	93
<i>Jurgita Kapočiūtė-Dzikienė, Andrius Utkā, and Ligita Šarkutė</i>	
Processing of Quantitative Expressions with Measurement Units in the Nominative, Genitive, and Accusative Cases for Belarusian and Russian . . . . .	101
<i>Yury Hetsevich and Alena Skopinava</i>	
Document Classification with Deep Rectifier Neural Networks and Probabilistic Sampling . . . . .	108
<i>Tamás Grósz and István Nagy T.</i>	
Language Independent Evaluation of Translation Style and Consistency: Comparing Human and Machine Translations of Camus' Novel "The Stranger" . . . . .	116
<i>Mahmoud El-Haj, Paul Rayson, and David Hall</i>	
Bengali Named Entity Recognition Using Margin Infused Relaxed Algorithm . . . . .	125
<i>Somnath Banerjee, Sudip Kumar Naskar, and Sivaji Bandyopadhyay</i>	
Score Normalization Methods Applied to Topic Identification . . . . .	133
<i>Lucie Skorkovská and Zbyněk Zajíc</i>	
Disambiguation of Japanese Onomatopoeias Using Nouns and Verbs . . . . .	141
<i>Hironori Fukushima, Kenji Araki, and Yuzu Uchida</i>	
Continuous Distributed Representations of Words as Input of LSTM Network Language Model . . . . .	150
<i>Daniel Soutner and Luděk Müller</i>	
NERC-fr: Supervised Named Entity Recognition for French . . . . .	158
<i>Andoni Azpeitia, Montse Cuadros, Seán Gaines, and German Rigau</i>	
Semantic Classes and Relevant Domains on WSD . . . . .	166
<i>Rubén Izquierdo, Sonia Vázquez, and Andrés Montoyo</i>	
An MLU Estimation Method for Hungarian Transcripts . . . . .	173
<i>György Orosz and Kinga Mátyus</i>	
Using Verb-Noun Patterns to Detect Process Inputs . . . . .	181
<i>Munshi Asadullah, Damien Nouvel, and Patrick Paroubek</i>	
Divergences in the Usage of Discourse Markers in English and Mandarin Chinese . . . . .	189
<i>David Steele and Lucia Specia</i>	



Sentence Similarity by Combining Explicit Semantic Analysis and Overlapping N-Grams . . . . .	201
<i>Hai Hieu Vu, Jeanne Villaneau, Farida Saïd, and Pierre-François Marteau</i>	
Incorporating Language Patterns and Domain Knowledge into Feature-Opinion Extraction . . . . .	209
<i>Erqiang Zhou, Xi Luo, and Zhiguang Qin</i>	
BFQA: A Bengali Factoid Question Answering System . . . . .	217
<i>Somnath Banerjee, Sudip Kumar Naskar, and Sivaji Bandyopadhyay</i>	
Dictionary-Based Problem Phrase Extraction from User Reviews . . . . .	225
<i>Valery Solovyev and Vladimir Ivanov</i>	
RelANE: Discovering Relations between Arabic Named Entities . . . . .	233
<i>Ines Boujelben, Salma Jamoussi, and Abdelmajid Ben Hamadou</i>	
Building an Arabic Linguistic Resource from a Treebank: The Case of Property Grammar . . . . .	240
<i>Raja Bensalem Bahloul, Marwa Elkarwi, Kais Haddar, and Philippe Blache</i>	
Aranea: Yet Another Family of (Comparable) Web Corpora . . . . .	247
<i>Vladimír Benko</i>	
Towards a Unified Exploitation of Electronic Dialectal Corpora: Problems and Perspectives . . . . .	257
<i>Nikitas N. Karanikolas, Eleni Galiotou, and Angela Ralli</i>	
Named Entity Recognition for Highly Inflectional Languages: Effects of Various Lemmatization and Stemming Approaches . . . . .	267
<i>Michal Konkol and Miloslav Konopík</i>	
An Experiment with Theme–Rheme Identification . . . . .	275
<i>Karel Pala and Ondřej Svoboda</i>	
Self Training Wrapper Induction with Linked Data . . . . .	285
<i>Anna Lisa Gentile, Ziqi Zhang, and Fabio Ciravegna</i>	
Paraphrase and Textual Entailment Generation . . . . .	293
<i>Zuzana Nevěřilová</i>	
Clustering in a News Corpus . . . . .	301
<i>Richard Elling Moe</i>	
Partial Grammar Checking for Czech Using the SET Parser . . . . .	308
<i>Vojtěch Kovář</i>	

Russian Learner Translator Corpus: Design, Research Potential and Applications . . . . .	315
<i>Andrey Kutuzov and Maria Kunilovskaya</i>	
Development of a Semantic and Syntactic Model of Natural Language by Means of Non-negative Matrix and Tensor Factorization . . . . .	324
<i>Anatoly Anisimov, Oleksandr Marchenko, Volodymyr Taranukha, and Taras Vozniuk</i>	
Partial Measure of Semantic Relatedness Based on the Local Feature Selection . . . . .	336
<i>Maciej Piasecki and Michał Wendelberger</i>	
A Method for Parallel Non-negative Sparse Large Matrix Factorization . . . . .	344
<i>Anatoly Anisimov, Oleksandr Marchenko, Emil Nasirov, and Stepan Palamarchuk</i>	
Using Graph Transformation Algorithms to Generate Natural Language Equivalents of Icons Expressing Medical Concepts . . . . .	353
<i>Pascal Vaillant and Jean-Baptiste Lamy</i>	

**Speech**

GMM Classification of Text-to-Speech Synthesis: Identification of Original Speaker's Voice . . . . .	365
<i>Jiří Přibíl, Anna Přibílová, and Jindřich Matoušek</i>	
Phonation and Articulation Analysis of Spanish Vowels for Automatic Detection of Parkinson's Disease . . . . .	374
<i>Juan Rafael Orozco-Arroyave, Elkyn Alexander Belalcázar-Bolaños, Julián David Arias-Londoño, Jesús Francisco Vargas-Bonilla, Tino Haderlein, and Elmar Nöth</i>	
Speaker Identification by Combining Various Vocal Tract and Vocal Source Features . . . . .	382
<i>Yuta Kawakami, Longbiao Wang, Atsuhiko Kai, and Seiichi Nakagawa</i>	
Inter-Annotator Agreement on Spontaneous Czech Language: Limits of Automatic Speech Recognition Accuracy . . . . .	390
<i>Tomáš Valenta, Luboš Šmídl, Jan Švec, and Daniel Soutner</i>	
Minimum Text Corpus Selection for Limited Domain Speech Synthesis . . . . .	398
<i>Markéta Jůzová and Daniel Tihelka</i>	

Tuning Limited Domain Speech Synthesis Using General Text-to-Speech System . . . . .	408
<i>Markéta Jůzová and Daniel Tihelka</i>	
Study on Phrases Used for Semi-automatic Text-Based Speakers' Names Extraction in the Czech Radio Broadcasts News . . . . .	416
<i>Michaela Kuchařová, Svatava Škodová, Ladislav Šeps, and Marek Boháč</i>	
Development of a Large Spontaneous Speech Database of Agglutinative Hungarian Language . . . . .	424
<i>Tilda Neuberger, Dorottya Gyarmathy, Tekla Etelka Gráczi, Viktória Horváth, Mária Gósy, and András Beke</i>	
Unit Selection Cost Function Exploration Using an A* Based Text-to-Speech System . . . . .	432
<i>David Guennec and Damien Lolive</i>	
LIUM and CRIM ASR System Combination for the REPERE Evaluation Campaign . . . . .	441
<i>Anthony Rousseau, Gilles Boulianne, Paul Deléglise, Yannick Estève, Vishwa Gupta, and Sylvain Meignier</i>	
Anti-Models: An Alternative Way to Discriminative Training . . . . .	449
<i>Jan Vaněk and Josef Psutka</i>	
Modelling F <sub>0</sub> Dynamics in Unit Selection Based Speech Synthesis . . . . .	457
<i>Daniel Tihelka, Jindřich Matoušek, and Zdeněk Hanzlíček</i>	
Audio-Video Speaker Diarization for Unsupervised Speaker and Face Model Creation . . . . .	465
<i>Pavel Campr, Marie Kunešová, Jan Vaněk, Jan Čech, and Josef Psutka</i>	
Improving a Long Audio Aligner through Phone-Relatedness Matrices for English, Spanish and Basque . . . . .	473
<i>Aitor Álvarez, Pablo Ruiz, and Haritz Arzelus</i>	
Initial Experiments on Automatic Correction of Prosodic Annotation of Large Speech Corpora . . . . .	481
<i>Zdeněk Hanzlíček and Martin Grüber</i>	
Automatic Speech Recognition Texts Clustering . . . . .	489
<i>Svetlana Popova, Ivan Khodyrev, Irina Ponomareva, and Tatiana Krivosheeva</i>	
Impact of Irregular Pronunciation on Phonetic Segmentation of Nijmegen Corpus of Casual Czech . . . . .	499
<i>Petr Mizera, Petr Pollak, Alice Kolman, and Mirjam Ernestus</i>	

Parametric Speech Coding Framework for Voice Conversion Based on Mixed Excitation Model . . . . .	507
<i>Michał Lenarczyk</i>	
Captioning of Live TV Commentaries from the Olympic Games in Sochi: Some Interesting Insights . . . . .	515
<i>Josef V. Psutka, Aleš Pražák, Josef Psutka, and Vlasta Radová</i>	
Language Resources and Evaluation for the Support of the Greek Language in the MARY Text-to-Speech . . . . .	523
<i>Pepi Stavropoulou, Dimitrios Tsonos, and Georgios Kouroupetroglou</i>	
Intelligibility Assessment of the De-Identified Speech Obtained Using Phoneme Recognition and Speech Synthesis Systems . . . . .	529
<i>Tadej Justin, France Mihelič, and Simon Dobrišek</i>	
<b>Dialogue</b>	
Referring Expression Generation: Taking Speakers' Preferences into Account . . . . .	539
<i>Thiago Castro Ferreira and Ivandré Paraboni</i>	
Visualization of Intelligibility Measured by Language-Independent Features . . .	547
<i>Tino Haderlein, Catherine Middag, Andreas Maier, Jean-Pierre Martens, Michael Döllinger, and Elmar Nöth</i>	
Using Suprasegmental Information in Recognized Speech Punctuation Completion . . . . .	555
<i>Marek Boháč and Karel Blavka</i>	
Two-Layer Semantic Entity Detection and Utterance Validation for Spoken Dialogue Systems . . . . .	563
<i>Adam Chýlek, Jan Švec, and Luboš Šmídl</i>	
Ontology Based Strategies for Supporting Communication within Social Networks . . . . .	571
<i>Ivan Kopeček, Radek Ošlejšek, and Jaromír Plhák</i>	
A Factored Discriminative Spoken Language Understanding for Spoken Dialogue Systems . . . . .	579
<i>Filip Jurčiček, Ondřej Dušek, and Ondřej Plátek</i>	
Alex: A Statistical Dialogue Systems Framework . . . . .	587
<i>Filip Jurčiček, Ondřej Dušek, Ondřej Plátek, and Lukáš Žilka</i>	

Speech Synthesis and Uncanny Valley . . . . .	595
<i>Jan Romportl</i>	
Integration of an On-line Kaldi Speech Recogniser to the Alex Dialogue Systems Framework . . . . .	603
<i>Ondřej Plátek and Filip Jurčíček</i>	
<b>Author Index</b> . . . . .	611