

NEALT

Proceedings

Northern European Association for Language Technology



Editors

Stephan Oepen

Kristin Hagen

Janne Bondi Johannessen

Proceedings of the
19th Nordic Conference of Computational Linguistics

NODALIDA 2013

May 22-24, 2013 • Oslo, Norway

Linköping Electronic Conference Proceedings

Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)

Stephan Oepen, Kristin Hagen, and
Janne Bondi Johannessen (Editors)

May 22–24, 2013

Oslo University (Norway)

Published by

Linköping University Electronic Press, Sweden

Linköping Electronic Conference Proceedings #85

(ISSN 1650-3740; ISBN 978-91-7519-589-6)

Preface

The tradition of bi-annual Nordic conferences in Computational Linguistics and related disciplines dates back to 1977, well before our professional organization—The Northern European Association for Language Technology (NEALT; <http://omilia.uio.no/nealt/>)—was formally established. With a sense of tradition as well as pride, this volume comprises the proceedings of the 19th Nordic Conference on Computational Linguistics (NODALIDA 2013), held on the campus of the University of Oslo, Norway, between May 22 and May 24, 2013. On the first day of NODALIDA 2013, four topical workshops are held, each with its own set of organizers and programme committee; these workshops have compiled their own proceedings volumes, which are published in the same series and included on the media distributed at the conference.

NODALIDA addresses all aspects of speech recognition and synthesis, natural language processing, and computational linguistics—including work in closely related neighbouring disciplines (such as, for example, linguistics or psychology) that is sufficiently formalized or applied to bear relevance to speech and language technologies. Following the pattern of previous years, the Programme Committee invited paper submissions in four distinct tracks:

- *regular papers* on substantial, original, and unpublished research, including empirical evaluation results, where appropriate;
- *student papers* on completed or ongoing work, where at least the first author is a Master- or PhD-level student;
- *short papers* on smaller, focused contributions, work in progress, negative results, surveys, or opinion pieces; and
- *demonstration papers* summarizing a software system or language resource, to be accompanied by a live demonstration at the conference.

The conference received 60 submissions from all over Europe (and one each from Mexico and the US), of which 38 are collected in this volume and will be presented at the conference: 13 regular, 6 student, 12 short, and 7 demonstration papers. All submissions were reviewed by at least three experts in the field (two for demonstration papers), and the final selection was made by the Programme Committee. We are indebted to everyone who contributed to the reviewing and selection process. The conference programme is complemented by three invited keynotes by distinguished researchers from Denmark, Germany, and the US, as well as by a special session on High-Performance Computing for Natural Language Processing.

NODALIDA 2013 is made possible by the joint work of many dedicated individuals, in particular the Programme and Organizing Committees; we warmly acknowledge their enthusiasm and community spirit. From the Organizing Committee, Kristin Hagen deserves a special note of gratitude, as the untiring ‘heart and soul’ of the conference logistics. We are grateful to the Department of Linguistics and Scandinavian Studies and the Department of Informatics at the University of Oslo for generously making available infrastructure and staff time. The conference is financially supported by organizations listed on the back cover, who thus make an important contribution to keeping participation fees at quite reasonable levels (by Norwegian standards).

With just about two more weeks to go, we expect some 150 participants at the conference and much look forward to welcoming our colleagues and peers to Oslo.

Stephan Oepen (Programme Chair), Janne Bondi Johannessen (Organizing Chair)

Programme Committee

- Lars Ahrenberg, Linköping University, Sweden
- Heiki-Jaan Kaalep, University of Tartu, Estonia
- Mikko Kurimo, Aalto University, Finland
- Stephan Oepen (Programme Chair), University of Oslo, Norway
- Eva Pettersson, Uppsala University, Sweden
- Bolette Sandford Pedersen, University of Copenhagen, Denmark
- Victoria Rosén, University of Bergen, Norway

Organizing Committee

- Ruth Vatvedt Fjeld
- Kristin Hagen
- Janne Bondi Johannessen (Organizing Chair)
- Anders Nøklestad
- Erik Velldal
- Lilja Øvrelid

Reviewers

- Szymon Acedański, Poland
- Tanel Alumäe, Estonia
- Miguel Ballesteros, Spain
- Emily M. Bender, USA
- Gosse Bouma, Netherlands
- Johan Boye, Sweden
- Aoife Cahill, USA
- Stefanie Dipper, Germany
- Helge Dyvik, Norway
- Jakob Elming, Denmark
- Tomaž Erjavec, Slovenia
- Peter Exner, Sweden

- Mark Fishel, Estonia
- Victoria Fossum, USA
- Jennifer Foster, Irland
- Tatiana Gornostay, Latvia
- Gintarė Grigonytė, Switzerland
- Christian Hardmeier, Sweden
- Petter Haugereid, Norway
- Anna Hjalmarsson, Sweden
- Sofie Johansson Kokkinakis, Sweden
- Arne Jönsson, Sweden
- Reima Karhila, Finland
- Mare Koit, Estonia
- Jan Tore Lønning, Norway
- Bente Maegaard, Denmark
- Diana McCarthy, United Kingdom
- Beata Megyesi, Sweden
- Einar Meister, Estonia
- Magnus Merkel, Sweden
- Paul Meurer Norway
- Costanza Navarretta, Denmark
- Mattias Nilsson, Sweden
- Joakim Nivre, Sweden
- Pierre Nugues Lund, Sweden
- Petya Osenova, Bulgaria
- Patrizia Paggio, Denmark
- Barbara Plank, Italy
- Adam Przepiórkowski, Poland
- Jonathon Read, Norway
- Trial Reviewer, Norway

- Inguna Skadiņa, Latvia
- Anders Søgaard, Denmark
- Sara Stymne, Sweden
- Jörg Tiedemann, Sweden
- Andrius Utkla, Lithuania
- Erik Velldal, Norway
- Sumithra Velupillai, Sweden
- Yi Zhang, Germany
- Heike Zinsmeister, Germany

Table of Contents

Invited Keynotes

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

Special Session on HPC for NLP

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

Regular Papers

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

Costanza Navarretta, Patrizia Paggio Classifying Multimodal Turn Management in Danish Dyadic First Encounters	133
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 Nordic and Baltic Wordnets Aligned and Compared through “WordTies”	147
Eva Pettersson, Beáta Megyesi, Joakim Nivre Normalisation of Historical Text Using Context-Sensitive Weighted Levenshtein Distance and Compound Splitting	163
Teemu Ruokolainen, Miikka Silfverberg Modeling OOV Words With Letter N-Grams in Statistical Taggers: Preliminary Work in Biomedical Entity Recognition	181
Inguna Skadiņa, Andrejs Vasiljevs, Lars Borin, Krister Lindén, Gyri Losnegaard, Sussi Olsen, Bolette S. Pedersen, Roberts Rozis, Koenraad De Smedt Baltic and Nordic Parts of the European Linguistic Infrastructure	195

Student Papers

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

Short Papers

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

Samuel Läubli, Mark Fishel, Martin Volk, Manuela Weibel Combining Statistical Machine Translation and Translation Memories with Domain Adaptation	331
Sjur N. Moshagen, Tommi A. Pirinen, Trond Trosterud Building an Open-Source Development Infrastructure for Language Technology Projects	343
Gailius Raškiniš, Asta Kazlauskienė From Speech Corpus to Intonation Corpus: Clustering Phrase Pitch Contours of Lithuanian	353
Jonathon Read, Rebecca Dridan, Stephan Oepen Simple and Accountable Segmentation of Marked-up Text	365
Sara Stymne, Jörg Tiedemann, Christian Hardmeier, Joakim Nivre Statistical Machine Translation with Readability Constraints	375
Hideyuki Tanushi, Hercules Dalianis, Martin Duneld, Maria Kvist, Maria Skeppstedt, Sumithra Velupillai Negation Scope Delimitation in Clinical Text Using Three Approaches: NegEx, PyConTextNLP and SynNeg	387
Marcus Uneson Tone Restoration in Transcribed Kammu: Decision-List Word Sense Disambiguation for an Unwritten Language	399
Nynke Van Der Vliet, Gosse Bouma, Gisela Redeker The Automatic Identification of Discourse Units in Dutch Text	411
<i>Demonstration Papers</i>	
Liesbeth Augustinus, Vincent Vandeghinste, Ineke Schuurman, Frank Van Eynde Example-Based Treebank Querying with GrETEL - Now Also for Spoken Dutch	423
Malin Ahlberg, Lars Borin, Markus Forsberg, Martin Hammarstedt, Leif-Jöran Olsson, Olof Olsson, Johan Roxendal, Jonatan Uppström Korp and Karp – A Bestiary of Language Resources: The Research Infrastructure of Språkbanken	429
Lars Hellan, Tore Bruland, Elias Aamot, Mads H. Sandøy A Grammar Sparrer for Norwegian	435
Mans Hulden, Miikka Silfverberg, Jerid Francom Finite State Applications with Javascript	441
Emanuele Lapponi, Erik Velldal, Nikolay A. Vazov, Stephan Oepen HPC-ready Language Analysis for Human Beings	447
Paul Meurer, Helge Dyvik, Victoria Rosén, Koenraad De Smedt, Gunn Inger Lyse, Gyri Smørdal Losnegaard, Martha Thunes The INESS Treebanking Infrastructure	453
Per Erik Solberg Building Gold-Standard Treebanks for Norwegian	459

