

# **19th Nordic Conference of Computational Linguistics (NODALIDA 2013)**

NEALT Proceedings Series Volume 16

Oslo, Norway  
22 – 24 May 2013

## **Editors:**

**Stephan Oepen  
Kristin Hagen  
Janne Bondi Johannessen**

ISBN: 978-1-5108-3335-7

**Printed from e-media with permission by:**

Curran Associates, Inc.  
57 Morehouse Lane  
Red Hook, NY 12571



**Some format issues inherent in the e-media version may also appear in this print version.**

Copyright© (2013) by the Association for Computational Linguistics  
All rights reserved.

Printed by Curran Associates, Inc. (2017)

For permission requests, please contact the Association for Computational Linguistics  
at the address below.

Association for Computational Linguistics  
209 N. Eighth Street  
Stroudsburg, Pennsylvania 18360

Phone: 1-570-476-8006  
Fax: 1-570-476-0860

[acl@aclweb.org](mailto:acl@aclweb.org)

**Additional copies of this publication are available from:**

Curran Associates, Inc.  
57 Morehouse Lane  
Red Hook, NY 12571 USA  
Phone: 845-758-0400  
Fax: 845-758-2633  
Email: [curran@proceedings.com](mailto:curran@proceedings.com)  
Web: [www.proceedings.com](http://www.proceedings.com)

## Table of Contents

### *Invited Keynotes*

<b>Ron Kaplan</b>	
The Conversational User Interface	1
<b>Caroline Sporleder</b>	
Detecting and Processing Figurative Language in Discourse	3
<b>Anders Søgaard</b>	
6,909 Reasons to Mess Up Your Data	5

### *Special Session on HPC for NLP*

<b>Gudmund Høst</b>	
The Nordic e-Infrastructure Collaboration: Opportunities for Synergy Across Borders	7
<b>Stephan Oepen</b>	
Tidying up the Basement: A Tale of Large-Scale Parsing on National eInfrastructure	9
<b>Jörg Tiedemann</b>	
Experiences in Building the Let's MT! Portal on Amazon EC2	11

### *Regular Papers*

<b>Eckhard Bick</b>	
Using Constraint Grammar for Chunking	13
<b>Johan Falkenjack, Katarina Heimann Mühlenbock, Arne Jönsson</b>	
Features Indicating Readability in Swedish Text	27
<b>Katri Haverinen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Jenna Nyblom, Stina Ojala, Timo Viljanen, Tapio Salakoski, Filip Ginter</b>	
Towards a Dependency-Based PropBank of General Finnish	41
<b>Ryan Johnson, Lene Antonsen, Trond Trosterud</b>	
Using Finite State Transducers for Making Efficient Reading Comprehension Dictionaries	59
<b>Jurgita Kapočiūtė-Dzikiėnė, Anders Nøklestad, Janne Bondi Johannessen, Algis Krupavičius</b>	
Exploring Features for Named Entity Recognition in Lithuanian Text Corpus	73
<b>Hrafn Loftsson</b>	
Tagging the Past: Experiments using the Saga Corpus	89
<b>Hrafn Loftsson, Robert Östling</b>	
Tagging a Morphologically Complex Language Using an Averaged Perceptron Tagger: The Case of Icelandic	105
<b>Magnus Merkel, Jody Foo, Lars Ahrenberg</b>	
Iphractor: A Linguistically Informed System for Extraction of Term Candidates	121

<b>Costanza Navarretta, Patrizia Paggio</b> Classifying Multimodal Turn Management in Danish Dyadic First Encounters	133
<b>Bolette S. Pedersen, Lars Borin, Markus Forsberg, Neeme Kahusk, Krister Lindén, Jyrki Niemi, Niklas Nisbeth, Lars Nygaard, Heili Orav, Eirikur Rögnvaldsson, Mitchell Seaton, Kadri Vider, Kaarlo Voionmaa</b> Nordic and Baltic Wordnets Aligned and Compared through “WordTies”	147
<b>Eva Pettersson, Beáta Megyesi, Joakim Nivre</b> Normalisation of Historical Text Using Context-Sensitive Weighted Levenshtein Distance and Compound Splitting	163
<b>Teemu Ruokolainen, Miikka Silfverberg</b> Modeling OOV Words With Letter N-Grams in Statistical Taggers: Preliminary Work in Biomedical Entity Recognition	181
<b>Inguna Skadiņa, Andrejs Vasiļjevs, Lars Borin, Krister Lindén, Gyri Losnegaard, Sussi Olsen, Bolette S. Pedersen, Roberts Rozis, Koenraad De Smedt</b> Baltic and Nordic Parts of the European Linguistic Infrastructure	195

<i>Student Papers</i>
-----------------------

<b>Liesbeth Augustinus, Peter Dirix</b> The IPP Effect in Afrikaans: A Corpus Analysis	213
<b>Christopher Horn, Alisa Zhila, Alexander Gelbukh, Roman Kern, Elisabeth Lex</b> Using Factual Density to Measure Informativeness of Web Documents	227
<b>Tapio Luostarinen, Oskar Kohonen</b> Using Topic Models in Content-Based News Recommender Systems	239
<b>Bernd Opitz, Cäcilia Zirn</b> Bootstrapping an Unsupervised Approach for Classifying Agreement and Disagreement	253
<b>Pēteris Paikens, Laura Rituma, Lauma Pretkalniņa</b> Morphological Analysis with Limited Resources: Latvian Example	267
<b>Lauma Pretkalniņa, Laura Rituma</b> Statistical Syntactic Parsing for Latvian	279

<i>Short Papers</i>
---------------------

<b>Filip Ginter, Jenna Nyblom, Veronika Laippala, Samuel Kohonen, Katri Haverinen, Simo Vihjanen, Tapio Salakoski</b> Building a Large Automatically Parsed Corpus of Finnish	291
<b>Lars Hellan, Tore Bruland</b> Constructing a Multilingual Database of Verb Valence	301
<b>Jussi Karlgren</b> New Measures to Investigate Term Typology by Distributional Data	311
<b>Andreas Sjøeborg Kirkedal</b> Analysis of Phonetic Transcription for Danish Automatic Speech Recognition	321

<b>Samuel Läubli, Mark Fishel, Martin Volk, Manuela Weibel</b> Combining Statistical Machine Translation and Translation Memories with Domain Adaptation	331
<b>Sjur N. Moshagen, Tommi A. Pirinen, Trond Trosterud</b> Building an Open-Source Development Infrastructure for Language Technology Projects	343
<b>Gailius Raškinis, Asta Kazlauskienė</b> From Speech Corpus to Intonation Corpus: Clustering Phrase Pitch Contours of Lithuanian	353
<b>Jonathon Read, Rebecca Dridan, Stephan Oepen</b> Simple and Accountable Segmentation of Marked-up Text	365
<b>Sara Stymne, Jörg Tiedemann, Christian Hardmeier, Joakim Nivre</b> Statistical Machine Translation with Readability Constraints	375
<b>Hideyuki Tanushi, Hercules Dalianis, Martin Duneld, Maria Kvist, Maria Skeppstedt, Sumithra Velupillai</b> Negation Scope Delimitation in Clinical Text Using Three Approaches: NegEx, PyConTextNLP and SynNeg	387
<b>Marcus Uneson</b> Tone Restoration in Transcribed Kammu: Decision-List Word Sense Disambiguation for an Unwritten Language	399
<b>Nynke Van Der Vliet, Gosse Bouma, Gisela Redeker</b> The Automatic Identification of Discourse Units in Dutch Text	411

<i>Demonstration Papers</i>
-----------------------------

<b>Liesbeth Augustinus, Vincent Vandeghinste, Ineke Schuurman, Frank Van Eynde</b> Example-Based Treebank Querying with GrE TEL - Now Also for Spoken Dutch	423
<b>Malin Ahlberg, Lars Borin, Markus Forsberg, Martin Hammarstedt, Leif-Jöran Olsson, Olof Olsson, Johan Roxendal, Jonatan Uppström</b> Korp and Karp – A Bestiary of Language Resources: The Research Infrastructure of Språkbanken	429
<b>Lars Hellan, Tore Bruland, Elias Aamot, Mads H. Sandøy</b> A Grammar Sparrer for Norwegian	435
<b>Mans Hulden, Miikka Silfverberg, Jerid Francom</b> Finite State Applications with Javascript	441
<b>Emanuele Lapponi, Erik Velldal, Nikolay A. Vazov, Stephan Oepen</b> HPC-ready Language Analysis for Human Beings	447
<b>Paul Meurer, Helge Dyvik, Victoria Rosén, Koenraad De Smedt, Gunn Inger Lyse, Gyri Smørdal Losnegaard, Martha Thunes</b> The INESS Treebanking Infrastructure	453
<b>Per Erik Solberg</b> Building Gold-Standard Treebanks for Norwegian	459