COLING 2020

The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature

Co-located with the 28th International Conference on Computational Linguistics COLING'2020

Proceedings

December 12, 2020
Barcelona, Spain (Online)

Copyright of each paper stays with the respective authors (or their employers).

ISBN 978-1-952148-34-7

Preface

These are very strange times. LaTeCH-CLfL has joined the swelling ranks of virtual scientific meetings. We all have next to no experience with such events – and yes, we hope that next year we will not need that experience any more. The format of the workshop is an experiment. You can access on-line, in advance, all talks and all posters. We will not bother you with detailed introductions, other than to say that the range of topics of the accepted papers has met a good deal of the expectations in the call for papers.

The actual workshop will consist of an invited talk (thank you, Elke Teich), brief Q&A sessions for the oral presentations (which you will have watched by then), and a poster session during which you will be able to chat with any author you like.

Here is a bit of statistics for those who care about such numbers. We have received unusully many submissions (thanks, everyone). We have accepted 20 papers for the 42.5% acceptance rate. Let us express our deep appreciation for the work of our wonderful program committee: you rock!

Keep well.

Stefania, Nils, Stan, Anna

`https://sighum.wordpress.com/events/latech-clfl-2020/`

Invited Talk

Linguistic variation and the dynamics of language use

It is widely acknowledged that linguistic variation is a core feature of language, affecting all linguistic levels from the phonetic to the semantic level. Linguistic variation emerges and is reinforced through language use in context, continuously adapting to social and cognitive constraints. Language use thus provides excellent data for studying (changing) socio-cultural practices as well as the (general) mechanisms of human communication.

In my talk I focus on two opposing but complementary effects to be observed in the dynamics of language use: innovation and conventionalization. Innovation leads to an expansion of linguistic options by new linguistic coinages, e.g. new words entering language or known words being used in new contexts. Conventionalization leads to a reduction of options by convergence in linguistic usage, i.e. the tacit agreement on "how to say things" often associated with a specific style or register. I will show that while innovation and conventionalization pull in different directions, they interact in specific ways to keep language intact for communication.

The underlying approach is corpus-based, using data-driven methods. Language models (e.g. word embeddings) are combined with selected information-theoretic measures (entropy, surprisal), providing models of language use and indices of linguistic variation (here: with special regard of innovation and conventionalization). I will focus on the domain of scientific writing (English) from a diachronic perspective with side glimpses at translation in the domain of European Parliament.

About the speaker

Elke Teich
Department of Language Science and Technology
Saarland University

Elke Teich is a full professor of English Linguistics and Translation at the Department of Language Science and Technology, Saarland University, Saarbrücken, Germany. Since 2014 she has been the head of the Collaborative Research Center "SFB 1102 Information Density and Linguistic Encoding" funded by the German Research Foundation (DFG). She is currently a principal investigator on two projects in SFB 1102, one on diachronic language change and one on human translation, as well as the Saarbrücken Cluster of Excellence Multimodal Computing and Interaction (MMCI) and the German CLARIN project (Common Language Resources and Technology Infrastructure). Elke Teich is an editorial board member of several journals and book series, including 'Languages in Contrast' (Benjamins) and 'Linguistics and the Human Sciences (Equinox)'. She is a regular reviewer for national and international funding agencies, including Deutsche Forschungsgemeinschaft (DFG), Humboldt Foundation, Schweizer Nationalfonds and the Finnish Academy.

Teich's expertise ranges from descriptive grammar of English and German over (multi-lingual) register analysis with a special focus on scientific language to translatology. She worked on machine translation, automatic text generation, corpus linguistics and the digital humanities at the following academic institutions: Gesellschaft für Mathematik und Datenverarbeitung (Fraunhofer), Information Sciences Institute (ISI)/USC Los Angeles, University of Sydney, Macquarie University and Technical University Darmstadt. Her research focus in the last 10 years has been on developing computationally based approaches to modelling language variation and change.

Organizers:

Stefania Degaetano-Ortlieb, Department of Language Science and Technology, Universität des Saarlandes
Anna Kazantseva, National Research Council of Canada
Nils Reiter, Institute for Natural Language Processing (IMS), Stuttgart University / Institute for Digital Humanities (IDH), Cologne University
Stan Szpakowicz, School of Electrical Engineering and Computer Science, University of Ottawa

Program Committee:

Beatrice Alex, University of Edinburgh, United Kingdom
Melanie Andresen, Hamburg University, Germany
JinYeong Bak, Sungkyunkwan University, South Korea
Andre Blessing, University of Stuttgart, Germany
Gosse Bouma, University of Groningen, The Netherlands
Julian Brooke, University of British Columbia, Canada
Paul Buitelaar, National University of Ireland, Galway, Ireland
Miriam Butt, University of Konstanz, Germany
Gerard de Melo, Tsinghua University, China
Thierry Declerck, Deutsche Forschungszentrum für Künstliche Intelligenz GmbH, Germany
Stefanie Dipper, Ruhr-University, Bochum, Germany
Jacob Eisenstein, Georgia Institute of Technology, United States
Anna Feldman, Montclair State University, United States
Mark Finlayson, Florida International University, United States
Antske Fokkens, Vrije Universiteit Amsterdam, The Netherlands
Udo Hahn, Friedrich-Schiller-Universität Jena, Germany
Mika Hämäläinen, University of Helsinki, Finland
Serge Heiden, École normale supérieure de Lyon, France
Graeme Hirst, University of Toronto, Canada
Fotis Jannidis, Würzburg University, Germany
Adam Jatowt, Kyoto University, Japan
Mike Kestemont, University of Antwerp, Belgium
Dimitrios Kokkinakis, University of Gothenburg, Sweden
Stasinos Konstantopoulos, National Centre of Scientific Research “Demokritos”, Greece
Markus Krug, Würzburg University, Germany
Jonas Kuhn, University of Stuttgart, Germany
John Lee, City University of Hong Kong, Hong Kong
Chaya Liebeskind, Jerusalem College of Technology, Israel
Tom Lippincott, Johns Hopkins University, United States
Barbara McGillivray, The Alan Turing Institute, United Kingdom
Vivi Nastase, University of Stuttgart, Germany
Borja Navarro Colorado, University of Alicante, Spain
John Nerbonne, University of Freiburg, Germany
Pierre Nugues, Lund University, Sweden
Petya Osenova, Sofia University and IICT-BAS, Bulgaria
Andrew Piper, McGill University, Canada
Thierry Poibeau, CNRS Paris and Lattice, France

Georg Rehm, DFKI, Germany
Martin Reynaert, Tilburg University, Radboud University Nijmegen, The Netherlands
Pablo Ruiz Fabo, Université de Strasbourg, France
Marijn Schraagen, Utrecht University, The Netherlands
Eszter Simon, Petőfi Literary Museum, Hungary
Caroline Sporleder, Göttingen University, Germany
Elke Teich, Saarland University, Germany
Ulrich Tiedau, University College London, United Kingdom
Ted Underwood, University of Illinois, Urbana-Champaign, United States
Krishnapriya Vishnubhotla, University of Toronto, Canada
Rob Voigt, Northwestern University, United States
Menno van Zaanen, South African Centre for Digital Language Resources, Potchefstroom, South Africa
Kalliopi Zervanou, Utrecht University, The Netherlands
Heike Zinsmeister, University of Hamburg, Germany

Invited Speaker:

Elke Teich
Department of Language Science and Technology, Saarland University
"Linguistic variation and the dynamics of language use"

Table of Contents

Workshop program

Invited talk

Linguistic variation and the dynamics of language use
Elke Teich

Regular talks

Automatic Topological Field Identification in (Historical) German Texts
Katrin Ortmann

Exhaustive Entity Recognition for Coptic: Challenges and Solutions
Amir Zeldes, Lance Martin and Sichang Tu

Neural Machine Translation of Artwork Titles Using Iconclass Codes
Nikolay Banar, Walter Daelemans and Mike Kestemont

Measuring the Effects of Bias in Training Data for Literary Classification
Sunyam Bagga and Andrew Piper

ERRANT: Assessing and Improving Grammatical Error Type Classification
Katerina Korre and John Pavlopoulos

Life still goes on: Analysing Australian WW1 Diaries through Distant Reading
Ashley Dennis-Henderson, Matthew Roughan, Lewis Mitchell and Jonathan Tuke

Interpretation of Sentiment Analysis in Aeschylus's Greek Tragedy
Vijaya Kumari Yeruva, Mayanka ChandraShekar, Yugyung Lee, Jeff Rydberg-Cox, Virginia Blanton and Nathan A Oyler

Finding and Generating a Missing Part for Story Completion
Yusuke Mori, Hiroaki Yamane, Yusuke Mukuta and Tatsuya Harada

Posters

History to Myths: Social Network Analysis for Comparison of Stories over Time
Clément Besnier

A Survey on Approaches to Computational Humor Generation
Miriam Amin and Manuel Burghardt

A Two-Step Approach for Automatic OCR Post-Correction
Robin Schaefer and Clemens Neudecker

"Shakespeare in the Vectorian Age" – An evaluation of different word embeddings and NLP parameters for the detection of Shakespeare quotes
Bernhard Liebl and Manuel Burghardt

Vital Records: Uncover the past from historical handwritten records
Herve Dejean and Jean-Luc Meunier

Zero-shot cross-lingual identification of direct speech using distant supervision
Murathan Kurfalı and Mats Wirén

Twenty-two Historical Encyclopedias Encoded in TEI: a New Resource for the Digital Humanities
Thora Hagen, Erik Ketzan, Fotis Jannidis and Andreas Witt

Results of a Single Blind Literary Taste Test with Short Anonymized Novel Fragments
Andreas van Cranenburgh and Corina Koolen

Geometric Deep Learning Models for Linking Character Names in Novels
Marek Kubis

Sonnet Combinatorics
Thierry Poibeau, Mylène Maignant, Frédérique Mélanie-Becquet, Clément Plancq, Matthieu Raffard and Mathilde Roussel

Towards Olfactory Information Extraction from Text: A Case Study on Detecting Smell Experiences in Novels
Ryan Brate, Paul Groth and Marieke van Erp

TL-Explorer: A Digital Humanities Tool for Mapping and Analyzing Translated Literature
Alex Zhai, Zheng Zhang, Amel Fraisse, Ronald Jenn, Shelley Fisher Fishkin and Pierre Zweigenbaum