Proceedings of the

8th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2019)

edited by

David Alfter, Elena Volodina, Lars Borin, Ildikó Pilán and Herbert Lange

Front cover photo by ketsa49@pixabay
Licensed under a Pixabay license:
`https://pixabay.com/de/service/license/`

Linköping Electronic Conference Proceedings 164
eISSN 1650-3740 (Online) • ISSN 1650-3686 (Print) 2019
ISBN 978-91-7929-998-9

Preface

The workshop series on Natural Language Processing (NLP) for Computer-Assisted Language Learning (NLP4CALL) is a meeting place for researchers working on the integration of Natural Language Processing and Speech Technologies in CALL systems and exploring the theoretical and methodological issues arising in this connection. The latter includes, among others, insights from Second Language Acquisition (SLA) research, on the one hand, and promote development of “Computational SLA” through setting up Second Language research infrastructure(s), on the other.

The intersection of Natural Language Processing (or Language Technology / Computational Linguistics) and Speech Technology with Computer-Assisted Language Learning (CALL) brings “understanding” of language to CALL tools, thus making CALL intelligent. This fact has given the name for this area of research – Intelligent CALL, ICALL. As the definition suggests, apart from having excellent knowledge of Natural Language Processing and/or Speech Technology, ICALL researchers need good insights into second language acquisition theories and practices, as well as knowledge of second language pedagogy and didactics. This workshop invites therefore a wide range of ICALL-relevant research, including studies where NLP-enriched tools are used for testing SLA and pedagogical theories, and vice versa, where SLA theories, pedagogical practices or empirical data are modeled in ICALL tools. The NLP4CALL workshop series is aimed at bringing together competences from these areas for sharing experiences and brainstorming around the future of the field.

We invited papers:

- that describe research directly aimed at ICALL;
- that demonstrate actual or discuss the potential use of existing Language and Speech Technologies or resources for language learning;
- that describe the ongoing development of resources and tools with potential usage in ICALL, either directly in interactive applications, or indirectly in materials, application or curriculum development, e.g. learning material generation, assessment of learner texts/responses, individualized learning solutions, provision of feedback;
- that discuss challenges and/or research agenda for ICALL;
- that describe empirical studies on language learner data

As in the previous edition of the workshop, a special focus was given to the established and upcoming infrastructures aimed at SLA and learner corpus research, covering questions such as data collection, legal issues, reliability of annotation, annotation tool development, search environments for SLA-relevant data, etc. We encouraged paper presentations and software demonstrations describing the above-mentioned themes primarily, but not exclusively, for the Nordic languages.

This year, we had the pleasure to welcome two invited speakers: Thomas François (Université catholique de Louvain) and Egon Stemle (Eurac Research).

David Alfter, Elena Volodina, Lars Borin, Ildikó Pilán and Herbert Lange 2019. Preface. *Proceedings of the 8th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2019).*
Linköping Electronic Conference Proceedings 164: i–v.

Thomas François is Assistant Professor in Applied Linguistics and Natural Language Processing at UCLouvain (Cental). His work focuses on automatic assessment of text readability, automatic text simplification, complex word identification, efficient communication in business, and the use of French as a professional language. He has been an invited researcher at IRCS (University of Pennsylvania) as a Fulbright and BAEF fellow and, later, has been a FNRS post-doctoral researcher. He has led research projects such as CEFRLex[1], a CEFR-graded lexicon for foreign language learning or AMesure[2], a platform to support simple writing. His work on readability for French as a foreign language has been awarded the best thesis Award by the ATALA in 2012 and the best paper in the TALN2016 conference.

In this talk entitled *Assessing language complexity for L2 readers with NLP techniques and corpora*, he summarized the main trends regarding the automatic assessment of language complexity for L2 readers and focus on three research projects. To illustrate the readability approach, the DMesure project was presented. It is the first computational readability formula specialized for readers of French as a foreign language. Secondly, the talk discussed the use of corpora to assess language complexity through CEFRLex, an international project providing, for some of the main European languages, lexical resources describing the frequency distributions of words across the six levels of competence of the Common European Framework of Reference for Languages (CEFR). These distributions have been estimated on corpora of pedagogical materials intended for L2 purposes such as textbooks and simplified readers. The resulting resources have been manually checked and are machine-readable and open-licensed. The project also offers an interface allowing to automatically assess difficult words in a text in accordance with CEFRLex knowledge. Thirdly, the Predicomplex project illustrated the use of learner data. It consists in a personalized approach of vocabulary knowledge prediction using machine learning algorithms. He concluded his talk by highlighting some of the current challenges and research opportunities relative to language difficulty assessment for L2 learners.

Egon Stemle is a researcher in the Institute for Applied Linguistics at Eurac Research, Bolzano, Italy. He is a cognitive scientist with a focus in the area where computational linguistics and artificial intelligence converge. He works on the creation, standardisation, and interoperability of tools for editing, processing, and annotating linguistic data and enjoys working together with other scientists on their data but also collects or helps to collect new data from the Web, from computer-mediated communication and social media, and from language learners. He is an advocate of open science to make research and data available for others to consult or reuse in new research.

In recent years, the reproducibility of scientific research has become increasingly important, both for external stakeholders and for the research communities themselves. They all demand that empirical data collected and used for scientific research is managed and preserved in a way that research results are reproducible. In order to account for this, the FAIR guiding principles for data stewardship have been established as a framework for good data management aiming at the findability, accessibility, interoperability, and reusability of research data. A special role is played by natural language processing and its methods, which are an integral part of many other disciplines working with language data: Language corpora are often living objects – they are

[1] http://cental.uclouvain.be/cefrlex/

[2] http://cental.uclouvain.be/amesure/

constantly being improved and revised, and at the same time the processing tools are also regularly updated, which can lead to different results for the same processing steps.

In his talk entitled *Towards an infrastructure for FAIR language learner corpora*, he first investigated CMC corpora, which resemble language learner corpora in some core aspects, with regard to their compliance with the FAIR principles and discuss to what extent the deposit of research data in repositories of data preservation initiatives such as CLARIN, Zenodo or META-SHARE can assist in the provision of FAIR corpora. Second, he showed some modern software technologies and how they make the process of software packaging, installation, and execution and, more importantly, the tracking of corpora throughout their life cycle reproducible. This in turn makes changes to raw data reproducible for many subsequent analyses.

Previous workshops

This workshop follows a series of workshops on NLP4CALL organized by the NEALT Special Interest Group on Intelligent Computer-Assisted Language Learning (SIG-ICALL[3]). The workshop series has previously been financed by the Center for Language Technology[4] at the University of Gothenburg, and the Swedish Research Council's conference grant.

Submissions to the eight workshop editions have targeted a wide range of languages, ranging from well-resources languages (Chinese, German, English, French, Portuguese, Russian, Spanish) to lesser-resourced languages (Erzya, Arabic, Estonian, Irish, Komi-Zyrian, Meadow Mari, Saami, Udmurt, Võro). Among these, several Nordic languages have been targeted, namely Danish, Estonian, Finnish, Icelandic, Norwegian, Saami, Swedish and Võro. The wide scope of the workshop is also evident in the affiliations of the participating authors as illustrated in Table 1.

COUNTRY	**NUMBER OF AUTHORS**
Australia	2
Belgium	4
Canada	4
Denmark	2
Estonia	3
Finland	9
France	6
Germany	77
Iceland	3
Ireland	2
Japan	2
Netherlands	1
Norway	12
Portugal	5
Russia	10
Slovakia	1
Spain	3
Sweden	62

[3] https://spraakbanken.gu.se/swe/forskning/ICALL/SIG-ICALL
[4] http://clt.gu.se

Switzerland	10
UK	1
US	5

Table 1: Authors by affiliation country, 2012-2019

The acceptance rate has varied between 50% and 77%, the average being 64% (see Table 2). Although the acceptance rate is rather high, the reviewing process has always been very rigorous with two to three double-blind reviews per submission. This indicates that submissions to the workshop have usually been of high quality.

WORKSHOP YEAR	**SUBMITTED**	**ACCEPTED**	**ACCEPTANCE RATE**
2012	12	8	67%
2013	8	4	50%
2014	13	10	77%
2015	9	6	67%
2016	14	10	72%
2017	13	7	54%
2018	16	11	69%
2019	16	10	63%

Table 2: Submissions and acceptance raters, 2012-2019

We would like to thank our Program Committee for providing detailed feedback on the reviewed papers.

- Lars Ahrenberg, Linköping University, Sweden
- David Alfter, University of Gothenburg, Sweden
- Lisa Beinborn, University of Amsterdam, Netherlands
- Eckhard Bick, University of Southern Denmark, Denmark
- Lars Borin, University of Gothenburg, Sweden
- António Branco, University of Lisbon, Portugal
- Jill Burstein, Educational Testing Service, USA
- Andrew Caines, University of Cambridge, UK
- Simon Dobnik, University of Gothenburg, Sweden
- Thomas François, UCLouvain, Belgium
- Johannes Graën, University of Gothenburg, Sweden
- Andrea Horbach, University of Duisburg-Essen, Germany
- Herbert Lange, University of Gothenburg and Chalmers University of Technology, Sweden
- John Lee, City University of Hong Kong, China
- Peter Ljunglöf, University of Gothenburg and Chalmers University of Technology, Sweden
- Montse Maritxalar, University of the Basque Country, Spain
- Beata Megyesi, Uppsala University, Sweden
- Detmar Meurers, University of Tübingen, Germany
- Ildikó Pilán, City University of Hong Kong, China and University of Oslo, Norway
- Martí Quixal, Universitat Oberta de Catalunya, Spain
- Robert Reynolds, Brigham Young University, USA

- Gerold Schneider, University of Zurich, Switzerland
- Irina Temnikova, Sofia University, Bulgaria
- Cornelia Tschichold, Swansea University, UK
- Francis M. Tyers, Indiana University Bloomington, USA
- Sowmya Vajjala, National Research Council Canada, Canada
- Elena Volodina, University of Gothenburg, Sweden
- Mats Wirén, Stockholm University, Sweden
- Victoria Yaneva, University of Wolverhampton, UK
- Torsten Zesch, University of Duisburg-Essen, Germany
- Robert Östling, Stockholm University, Sweden

We intend to continue this workshop series, which so far has been the only ICALL-relevant recurring event based in the Nordic countries. Our intention is to co-locate the workshop series with the two major LT events in Scandinavia, namely SLTC (the Swedish Language Technology Conference) and NoDaLiDa (Nordic Conference on Computational Linguistics), thus making this workshop an annual event. Through this workshop, we intend to profile ICALL research in Nordic countries as well as beyond, and we aim at providing a dissemination venue for researchers active in this area.

Workshop website:

https://spraakbanken.gu.se/eng/research-icall/8th-nlp4call

Workshop organizers

David Alfter[1], Elena Volodina[1], Ildikó Pilán[2], Herbert Lange[3], Lars Borin[1]

[1] Språkbanken, University of Gothenburg
[2] City University of Hong Kong and University of Oslo
[3] Department of Computer Science and Engineering, University of Gothenburg and Chalmers University of Technology

Acknowledgements

We gratefully acknowledge Riksbankens Jubileumsfond for the financial support for the two invited speakers: from the project on Development of lexical and grammatical competences in immigrant Swedish[5] (through a grant P17-0716:1) that provided funding for the invited speaker Thomas François, and from the SweLL research infrastructure project on Swedish Second language[6] (through a grant IN16-0464:1) regarding the invited speaker Egon Stemle.

[5] https://rj.se/en/anslag/2017/utveckling-av-lexikala-och-grammatiska-kompetenser-i-invandrarsvenska/
[6] https://rj.se/en/anslag/2016/swell---electronic-research-infrastructure-on-swedish-learner-language/

Contents