

# NEALT Proceedings Series Vol. 30

Proceedings of the Joint

6th Workshop on NLP for Computer Assisted Language Learning  
and  
2nd Workshop on NLP for Research on Language Acquisition

at NoDaLiDa, Gothenburg, 22nd May 2017



# NoDaLiDa

22-24 MAY 2017

GOTHENBURG SWEDEN  
CONFERENCE CENTRE WALLENBERG



Linköping Electronic Conference Proceedings 134

eISSN 1650-3740 (Online) • ISSN 1650-3686 (Print)

2017



Proceedings of the Joint

6th Workshop on NLP for Computer Assisted  
Language Learning  
and

2nd Workshop on NLP for Research on Language  
Acquisition

at NoDaLiDa 2017

Gothenburg, 22nd May 2017

edited by

Elena Volodina, Ildikó Pilán, Lars Borin,  
Gintarė Grigonytė and Kristina Nilsson Björkenstam

**Front cover photo:** Kjell Holmner/Göteborg & Co

Linköping Electronic Conference Proceedings  
eISSN 1650-3740 (Online) • ISSN 1650-3686 (Print)  
ISBN 978-91-7685-502-7

134  
2017



## Preface

For the second year in a row we have brought the two related themes of NLP for Computer-Assisted Language Learning and NLP for Language Acquisition together under one umbrella. The goal of organizing these joint workshops is to provide a meeting place for researchers working on language learning issues including both *empirical and experimental studies* and *NLP-based applications*.

The **theme 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 intersection of Natural Language Processing and Speech/Dialogue 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/Dialogue Technology, ICALL researchers need good insights into the second language acquisition (SLA) theories and practices, as well as knowledge of second language pedagogy and didactics. Hence, this workshop covers all 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 are modeled in ICALL tools.

The **workshop on Natural Language Processing (NLP) for Research in Language Acquisition (NLP4LA)** broadens the scope of this joint workshop to also include theoretical, empirical, and experimental investigation of first, second and bilingual language acquisition.

We believe that this field will benefit from collaboration between the NLP, linguistics, psychology and cognitive science communities. The workshop is targeted at anyone interested in the relevance of computational techniques for first, second and bilingual language acquisition. Therefore, our aim is to bring together researchers from different fields with a shared interest in language acquisition.

### For the two tracks we invited submissions:

- that describe research directly aimed at ICALL
- that demonstrate actual or discuss the potential use of existing Speech Technologies, NLP tools 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. collecting and annotating ICALL-relevant corpora; developing tools and algorithms for readability

---

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

Elena Volodina, Lars Borin, Ildikó Pilán, Gintarė Grigonytė and Kristina Nilsson Björkenstam 2017. Preface. *Proceedings of the Joint 6th Workshop on NLP for Computer Assisted Language Learning and 2nd Workshop on NLP for Research on Language Acquisition at NoDaLiDa 2017*. Linköping Electronic Conference Proceedings 134: i–vi.



- analysis, selecting optimal corpus examples, etc.
- that discuss challenges and/or research agenda for ICALL
- that describe empirical studies on language learner data
- that describe computational models of first, second and bilingual language acquisition
- that describe empirical or experimental studies, or computational models of various aspects of language and their effect in language comprehension and acquisition
- that demonstrate actual or discuss the potential use of Speech Technologies, NLP tools or resources for investigating language acquisition
- that describe psycholinguistic and socio-linguistic investigations on first, second and bilingual language acquisition

We have encouraged paper presentations and software demonstrations describing the above-mentioned themes for the Nordic languages; and papers that focus on different age groups, cultures, and language variation.

This year we had the pleasure to welcome invited speakers from the two research areas: Torsten Zesch (University of Duisburg-Essen) and Bente Ailin Svendsen (University of Oslo).

**Torsten Zesch** leads the Language Technology Lab<sup>1</sup> at University of Duisburg-Essen, Germany. He holds a doctoral degree in Computer Science from Technische Universität Darmstadt and has worked as a substitute professor at the German Institute for International Pedagogical Research. His research interests include the processing of non-standard, error-prone language as found in social media and learner language. He also focuses on exercise generation for computer-assisted language learning and automatic assessment of free-text answers and essays.

**In his talk** *Automatically \_\_\_\_\_ gap-fill exercise items* he gave an overview of the work done by him and his colleagues on automatically generating and scoring gap-fill exercise items. This covered early experiments on trying to find low-ambiguity contexts, the follow-up work on generating challenging distractors, and finally the recently introduced gap-fill bundles.

**Bente Ailin Svendsen** is Professor of Second Language Acquisition and Scandinavian Linguistics. She initiated and co-developed MultiLing Multilingualism in Society across the Lifespan<sup>2</sup>, a Center of Excellence funded by the Research Council of Norway (RCN), where she was the Deputy Director 2013-2015. She has carried out research on multilingual socialisation, competence and use among children and adults; and on linguistic practices and identity constructions among young people in multilingual urban spaces. Her publications include the book *Language, Youth and Identity in the 21st Century. Linguistic Practices across Urban Spaces* (co-edited with Jacomine Nortier, Cambridge UP, 2015), *Multilingual Urban Scandinavia: New Linguistic Practices* (co-edited with Pia Quist, *Multilingual Matters*, 2010), as well as articles in the *European Journal of Applied Linguistics*, *International Journal of Bilingualism* and *Nordic and Norwegian journals and books*.

1 <http://www.ltl.uni-due.de/>

2 <http://www.hf.uio.no/multiling/english/>

**In her talk** *The dynamics of citizen science in exploring language diversity* she explored the dynamics of citizen science (CS) in sociolinguistics, i.e. the involvement of non-professionals in doing sociolinguistic research, coined as Citizen Sociolinguistics (Rymes and Leone 2014, Svendsen, pending for review). In 2014, Norwegian pupils in all grades were invited to become citizen scientists through a national research campaign. Data from 4500 pupils reveal a vast linguistic diversity, an eagerness to learn languages, and a widespread use of English on a daily basis. The results, however, reflect prevailing hierarchical language regimes, firstly in the selection of specific ‘foreign’ languages offered and the desire of pupils to learn them, and secondly in the fact that the pupils’ home languages are not actively used in the classroom. In the talk, Bente argued that one of the main advantages of Citizen Sociolinguistics is its wide-reaching potential and that it represents a method suited for collecting big data sets. Secondly, it is argued, based on a media analysis of the above CS-study, that CS has a potential to increase linguistic awareness and thus stimulating linguistic stewardship. However, CS raises some challenges for sociolinguistic research, ethically, as well as ontologically and epistemologically: what do CS-data represent and what claims can be made from them? Epistemologically, with a CS methodology, we are decentralising authority on who holds legitimate knowledge about language. Citizen Sociolinguistics is about opening the dialogue between ‘the academy’ and the ‘citizens’, it stimulates public engagement and it has a potential to advance the social impact of sociolinguistics.

Rymes, Betsy and Andrea R. Leone. 2014. Citizen sociolinguistics: A new media methodology for understanding language and social life. *Working Papers in Educational Linguistics* 29(2): 25–43.

Svendsen, Bente Ailin, pending for review. The dynamics of citizen science in sociolinguistics. *Journal of Sociolinguistics*.

Svendsen, Bente Ailin, Else Ryen and Kristin Vold Lexander. 2015. *Rapport fra Forskningskampanjen 2014: Ta tempen på språket!*<sup>3</sup> [Report on the Research Campaign 2014: Taking the temperature of language!]. Oslo: Norwegian Research Council.

## Previous workshops

This workshop follows a series of workshops on NLP for CALL – and lately in combination with NLP for LA – organized by the NEALT Special Interest Group on Intelligent Computer-Assisted Language Learning (SIG-ICALL<sup>4</sup>). The workshop series has previously been financed by the Center for Language Technology<sup>5</sup> at the University of Gothenburg, and Swedish Research Council's conference grant.

Submissions to the six workshop editions have targeted a wide variety of languages, ranging from well-resourced languages (Chinese, German, English, French, Portuguese, Russian, Spanish) to less-resourced ones (Erzya, Estonian, Irish, Komi-Zyrian, Meadow Mari, Saami, Udmurt, Võro), among which several Nordic languages have been targeted: Danish, Estonian, Finnish, Icelandic, Norwegian, Saami, Swedish, and Võro.

The wide scope is also evident in the affiliations of the participating authors as shown in

---

3 [https://www.miljolare.no/innsendt/oppslag/1336/5502d9f97a260/rapport\\_fd2014.pdf](https://www.miljolare.no/innsendt/oppslag/1336/5502d9f97a260/rapport_fd2014.pdf)

4 <http://spraakbanken.gu.se/swe/forskning/ICALL/SIG-ICALL>

5 <http://clt.gu.se/>

Table 1:

Country	2012–2017
Australia	2
Belgium	4
Canada	3
Denmark	1
Estonia	3
Finland	6
France	3
Germany	42
Iceland	3
Ireland	2
Japan	2
Norway	10
Portugal	4
Russia	10
Slovakia	1
Spain	3
Sweden	51
Switzerland	8
UK	1
US	3

Table 1. Authors by affiliations, 2012-2017

So far, acceptance rate has varied between 50% and 77%, the average being 60% (see Table 2). The acceptance rate is rather high, however, the reviewing process has always been very strict with two-three double reviews per submission. This indicates that submissions to the workshops have always 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	71.5%
2017	13	7	54%

Table 2. Submissions and acceptance rates, 2012–2017



## Acknowledgements

We gratefully acknowledge the financial support from the Department of Linguistics at Stockholm University regarding the keynote speaker Bente Ailin Svendsen, and to the Department of Swedish Language at the University of Gothenburg, that provided funding for the invited speaker Torsten Zesch. We would also like to thank our Programme Committee who did a wonderful job reviewing papers:

- Lars Ahrenberg, Linköping University, Sweden
- Florencia Alam, CONICET, Argentina
- Christina Bergmann, Centre National de la Recherche Scientifique, France
- Eckhard Bick, University of Southern Denmark, Denmark
- Lars Borin, University of Gothenburg, Sweden
- António Branco, University of Lisboa, Portugal
- Jill Burstein, Educational Testing Service, USA
- Piet Desmet, KU Leuven Kulak, Belgium
- Simon Dobnik, University of Gothenburg, Sweden
- Thomas François, UCLouvain, Belgium
- Gintarė Grigonytė, Stockholm University, Sweden
- Anna Gudmundsson, Stockholm University, Sweden
- Björn Hammarberg, Stockholm University, Sweden
- Katarina Heimann Mühlenbock, DART, Sahlgrenska Universitetssjukhuset, Sweden
- Sofie Johansson, University of Gothenburg, Sweden
- Chigusa Kurumada, University of Rochester, USA
- Peter Ljunglöf, Chalmers University of Technology, Sweden
- Ellen Marklund, Stockholm University, Sweden
- Detmar Meurers, University of Tübingen, Germany
- Kristina Nilsson Björkenstam, Stockholm University, Sweden
- John K. Pate, The University at Buffalo, USA
- Martí Quixal, Univeritat Oberta de Catalunya, Spain
- Lena Renner, Stockholm University, Sweden
- Gerold Schneider, University of Konstanz, Germany
- Mathias Schulze, University of Waterloo, Canada
- Philip Shaw, Stockholm University, Sweden
- Jennifer Spenader, University of Groningen, Netherlands
- Sofia Strömbergsson, Karolinska Institutet, Sweden
- Joel Tetreault, Yahoo! Labs, USA
- Cornelia Tschichold, Swansea University, UK
- Francis Tyers, The Arctic University of Norway, Norway
- Sowmya Vajjala, Iowa State University, US
- Paul Vogt, Tilburg University, Netherlands
- Elena Volodina, University of Gothenburg, Sweden
- 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- and LA-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, SLTC (the Swedish Language Technology Conference) and Nodalida, thus making this workshop an annual event. Through this workshop, we intend to profile ICALL and LA research in Nordic countries and beyond, and to provide a dissemination venue for researchers active in this area.

Workshop website:

<https://spraakbanken.gu.se/eng/icall/joint6thNLP4CALL-2ndNLP4LA2017>

*Workshop organizers*

*Elena Volodina, Ildikó Pilán, Lars Borin (University of Gothenburg)*

*Gintarė Grigonytė, Kristina Nilsson Björkenstam (University of Stockholm)*

## Contents

Preface	i
<i>Elena Volodina, Lars Borin, Ildikó Pilán, Gintarė Grigonytė and Kristina Nilsson Björkenstam</i>	
Learning with learner corpora: Using the TLE for native language identification	1
<i>Allison Adams and Sara Stymne</i>	
Challenging learners in their individual zone of proximal development using pedagogic developmental benchmarks of syntactic complexity	8
<i>Xiaobin Chen and Detmar Meurers</i>	
Crossing the border twice: Reimporting prepositions to alleviate L1-specific transfer errors	18
<i>Johannes Graën and Gerold Schneider</i>	
Revita: a system for language learning and supporting endangered languages	27
<i>Anisia Katinskaia, Javad Nouri and Roman Yangarber</i>	
Developing a web-based workbook for English supporting the interaction of students and teachers	36
<i>Björn Rudzewitz, Ramon Ziai, Kordula De Kuthy and Detmar Meurers</i>	
Annotating errors in student texts: First experiences and experiments	47
<i>Sara Stymne, Eva Pettersson, Beáta Megyesi and Anne Palmér</i>	
Building and using language resources and infrastructure to develop e-learning programs for a minority language	61
<i>Heli Uibo, Jack Rueter and Sulev Iva</i>	