Selected papers from the
CLARIN Annual Conference 2016

Aix-en-Provence, 26–28 October 2016

Linköping Electronic Conference Proceedings 136
eISSN 1650-3740 (Online) • ISSN 1650-3686 (Print) 2017
ISBN 978-91-7685-499-0

Selected papers from the

CLARIN Annual Conference 2016

Aix-en-Provence, 26–28 October 2016

edited by Lars Borin

Front cover illustration:
Group photo by Quirijn Backx, picture composition by Karolina Badzmierowska • CLARIN ERIC

Linköping Electronic Conference Proceedings 136
eISSN 1650-3740 (Online) • ISSN 1650-3686 (Print) 2017
ISBN 978-91-7685-499-0

Preface

These proceedings present the highlights of the CLARIN Annual Conference 2016 that took place in Aix-en-Provence, France. As the third volume in the proceedings series it illustrates that CLARIN has developed into a community with an ambition that goes beyond the realization of a research infrastructure for language resources. The multiannual record of CLARIN's progression also demonstrates the ambition of coupling a data infrastructure to a knowledge sharing infrastructure: a common platform that will guide researchers from the Humanities and Social Sciences in making optimal use of the infrastructure and benefitting from the experience of other researchers and the best practices applied across the case studies collected.

The papers selected for this volume present the results of a number of projects conducted within and between CLARIN's national consortia, but the proceedings of 2016 also contain papers with contributions from authors outside the CLARIN consortium. Furthermore, the fact that France generously hosted the conference already in the year before it actually joined CLARIN as an observer highlights the potential for growth and for pan-European collaboration.

CLARIN provides sustainable access to language resources in all forms, analysis services for the processing of language materials, and a platform that can stimulate the use, reuse and repurposing of the available data. This contributes to realizing the vision associated with the Open Science agenda and to strengthening Europe's capacity to lower the barriers for researchers to entry digital scholarship and cutting edge research. As shown by the range of topics addressed in the proceedings, language resources can play a multitude of roles, including carrier of information, record of the past, means of literary expression, social signal, or object of linguistic study.

Due to the diversity of the data types supported, the communities of use to be served by CLARIN are also diverse. Combined with the multitude of languages covered, CLARIN can help to realize a multilingual European Research Area for digital research in the Humanities and Social Sciences, to turn Europe's multilingualism into a basis for the comparative investigation of a wide range of intellectual and societal phenomena and to ensure that the multidisciplinary research agendas addressing societal challenges will have impact.

The CLARIN Annual Conference is one of the communication instruments between those who build and maintain the infrastructure, those who provide data and tools, and those who use the CLARIN infrastructure in their scholarly projects. A similar role is played by the workshops focused on specific data types organized in 2016 and 2017. All these CLARIN events have demonstrated the importance of coordination in sharing the insights into problems, solutions, failures and successes across national and linguistic borders. Hopefully this volume will help to attract new categories of scholars, with ideas and requirements for use cases that can help us identify the directions and next steps to take in the further development of the CLARIN infrastructure as a pillar of Europe's Open Science policies.

Utrecht, 14 May 2017

Franciska de Jong
Executive Director CLARIN ERIC

Franciska de Jong 2017. Preface. *Selected papers from the CLARIN Annual Conference 2016*. Linköping Electronic Conference Proceedings 136: i–i.

Introduction

This volume contains a selection of papers presented at the CLARIN Annual Conference 2016 which was held in Aix-en-Provence, France, on 26–28 October 2016.

This was the fifth edition of the conference. It started in 2012 as an internal event, where members of the national CLARIN consortia came together to share their experiences of and thoughts on the development of the CLARIN ERIC infrastructure.

In 2014, it was felt that the time was ripe to change the format of the conference into an event with an open call for contributions, in order to include also the humanities and social-science research communities – the intended users of the infrastructure – in the exchange of ideas and experiences on the CLARIN infrastructure. This includes its design, construction and operation, the data and services that it contains or should contain, its actual use by researchers, its relation to other infrastructures and projects, and the CLARIN Knowledge Sharing Infrastructure.

As a result of the 2016 call for papers we received 34 anonymous extended abstracts, each of which was anonymously reviewed by at least three members of the program committee, which as always consisted of the members of the CLARIN ERIC National Coordinators' Forum, i.e., one member from each participating country or NGO. In order to avoid conflicts of interest, no PC member reviewed submissions from their own country. As a result of the reviewing process, a total of 25 submissions were accepted for presentation at the conference, 14 as oral presentations and 11 as posters.

In addition to the submitted presentations, the conference featured two invited speakers. The keynote on the first day was presented by professor Ian Gregory from Lancaster University, under the title *Texts, language and geography: Understanding literature using geographical text analysis*, and on the second day, professor Sally Wyatt, Maastricht University talked about *Why technologies are not neutral, and why it matters for linguists*.

As a new feature, the CLARIN 2016 call for papers included a call for submissions to a thematic session, focusing on *Language resources and historical sources*. The general area of interest for the thematic session was stated in the call for papers as CLARIN-related research in the historical sciences, understood in a wide sense to encompass fields such as History, "History of ..."/"... history" (e.g., History of science, Rhetorical history), as well as the various historically oriented subfields of linguistics (e.g., Historical linguistics, Historical pragmatics, etc.), and philology. We invited submissions on two separate but overlapping aspects that we construed this theme to encompass:

(1) The historical aspect in a narrower sense: Processing historical language stages in the form of text or speech, with the concomitant issues of digitization, non-standardized language, etc.

(2) The diachronic aspect: Discovering, characterizing and tracking change through time, both linguistic changes and changes in the world as reflected in the content of text.

Two of the oral presentations and several posters addressed this theme. Ian Gregory's keynote speech together with the two oral presentations were organized into a thematic session scheduled at the very beginning of the conference program.

The conference was video recorded; see the YouTube playlist:
https://www.youtube.com/playlist?list=PLlKmS5dTMgw2pP-uvhKNVSgOuuZjvmLwy

Following the conference, authors of the accepted papers were invited to submit full versions of their papers to be considered for the conference proceedings volume,

Lars Borin 2017. Introduction. *Selected papers from the CLARIN Annual Conference 2016*. Linköping Electronic Conference Proceedings 136: iii–iv.

although this time the submissions were not anonymous. Again the papers were reviewed (anonymously) by two to four PC members, at least one of which had not reviewed the original abstract submitted for the conference. We received 14 full-length submissions, out of which 10 were accepted for this volume. Most of these address core CLARIN issues dealing with the construction, maintenance and use of the European infrastructure coordinated in the framework of the CLARIN ERIC, such as search engine design, resource discovery, metadata quality, researcher training in infrastructure use, and design of specific tools and resources. There is one "pure" research paper in this volume – by Hinrichs, Erdmann and Joseph – but many of the contributions refer to research conducted using the CLARIN infrastructure. In two cases – the papers by Beißwenger et al. and by MacWhinney – the focus is on resource-building with specific research questions or a specific research field in mind, where the research and infrastructure-building activities feed into each other and actually become hard to disentangle.

I would like to thank the reviewers for the dedicated efforts they put down in evaluating the submissions, and also Peter Berkesand at Linköping University Electronic Press, who (as usual) has ensured that the digital publication of this volume went smoothly and painlessly.

Lars Borin
University of Gothenburg
Program committee chair

Program committee for the CLARIN Annual Conference 2016

Members of the CLARIN ERIC National Coordinators' Forum

Jan Theo Bakker
Lars Borin
António Branco
Koenraad De Smedt
Tomaž Erjavec
Eva Hajičová
Erhard Hinrichs
Krister Lindén
Bente Maegaard
Monica Monachini
Karlheinz Mörth
Jan Odijk
Maciej Piasecki
Stelios Piperidis
Kiril Simov
Inguna Skadiņa
Jurgita Vaičenonienė
Kadri Vider
Martin Wynne

Additional reviewer

Paul Meurer

Contents