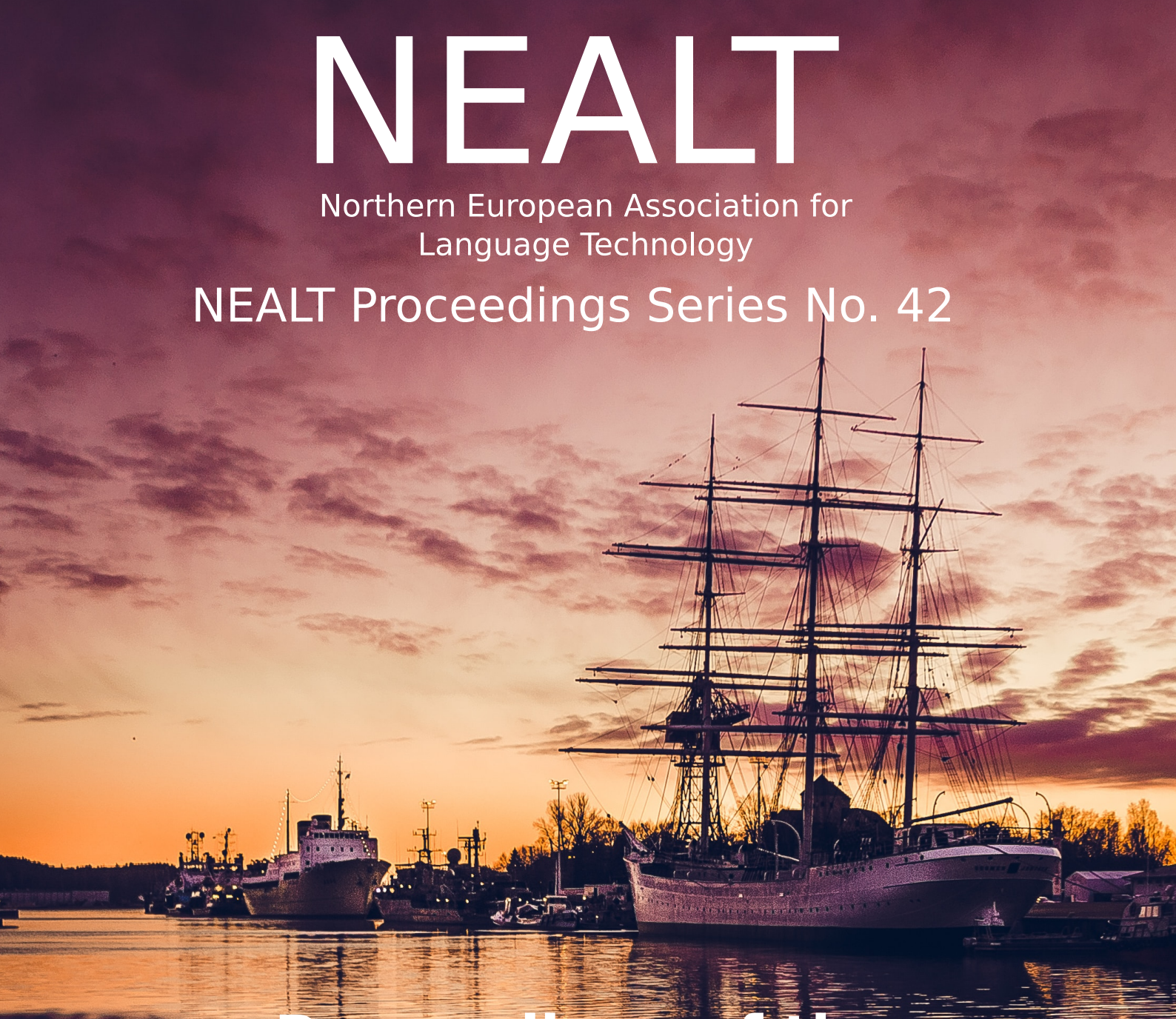
NEALT

Northern European Association for Language Technology

NEALT Proceedings Series No. 42



Proceedings of the 22nd Nordic Conference on Computational Linguistics (NoDaLiDa)

September 30 - October 2, 2019
University of Turku
Turku, Finland

Editors: Mareike Hartmann and Barbara Plank

NoDaLiDa 2019

22nd Nordic Conference on Computational Linguistics (NoDaLiDa)

Proceedings of the Conference

September 30–October 2, 2019
University of Turku
Turku, Finland

Frontcover photo by Patrick Selin on Unsplash

Published by
Linköping University Electronic Press, Sweden
Linköping Electronic Conference Proceedings, No. 167
NEALT Proceedings Series, No. 42
Indexed in the ACL anthology

ISBN: 978-91-7929-995-8
ISSN: 1650-3686
eISSN: 1650-3740

Sponsors

TURKU

Introduction

Welcome to the 22nd Nordic Conference on Computational Linguistics (NoDaLiDa 2019) held at the University of Turku in the beautiful city of Turku in Finland, on September 30-October 2, 2019. The aim of NoDaLiDa is to bring together researchers in the Nordic countries interested in any aspect related to human language and speech technologies. It is a great honor for me to serve as the general chair of NoDaLiDa 2019.

NoDaLiDa has a very long tradition. It stems from a working group initiative led by Sture Allèn, Kolbjörn Heggstad, Baldur Jönsson, Viljo Kohonen and Bente Maegaard (as the preface of the oldest workshop proceedings in the ACL anthology reveals).[1] They organized the first NoDaLiDa ("Nordiska datalingvistikdagar") in Gothenburg on October 10-11, 1977. In 2006, NEALT, the Northern European Association for Language Technology was founded. We are very honored to bring this bi-annual conference after 42 years to Turku this fall.

We solicited three different types of papers (long, short, demo papers) and received 78 valid submissions. In total, we accepted 49 papers, which will be presented as 34 oral presentations, 10 posters and 5 demo papers. A total of 4 submissions were withdrawn in the process. Each paper was reviewed by three experts. We are extremely grateful to the Programme Committee members for their detailed and helpful reviews. Overall, there are 10 oral sessions with talks and one poster session organized into themes over the two days, starting each day with a keynote talk.

We would like to thank our two keynote speakers for travel to Turku and sharing their work. Marie-Catherine de Marneffe from Ohio State University will talk about "Do you know that there's still a chance? Identifying speaker commitment for natural language understanding". Grzegorz Chrupała from Tilburg University will talk about "Investigating neural representations of speech and language". We are also very grateful to Fred Karlsson, who accepted to share his insights into the Finnish language in the traditional NoDaLiDa language tutorial.

The conference is preceded by 5 workshops on a diverse set of topics: deep learning for natural language processing, NLP for Computer-Assisted Language Learning, Constraint Grammar Methods, Tools and Applications, NLP and pseudonymisation and Financial Narrative Processing. This shows the breadth of topics that can be found in language technology these days, and we are extremely happy and grateful to the workshop organizers for complementing the main program this way.

There will be two social events. A reception which is sponsored by the City of Turku and held at the Old Town Hall in Turku. A conference dinner will be held in the Turku Castle in the King's hall. Two fantastic evenings are awaiting.

I would like to thank the entire team that made NoDaLiDa 2019 possible in the first place. First of all, I would like to thank Beáta Megyesi for inviting me to take up this exciting (and admittedly at times demanding) role and all her valuable input regarding NEALT and previous editions of NoDaLiDa. Jörg Tiedemann, for the smooth transition from the previous NoDaLiDa edition and his input and work as program chair; the program chair committee Jurgita Kapočiūtė-Dzikienė, Hrafn Loftsson, Patrizia Paggio, and Erik Velldal, for working hard on putting the program together. I am particularly grateful to Jörg Tiedemann, Jurgita Kapočiūtė-Dzikienė, Kairit Sirts and Patrizia Paggio for leading the reviewing process. Special thanks goes to the workshop chairs Richard Johansson and Kairit Sirts, who have done an invaluable job with leading the workshop selection and organization. A big thanks also to Miryam

[1] https://www.aclweb.org/anthology/events/ws-1977/

de Lhoneux for her work as social media chair and Mareike Hartmann for leading the publication efforts that led to this volume, as well as the coordination of the workshop proceedings. Thank you! Finally, my ultimate thanks goes to the amazing local organization committee and team. Thank you, Filip Ginter and Jenna Kanerva. With your infinite support and pro-active engagement in organizing NoDaLiDa you are the ones that make NoDaLiDa possible and surely an unforgettable experience. Thanks also to the entire local team (with special thanks to Hans Moen for help with the program): Li-Hsin Chang, Rami Ilo, Suwisa Kaewphan, Kai Hakala, Roosa Kyllönen, Veronika Laippala, Akseli Leino, Juhani Luotolahti, Farrokh Mehryary, Hans Moen, Maria Pyykönen, Sampo Pyysalo, Samuel Rönnqvist, Antti Saloranta, Antti Virtanen, Sanna Volanen. NoDaLiDa 2019 has received financial support from our generous sponsors, which we would also like to thank here.

This is the usual place for the greetings from the local organizers, but as we set out to write it, it turns out that Barbara already said it all. So we really only need to add one thing: huge thanks to Barbara for all the hard work she put into NoDaLiDa. We can only wonder where you found the time for all this. We hope the Turku edition of NoDaLiDa will be a success, at least we tried our best to make it so. In two weeks we will know. — Filip, Jenna, and the local team

Danke - kiitos!

We very much hope that you will have an enjoyable and inspiring time at NoDaLiDa 2019 in Turku.

Barbara Plank

København

September 2019

General Chair

Barbara Plank, IT University of Copenhagen, Denmark

Program Committee

Jurgita Kapočiūtė-Dzikienė, Vytautas Magnus University, Lithuania
Hrafn Loftsson, Reykjavík University, Iceland
Patrizia Paggio, University of Copenhagen, Denmark
Jörg Tiedemann, University of Helsinki, Finland
Erik Velldal, University of Oslo, Norway

Organizing Committee

Publication Chair: Mareike Hartmann, University of Copenhagen, Denmark
Social Media Chair: Miryam de Lhoneux, Uppsala University, Sweden
Workshop Chair: Richard Johansson, Chalmers Technical University and University of Gothenburg, Sweden
Workshop Chair: Kairit Sirts, University of Tartu, Estonia
Local Chair: Filip Ginter, University of Turku, Finland
Local Chair: Jenna Kanerva, University of Turku, Finland

Invited Speakers

Marie-Catherine de Marneffe, Ohio State University
Grzegorz Chrupała, Tilburg University

Reviewers

Mostafa Abdou, University of Copenhagen
Yvonne Adesam, Department of Swedish, University of Gothenburg
Lars Ahrenberg, Linköping University
Laura Aina, Pompeu Fabra University
Eivind Alexander Bergem, UiO
Krasimir Angelov, University of Gothenburg and Chalmers University of Technology
Maria Barrett, University of Copenhagen
Valerio Basile, University of Turin
Joachim Bingel, University of Copenhagen
Arianna Bisazza, University of Amsterdam
Kristín Bjarnadóttir, HI.is
Anna Björk Nikulásdóttir, Grammatek ehf
Marcel Bollmann, University of Copenhagen
Gerlof Bouma, University of Gothenburg
Gosse Bouma, Rijksuniversiteit Groningen
Hande Celikkanat, University of Helsinki
Lin Chen, UIC
Jeremy Claude Barnes, University of Oslo
Mathias Creutz, University of Helsinki
Hercules Dalianis, DSV-Stockholm University
Miryam de Lhoneux, Uppsala University
Koenraad De Smedt, University of Bergen
Rodolfo Delmonte, Universita' Ca' Foscari
Leon Derczynski, ITU Copenhagen
Stefanie Dipper, Bochum University
Senka Drobac, University of Helsinki
Jens Edlund, KTH Royal Institute of Technology
Raquel Fernández, University of Amsterdam
Björn Gambäck, Norwegian University of Science and Technology
Filip Ginter, University of Turku
Jon Gudnason, Reykjavik University
Mika Hämäläinen, University of Helsinki
Daniel Hardt, Copenhagen Business School
Petter Haugereid, Western Norway University of Applied Sciences
Daniel Hershcovich, University of Copenhagen
Angelina Ivanova, University of Oslo
Tommi Jauhiainen, University of Helsinki
Anders Johannsen, Apple Inc
Sofie Johansson, Institutionen för svenska språket
Jenna Kanerva, University of Turku
Jussi Karlgren, Gavagai and KTH Royal Institute of Technology
Roman Klinger, University of Stuttgart
Mare Koit, University of Tartu
Artur Kulmizev, Uppsala University

Andrey Kutuzov, University of Oslo
Veronika Laippala, University of Turku
Krister Lindén, University of Helsinki
Nikola Ljubešić, Faculty of Humanities and Social Sciences
Jan Tore Loenning, University of Oslo
Hrafn Loftsson, Reykjavik University
Diego Marcheggiani, Amazon
Bruno Martins, IST and INESC-ID - Instituto Superior Técnico, University of Lisbon
Hans Moen, University of Turku
Costanza Navarretta, University of Copenhagen
Mattias Nilsson, Karolinska Institutet, Department of Clinical Neuroscience
Joakim Nivre, Uppsala University
Farhad Nooralahzadeh, UiO
Pierre Nugues, Lund University, Department of Computer Science Lund, Sweden
Emily Öhman, University of Helsinki
Robert Östling, Department of Linguistics, Stockholm University
Lilja Øvrelid, University of Oslo
Viviana Patti, University of Torino
Eva Pettersson, Uppsala University
Ildikó Pilán, University of Gothenburg
Tommi A Pirinen, University of Hamburg
Alessandro Raganato, University of Helsinki
Taraka Rama, University of Oslo
Vinit Ravishankar, University of Oslo
Marek Rei, University of Cambridge
Nils Rethmeier, DFKI LT-Lab
Corentin Ribeyre, Etermind
Fabio Rinaldi, University of Zurich
Samuel Rönnqvist, University of Turku
Jack Rueter, University of Helsinki
Rune Sætre, Dep. of Computer Science (IDI), Norwegian University of Science and Technology (NTNU) in Trondheim
Magnus Sahlgren, RISE AI
Marina Santini, SICS East ICT
Yves Scherrer, University of Helsinki
Natalie Schluter, IT University of Copenhagen
Ravi Shekhar, University of Trento
Miikka Silfverberg, University of Colorado Boulder
Raivis Skadiņš, Tilde
Aaron Smith, Google
Steinþór Steingrímsson, The Árni Magnússon Institute for Icelandic Studies
Torbjørn Svendsen, Norwegian University of Science and Technology
Nina Tahmasebi, University of Gothenburg
Aarne Talman, University of Helsinki
Samia Touileb, University of Oslo
Francis M. Tyers, Indiana University Bloomington

Martti Vainio, University of Helsinki, Institute of Behavioural Sciences
Rob van der Goot, RuG
Raul Vazquez, University of Helsinki
Erik Velldal, University of Oslo
Sumithra Velupillai, TCS, School of Computer Science and Communication, KTH Royal Institute of Technology
Martin Volk, University of Zurich
Atro Voutilainen, University of Helsinki
Jürgen Wedekind, University of Copenhagen
Mats Wirén, Stockholm University
Anssi Yli-Jyrä, University of Helsinki
Marcos Zampieri, University of Wolverhampton
Heike Zinsmeister, University of Hamburg

Invited Talks

Marie-Catherine de Marneffe: Do you know that there's still a chance? Identifying speaker commitment for natural language understanding.

When we communicate, we infer a lot beyond the literal meaning of the words we hear or read. In particular, our understanding of an utterance depends on assessing the extent to which the speaker stands by the event she describes. An unadorned declarative like "The cancer has spread" conveys firm speaker commitment of the cancer having spread, whereas "There are some indicators that the cancer has spread" imbues the claim with uncertainty. It is not only the absence vs. presence of embedding material that determines whether or not a speaker is committed to the event described: from (1) we will infer that the speaker is committed to there *being* war, whereas in (2) we will infer the speaker is committed to relocating species *not being* a panacea, even though the clauses that describe the events in (1) and (2) are both embedded under "(s)he doesn't believe".

(1) The problem, I'm afraid, with my colleague here, he really doesn't believe that it's war.

(2) Transplanting an ecosystem can be risky, as history shows. Hellmann doesn't believe that relocating species threatened by climate change is a panacea.

In this talk, I will first illustrate how looking at pragmatic information of what speakers are committed to can improve NLP applications. Previous work has tried to predict the outcome of contests (such as the Oscars or elections) from tweets. I will show that by distinguishing tweets that convey firm speaker commitment toward a given outcome (e.g., "Dunkirk will win Best Picture in 2018") from ones that only suggest the outcome (e.g., "Dunkirk might have a shot at the 2018 Oscars") or tweets that convey the negation of the event ("Dunkirk is good but not academy level good for the Oscars"), we can outperform previous methods. Second, I will evaluate current models of speaker commitment, using the CommitmentBank, a dataset of naturally occurring discourses developed to deepen our understanding of the factors at play in identifying speaker commitment. We found that a linguistically informed model outperforms a LSTM-based one, suggesting that linguistic knowledge is needed to achieve robust language understanding. Both models however fail to generalize to the diverse linguistic constructions present in natural language, highlighting directions for improvement.

Grzegorz Chrupała: Investigating Neural Representations of Speech and Language

Learning to communicate in natural language is one of the unique human abilities which are at the same time extraordinarily important and extraordinarily difficult to reproduce in silico. Substantial progress has been achieved in some specific data-rich and constrained cases such as automatic speech recognition or machine translation. However the general problem of learning to use natural language with weak and noisy supervision in a grounded setting is still open. In this talk, I will present recent work which addresses this challenge using deep recurrent neural network models. I will then focus on analytical methods which allow us to better understand the nature and localization of representations emerging in such architectures.

Table of Contents

Long Papers

Short Papers

Demo Papers

Conference Program

Monday, September 30, 2019 Workshops

08:00- **Registration**

09:00-17:00 **The First NLPL Workshop on Deep Learning for Natural Language Processing**
Location: PUB2

09:00-17:30 **The 8th Workshop on Natural Language Processing for Computer-Assisted Language Learning (NLP4CALL)**
Location: PUB5

09:00-15:30 **Constraint Grammar - Methods, Tools and Applications**
Location: PUB 209

14:00-17:00 **The Workshop on NLP and Pseudonymisation**
Location: PUB4

09:00-15:30 **The Second Financial Narrative Processing Workshop (FNP 2019)**
Location: PUB 126

10:00-10:30 *Coffee break*

12:00-14:00 *Lunch break*
Location: Holiday Club Caribia

15:00-15:30 *Coffee break*

19:00 *Welcome Reception*
Location: Turku City Hall

Tuesday, October 1, 2019

09:00-09:15 **Opening**
Location: PUB1

09:15-10:05 **Keynote by Marie-Catherine de Marneffe: Do you know that there's still a chance? Identifying speaker commitment for natural language understanding**
Chair: Joakim Nivre
Location: PUB1

10:05-10:35 *Coffee break*

10:35-12:15 **Parallel session A: Multilinguality and Machine Translation**
Chair: Jörg Tiedemann
Location: PUB1

10:35-11:00 Comparison between NMT and PBSMT Performance for Translating Noisy User-Generated Content
José Carlos Rosales Nuñez, Djamé Seddah and Guillaume Wisniewski

11:00-11:25 Bootstrapping UD treebanks for Delexicalized Parsing
Prasanth Kolachina and Aarne Ranta

11:25-11:50 Lexical Resources for Low-Resource PoS Tagging in Neural Times
Barbara Plank and Sigrid Klerke

11:50-12:15 Toward Multilingual Identification of Online Registers
Veronika Laippala, Roosa Kyllönen, Jesse Egbert, Douglas Biber and Sampo Pyysalo

10:35-12:15 **Parallel session B: Embeddings, Biases and Language Change**
Chair: Richard Johansson
Location: PUB3

10:35-11:00 Gender Bias in Pretrained Swedish Embeddings
Magnus Sahlgren and Fredrik Olsson

11:00-11:25 A larger-scale evaluation resource of terms and their shift direction for diachronic lexical semantics
Astrid van Aggelen, Antske Fokkens, Laura Hollink and Jacco van Ossenbruggen

11:25-11:50 Some steps towards the generation of diachronic WordNets
Yuri Bizzoni, Marius Mosbach, Dietrich Klakow and Stefania Degaetano-Ortlieb

11:50-12:15 An evaluation of Czech word embeddings
Karolína Hořeňovská

12:15-13:45 *Lunch break*
Location: Holiday Club Caribia

13:45-15:00 **Parallel session A: Semantics**
Chair: Marianna Apidianaki
Location: PUB1

13:45-14:10 Language Modeling with Syntactic and Semantic Representation for Sentence Acceptability Predictions
Adam Ek, Jean-Philippe Bernardy and Shalom Lappin

14:10-14:35 Comparing linear and neural models for competitive MWE identification
Hazem Al Saied, Marie Candito and Mathieu Constant

14:35-15:00 A Wide-Coverage Symbolic Natural Language Inference System
Stergios Chatzikyriakidis and Jean-Philippe Bernardy

13:45-15:00 **Parallel session B: Morphology and Syntax**
Chair: Kairit Sirts
Location: PUB3

13:45-14:10 Ensembles of Neural Morphological Inflection Models
Ilmari Kylliäinen and Miikka Silfverberg

14:10-14:35 Nefnir: A high accuracy lemmatizer for Icelandic
Svanhvít Lilja Ingólfsdóttir, Hrafn Loftsson, Jón Friðrik Daðason and Kristín Bjarnadóttir

14:35-15:00 Syntax-based identification of light-verb constructions
Silvio Ricardo Cordeiro and Marie Candito

15:00-15:30 *Coffee Break*

15:30-16:45 **Parallel session A: Machine Learning Applications, Text Classification**
Chair: Jenna Kanerva
Location: PUB1

15:30-15:55 Natural Language Processing in Policy Evaluation: Extracting Policy Conditions from IMF Loan Agreements
Joakim Åkerström, Adel Daoud and Richard Johansson

15:55-16:20 Comparing the Performance of Feature Representations for the Categorization of the Easy-to-Read Variety vs Standard Language
Marina Santini, Benjamin Danielsson and Arne Jönsson

16:20-16:45 Unsupervised Inference of Object Affordance from Text Corpora
Michele Persiani and Thomas Hellström

15:30-16:45 **Parallel session B: Language Resources and Applications**
Chair: Elena Volodina
Location: PUB3

15:30-15:55 Annotating evaluative sentences for sentiment analysis: a dataset for Norwegian
Petter Mæhlum, Jeremy Barnes, Lilja Øvrelid and Erik Velldal

15:55-16:20 Interconnecting lexical resources and word alignment: How do learners get on with particle verbs?
David Alfter and Johannes Graën

16:20-16:45 An Unsupervised Query Rewriting Approach Using N-gram Co-occurrence Statistics to Find Similar Phrases in Large Text Corpora
Hans Moen, Laura-Maria Peltonen, Henry Suhonen, Hanna-Maria Matinolli, Riitta Mieronkoski, Kirsi Telen, Kirsi Terho, Tapio Salakoski and Sanna Salanterä

16:45-17:45 **Poster and demo session**
Location: Entrance hall

16:45-17:45 **Posters:**

Compiling and Filtering ParIce: An English-Icelandic Parallel Corpus
Starkaður Barkarson and Steinþór Steingrímsson

May I Check Again? —A simple but efficient way to generate and use contextual dictionaries for Named Entity Recognition. Application to French Legal Texts.
Valentin Barriere and Amaury Fouret

Predicates as Boxes in Bayesian Semantics for Natural Language
Jean-Philippe Bernardy, Rasmus Blanck, Stergios Chatzikyriakidis, Shalom Lappin and Aleksandre Maskharashvili

DIM: The Database of Icelandic Morphology
Kristín Bjarnadóttir, Kristín Ingibjörg Hlynsdóttir and Steinþór Steingrímsson

Bornholmsk Natural Language Processing: Resources and Tools
Leon Derczynski and Alex Speed Kjeldsen

Morphosyntactic Disambiguation in an Endangered Language Setting
Jeff Ens, Mika Hämäläinen, Jack Rueter and Philippe Pasquier

Tagging a Norwegian Dialect Corpus
Andre Kåsen, Anders Nøklestad, Kristin Hagen and Joel Priestley

The Lacunae of Danish Natural Language Processing
Andreas Kirkedal, Barbara Plank, Leon Derczynski and Natalie Schluter

Tools for supporting language learning for Sakha
Sardana Ivanova, Anisia Katinskaia and Roman Yangarber

Inferring morphological rules from small examples using 0/1 linear programming
Ann Lilliestrom, Koen Claessen and Nicholas Smallbone

16:45-17:45 **Demos:**

LEGATO: A flexible lexicographic annotation tool
David Alfter, Therese Lindström Tiedemann and Elena Volodina

The OPUS Resource Repository: An Open Package for Creating Parallel Corpora and Machine Translation Services
Mikko Aulamo and Jörg Tiedemann

Garnishing a phonetic dictionary for ASR intake
Iben Nyholm Debess, Sandra Saxov Lamhauge and Peter Juel Henrichsen

Docria: Processing and Storing Linguistic Data with Wikipedia
Marcus Klang and Pierre Nugues

UniParse: A universal graph-based parsing toolkit
Daniel Varab and Natalie Schluter

19:30-23:59 *Conference Dinner*
Location: Turku Castle

Wednesday, October 2, 2019

09:00-09:50 **Keynote by Grzegorz Chrupała: Investigating Neural Representations of Speech and Language**
Chair: Lilja Øvrelid
Location: PUB1

09:50-10:20 *Coffee break*

10:20-12:00 **Parallel session A: Sentiment Analysis and Stance**
Chair: Mathias Creutz
Location: PUB1

10:20-10:45 Lexicon information in neural sentiment analysis: a multi-task learning approach
Jeremy Barnes, Samia Touileb, Lilja Øvrelid and Erik Velldal

10:45-11:10 Aspect-Based Sentiment Analysis using BERT
Mickel Hoang, Oskar Alija Bihorac and Jacobo Rouces

11:10-11:35 Political Stance Detection for Danish
Rasmus Lehmann and Leon Derczynski

11:35-12:00 Joint Rumour Stance and Veracity Prediction
Anders Edelbo Lillie, Emil Refsgaard Middelboe and Leon Derczynski

10:20-12:00 **Parallel session B: Named Entity Recognition**
Chair: Manex Agirrezabal
Location: PUB3

10:20-10:45 Towards High Accuracy Named Entity Recognition for Icelandic
Svanhvít Lilja Ingólfsdóttir, Sigurjón Þorsteinsson and Hrafn Loftsson

10:45-11:10 Named-Entity Recognition for Norwegian
Bjarte Johansen

11:10-11:35 Neural Cross-Lingual Transfer and Limited Annotated Data for Named Entity Recognition in Danish
Barbara Plank

11:35-12:00 Projecting named entity recognizers without annotated or parallel corpora
Jue Hou, Maximilian Koppatz, José María Hoya Quecedo and Roman Yangarber

12:00-13:00 *Lunch break*
Location: Holiday Club Caribia

13:00-14:00 **NEALT business meeting**
Location: PUB1

14:00-15:15 **Parallel session A: Text Generation and Language Model Applications**
Chair: Leon Derczynski
Location: PUB1

14:00-14:25 Template-free Data-to-Text Generation of Finnish Sports News
Jenna Kanerva, Samuel Rönnqvist, Riina Kekki, Tapio Salakoski and Filip Ginter

14:25-14:50 Matching Keys and Encrypted Manuscripts
Eva Pettersson and Beata Megyesi

14:50-15:15 The Seemingly (Un)systematic Linking Element in Danish
Sidsel Boldsen and Manex Agirrezabal

14:00-15:15 **Parallel session B: Speech**
Chair: Grzegorz Chrupała
Location: PUB3

14:00-14:25 Perceptual and acoustic analysis of voice similarities between parents and young children
Evgeniia Rykova and Stefan Werner

14:25-14:50 Enhancing Natural Language Understanding through Cross-Modal Interaction: Meaning Recovery from Acoustically Noisy Speech
Ozge Alacam

14:50-15:15 Predicting Prosodic Prominence from Text with Pre-trained Contextualized Word Representations
Aarne Talman, Antti Suni, Hande Celikkanat, Sofoklis Kakouros, Jörg Tiedemann and Martti Vainio

15:15-15:45 *Coffee Break*

15:45-16:25 **Tutorial on Finnish by Fred Karlsson**
Chair: Filip Ginter
Location: PUB1

16:25-16:35 **Closing**
Location: PUB1