LChange'21

The 2nd International Workshop on Computational Approaches to Historical Language Change 2021

Proceedings of the Workshop

August 6, 2021
Bangkok, Thailand (online)

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-954085-60-2

Message from Organizers

Welcome to the 2nd International Workshop on Computational Approaches to Historical Language Change (LChange'21) co-located with ACL-IJCNLP 2021 on August 6, 2021 and held virtually. Following the success of the first workshop organized at ACL 2019, we are pleased to present the proceedings of the second installment of this workshop series.

The sociocultural and technological development in the world is closely connected to our language as a means to facilitate efficient communication. As a consequence, human language changes over time. Traces of these changes are apparent in our texts and important to anyone who either directly studies changing phenomena, or who uses diachronic texts as a basis for their studies. This workshop explores the phenomena of language change found in written text, on the topics of computational methodologies, theories and digital text resources for exploring the time-varying nature of human language. Its aim is three-fold. First, we want to provide an outlet for pioneering researchers who work on computational methods, evaluation, and large-scale modelling of language change to disseminate cutting-edge research on topics concerning language change. We intended this workshop as a platform for sharing state-of-the-art research progress in this fundamental domain of natural language research. Second, we want to bring together domain experts across disciplines including but not restricted to linguistics, natural language processing, computer science, cognitive psychology, history and digital humanities. Third, the detection and modelling of language change using diachronic text and text mining raise fundamental theoretical and methodological challenges for future research in this area. We hope to engage corpus and computational linguists, (big-) data scientists, as well as humanities and social science scholars to address these open issues.

In response to the call we received 16 submissions.[1] Each of them was carefully evaluated by three members of the Program Committee, whom we believed to be most appropriate for each paper. Based on the reviewers' feedback we accepted 9 full and short papers as oral presentations or as poster papers. We had two distinguished keynote presentations: the first by Maria Koptjevskaja-Tamm (Stockholm University) and Tatiana Nikitina (LLACAN – "Languages and cultures of Africa", CNRS) who presented a talk titled "Linguistic diversity as a testing ground for the study of semantic change", and the second by Alexander Koplenig (Leibniz-Institute for the German Language in Mannheim) with the talk "Two challenges we face when analyzing diachronic corpora". Finally, we have invited 6 findings papers from ACL2021 to be presented either orally or as posters, which are not included in the workshop proceedings.

We hope that you will find the workshop papers insightful and inspiring. We would like to thank the keynote speakers for their stimulating talks, the authors of all papers for their interesting contributions and the members of the Program Committee for their insightful reviews. Our special thanks go to the emergency reviewers who stepped in to provide their expertise. We also express our gratitude to the ACL 2021 workshop chairs for their kind assistance during the organisation process, and for arranging the logistics and infrastructure allowing us to hold LChange'21 online. Finally, our thanks go towards our silver sponsor iguanodon.ai.

Nina Tahmasebi, workshop chair, University of Gothenburg (Sweden)
Adam Jatowt, University of Innsbruck (Austria)
Yang Xu, University of Toronto (Canada)
Simon Hengchen, University of Gothenburg (Sweden)
Syrielle Montariol, INRIA Paris (France)
Haim Dubossarsky, University of Cambridge (United Kingdom)
LChange'21 Workshop Chairs

[1] The number of submissions is a significant drop from the 2019 workshop, following a trend of lower workshop submissions as a consequence of the Covid pandemic.

Organizing Committee:

Nina Tahmasebi, workshop chair, University of Gothenburg (Sweden)
Adam Jatowt, University of Innsbruck (Austria)
Yang Xu, University of Toronto (Canada)
Simon Hengchen, University of Gothenburg (Sweden)
Syrielle Montariol, INRIA Paris (France)
Haim Dubossarsky, University of Cambridge (United Kingdom)

Program Committee:

Ehsaneddin Asgari, University of California, Berkeley (United States)
Zhenisbek Assylbekov, Nazarbayev University (Kazakhstan)
Pierpaolo Basile, Department of Computer Science, University of Bari Aldo Moro (Italy)
Christin Beck, University of Konstanz (Germany)
Barend Beekhuizen, University of Toronto (Canada)
Klaus Berberich, Saarbruecken University of Applied Sciences (Germany)
Chris Biemann, University of Hamburg (Germany)
Damián Blasi, Harvard University and Max Planck Institute for the Science of Human History (United States)
Annalina Caputo, Dublin City University, ADAPT Centre, I-Form Centre (Ireland)
Pierluigi Cassotti, University of Bari Aldo Moro (Italy)
Brady Clark, Northwestern University (United States)
Paul Cook, University of New Brunswick (Canada)
Yijun Duan, AIST (Japan)
Michael Färber, Karlsruhe Institute of Technology (Germany)
Lauren Fonteyn, Leiden University (Netherlands)
Karlien Franco, KU Leuven (Belgium)
Mario Giulianelli, University of Amsterdam (Netherlands)
Maurício Gruppi, Rensselaer Polytechnic institute (United States)
Mika Hämäläinen, University of Helsinki (Finland)
Ran Iwamoto, Keio University (Japan)
Vaibhav Jain, Independent Scholar (India)
Abhik Jana, University of Hamburg (Germany)
Tommi Jauhiainen, University of Helsinki (Finland)
Péter Jeszenszky, University of Bern (Switzerland)
Richard Johansson, University of Gothenburg (Sweden)
Jens Kaiser, University of Stuttgart (Germany)
Vani Kanjirangat, IDSIA (Switzerland)
Andres Karjus, University of Tartu (Estonia)
Tom Kenter, Google UK (United Kingdom)
Andrey Kutuzov, University of Oslo (Norway)
Anna Marakasova, TU Wien (Austria)
Matej Martinc, Jozef Stefan Institute (Slovenia)
Barbara McGillivray, The Alan Turing Institute and University of Cambridge (United Kingdom)
Filip Miletic, CLLE, CNRS /University of Toulouse (France)
Animesh Mukherjee, IIT Kharagpur (India)
Kjetil Norvag, NTNU (Norway)
Krzysztof Nowak, Institute of Polish Language (Poland)

Paul Nulty, University College Dublin (Ireland)
Maike Park, Leibniz-Institute for German Language (Germany)
Stefano De Pascale, KU Leuven (Belgium)
Xutan Peng, The University of Sheffield (United Kingdom)
Peter Petré, University of Antwerp (Belgium)
Lidia Pivovarova, University of Helsinki (Finland)
Martin Pömsl, Osnabrück University (Germany)
Pavel Přibáň, University of West Bohemia (Czech Republic)
Taraka Rama, University of North Texas at Denton (United States)
Julia Rodina, National Research University Higher School of Economics (Russia)
Eyal Sagi, Northwestern University (United States)
Tanja Säily, University of Helsinki (Finland)
Dominik Schlechtweg, University of Stuttgart (Germany)
Sandeep Soni, Georgia Institute of Technology (United States)
Andreas Spitz, University of Konstanz (Germany)
Ian Stewart, University of Michigan (United States)
Suzanne Stevenson, University of Toronto (Canada)
Ludovic Tanguy, CLLE: University of Toulouse and CNRS (France)
Stephen Taylor, University of West Bohemia (Czech Republic)
Rocco Tripodi, Alma Mater Studiorum - University of Bologna (Italy)
Melvin Wevers, DHLAB, KNAW Humanities Cluster (Netherlands)
Ekaterina Vylomova, University of Melbourne (Australia)
Frank D. Zamora-Reina, University of Chile (Chile)
Yihong Zhang, Osaka University (Japan)
Elaine Zosa, University of Helsinki (Finland)

Invited Speakers:

Alexander Koplenig, Leibniz-Institute for the German Language in Mannheim (Germany)
Maria Koptjevskaja-Tamm, Stockholm University (Sweden)
Tatiana Nikitina, LLACAN – “Languages and cultures of Africa”, CNRS (France)

Keynote abstracts:

Keynote 1
Speakers: Maria Koptjevskaja-Tamm, Stockholm University (Sweden) and Tatiana Nikitina, LLACAN – "Languages and cultures of Africa", CNRS (France)

Title of talk: **Linguistic diversity as a testing ground for the study of semantic change**
Abstract: There are between 6000 and 8000 languages currently spoken in the world. The majority of those still lack decent descriptions, not to mention any written tradition and sizeable documents to rely on while trying to trace semantic changes they have undergone in the past and understanding the mechanisms behind them. Understandably, but likewise regrettably, most of the theoretical thinking in linguistics and adjacent disciplines has been formed by research on a few very big languages with a long written tradition, and the same has to a large extent been carried over to computational approaches, including work on semantic change. In our talk we will focus on two big issues which we believe deserve more awareness and attention among researchers involved in computational approaches to historical language change:

- A crucial part in any theoretical work consists of formulating hypotheses, generalizations, laws etc. and explaining them, and work on semantic change is, of course, no exception. Linguistic diversity does not imply that any such generalizations are meaningless or premature before these have been studied for all the world's languages. It does imply, though, that such generalizations gain a lot from careful systematic cross-linguistic research that may unveil cross-linguistic regularities behind diversity – which is foundational for linguistic typology. Here we will discuss several cases whereby such research has questioned earlier generalizations on semantic change based on the familiar languages and/or has come up with new hypotheses.
- But given that the majority of the world's languages lack any written tradition and sizeable historical documents, how is it possible to study semantic changes they have undergone in the past? This is indeed a big challenge, but not an insurmountable one. We will discuss several methods which often combine a careful intragenetic comparison (i.e., comparison of closely related languages) and a broader cross-linguistic perspective and some of the results obtained by their application.

Keynote 2
Speaker: Alexander Koplenig, Leibniz-Institute for the German Language in Mannheim (Germany)

Title of talk: **Two challenges we face when analyzing diachronic corpora.**
Abstract: In my keynote, I want to discuss two important challenges for the quantitative analysis of diachronic corpora that I believe deserve more attention:

- The first challenge is the systematic influence of the sample size when it comes to basically all measures in quantitative linguistics (Baayen 2001). By analysing the lexical dynamics of the German weekly news magazine "Der Spiegel" (consisting of approximately 365,000 articles and 237,000,000 words that were published between 1947 and 2017), I show that this influence makes it difficult to quantify lexical dynamics and language change. I will also demonstrate that standard sampling approaches do not solve this problem. I will

suggest an approach that is able to break the sample size dependence but presupposes access to the full text data (Koplenig, Wolfer & Müller-Spitzer 2019).

- The second challenge is of methodological nature and relates to the problem of representativeness of diachronic corpora. Labov (1994) famously stated that "historical documents survive by chance, not by design, and the selection that is available is the product of an unpredictable series of historical accidents." By using both Google Books Ngram data (Michel et al. 2010; Koplenig 2015; Pechenick, Danforth & Dodds 2015) and publicly available data from the German National Bibliography, I will try to show that the problem is even more fundamental, because there is good reason to believe that composition of the body of published written works (from which a corresponding corpus is supposed to be sampled from) systematically changes as a function of time. This makes it difficult to disentangle actual language change from environmental changes in the textual habitat (Szmrecsanyi 2016).

Table of Contents

Conference Program

August 6, 2021 [UTC+0]

07:00-07:15 ***Introduction***

07:15-08:30 **Session 1**

07:15-07:40 *Time-Aware Ancient Chinese Text Translation and Inference*
Ernie Chang, Yow-Ting Shiue, Hui-Syuan Yeh and Vera Demberg LChange'21

07:40-08:05 *Three-part diachronic semantic change dataset for Russian*
Andrey Kutuzov and Lidia Pivovarova LChange'21

08:05-08:30 *The Corpora They Are a-Changing: a Case Study in Italian Newspapers*
Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti and Rossella Varvara LChange'21

08:30-09:05 ***Break***

09:05-09:30 *Linguistic change and historical periodization of Old Literary Finnish*
Niko Partanen, Khalid Alnajjar, Mika Hämäläinen and Jack Rueter LChange'21

09:30-10:00 *Studying the Evolution of Scientific Topics and their Relationships*
Ana Sabina Uban, Cornelia Caragea and Liviu P. Dinu Findings

10:00-10:30 *When Time Makes Sense: A Historically-Aware Approach to Targeted Sense Disambiguation*
Kaspar Beelen, Federico Nanni, Mariona Coll Arduany, Kasra Hosseini, Giorgia Tolfo and Barbara McGillivray Findings

10:30-11:30 ***Lunch break***

11:30-12:30 **Keynote 1:** *Linguistic diversity as a testing ground for the study of semantic change*
Maria Koptjevskaja-Tamm and Tatiana Nikitina

12:30-14:00 **Online poster session**

A diachronic evaluation of gender asymmetry in euphemism
Anna Kapron-King and Yang Xu LChange'21

The GLAUx corpus: methodological issues in designing a long-term, diverse, multi-layered corpus of Ancient Greek LChange'21
Alek Keersmaekers

Bhāṣācitra: Visualising the dialect geography of South Asia
Aryaman Arora, Adam Farris, Gopalakrishnan R and Samopriya Basu — LChange'21

Modeling the Evolution of Word Senses with Force-Directed Layouts of Co-occurrence Networks
Tim Reke, Robert Schwanhold and Ralf Krestel — LChange'21

Tracking Semantic Change in Cognate Sets for English and Romance Languages
Ana Sabina Uban, Alina Maria Cristea, Anca Dinu, Liviu P. Dinu, Simona Georgescu and Laurentiu Zoicas — LChange'21

Can Cognate Prediction Be Modelled as a Low-Resource Machine Translation Task?
Clementine Fourrier, Rachel Bawden, Benoit Sagot — Findings

Event Extraction from Historical Texts: A New Dataset for Black Rebellions
Viet Dac Lai, Minh Van Nguyen, Heidi Kaufman and Thien Huu Nguyen — Findings

Sequence Models for Computational Etymology of Borrowings
Winston Wu, Kevin Duh, David Yarowsky — Findings

A Formidable Ability: Detecting Adjectival Extremeness with DSMs
Farhan Samir, Barend Beekhuizen and Suzanne Stevenson — Findings

14:00-15:00 **Keynote 2:** *Two challenges we face when analyzing diachronic corpora*
Alexander Koplenig