





Selected papers from the  
CLARIN Annual Conference 2017  
Budapest, 18–20 September 2017

edited by Maciej Piasecki



**Front cover illustration:**

picture composition by Marcin Oleksy • CLARIN ERIC  
Licensed under Creative Commons Attribution 4.0 International:  
<https://creativecommons.org/licenses/by/4.0/>

Linköping Electronic Conference Proceedings  
eISSN 1650-3740 (Online) • ISSN 1650-3686 (Print)  
ISBN 978-91-7685-273-6

147  
2017



# Introduction

**Franciska de Jong**  
Executive Director CLARIN ERIC  
Universiteit Utrecht  
f.m.g.dejong@uu.nl

**Maciej Piasecki**  
CLARIN-PL  
and Faculty of Computer Science  
and Management  
Wrocław University of Science and Technology  
maciej.piasecki@pwr.edu.pl

This volume includes extended versions of the selected papers presented at the CLARIN Annual Conference 2017 held in Budapest, Hungary, on 18th–20th September 2017.

CLARIN ERIC (<http://clarin.eu>) is the European Research Infrastructure for Language Resources and Technology aimed at supporting researchers mostly from the domain of Social Sciences and Humanities (SS&H) in their use of language data and technologies. CLARIN works towards lowering barriers in doing research in those areas by offering widespread, advanced, user-friendly and effective applications giving access to language resources and enabling the analysis of textual data, speech recordings, as well as multimodal data in different research tasks.

The annual conference is a fruitful combination of an internal event and a venue open to a general community of researchers. On the one hand, the conference is the annual plenary meeting for the CLARIN consortia from all the participating countries. It is attended by selected delegates of the national consortia that are involved in building, maintaining and exploiting the infrastructure. Since the establishment of the ERIC in 2012, CLARIN has considerably grown in size. There are 20 member countries and more than 100 associated research institutions. As a consequence, only the subset of a large CLARIN community can participate in the annual conference. At the same time, the annual event is open for various communities of users – researchers from the SS&H domains, i.e. the people who are the *raison d’être* for CLARIN.

The topics that are in the focus of the CLARIN Annual Conference can be divided into five main areas:

- operation and use of the CLARIN infrastructure,
- aspects of its design and construction,
- knowledge sharing about the infrastructure and its use,
- relations with other infrastructures and projects,
- and, the most important, its applications in research in SS&H.

The conference hosted two invited talks given by renowned researchers from the field of Humanities. Professor Karina van Dalen-Oskam from University of Amsterdam / Huygens ING talked about “Literary translations and tools for stylometric research” and Professor Piek Vossen from Vrije Universiteit Amsterdam presented an overview of applications of multilingual language technology: “From multilingual to cross-lingual processing for Social Sciences and Humanities”.

Since 2015, the CLARIN annual conference has put a specific topic in focus. This topic is highlighted by organising a special thematic session. The theme chosen for 2017 edition was: “Multilingual Processing for Social Sciences and Humanities”.

The contributions were solicited through an open call. 37 submissions were registered in the reviewing system (provided by EasyChair, <http://easychair.org>) in the form of extended

---

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

abstracts. Each submission was reviewed by at least three reviewers, all of whom scientists with rich experience and significant achievements in Language Technology and/or Digital Humanities. 24 submissions were accepted as a result of the two reviewing and correction rounds: 13 for oral presentations, the remaining ones as posters. The selected contributions were published online in the Book of Abstracts:

<https://www.clarin.eu/content/abstracts-overview-clarin-annual-conference-2017>.

After the conference, the authors of all accepted presentations were invited to prepare significantly expanded versions of their abstracts as full paper proposals. The authors were strongly encouraged to take into account the outcome of the discussion of their presentations during the conference, as well as their further work performed after conference. 12 papers were submitted and reviewed by the Program Committee. Each paper was reviewed by two reviewers, during two iterations. The process was completed by a final review with the aim to check if all the requested changes had been introduced. On the basis of the improved versions sent after the second reviewing iteration, all 12 papers were accepted and included in this volume.

The accepted papers represent an overview of the most important topics discussed during the conference. The first two papers (Nicolas et al.) and (Pariſse, et al.) report on the process of the formation of national consortia and their initial development.

Next, Durco et al. revisit an important problem of constructing mapping between metadata formats from the point of view of building links between the CLARIN CMDI metadata standard and the metadata system in the PARTHENOS project. One of the goals of this project is to build a platform linking several major research infrastructures related to the area of SS&H. Metadata mapping is also close to a very challenging issue of the curation of metadata content.

Zinn presents further development of the Switchboard systems for effortless and automated linking of resources with language tools, as well as processing chains on the basis of resources content. A new mechanism of directly linking EUDAT's B2DROP cloud service to Switchboard is the main focus of the paper.

The next papers, namely of Sugimoto and Monachini et al., are about a problem of great importance for CLARIN (and every research infrastructure), i.e. studying and discovering users' needs. Sugimoto analyses users' behaviours through the observation of their activities in the already existing infrastructure and its systems. Monachini et al. report on an interesting survey-based study among the researchers in Classics Studies on the actual and intended use of digital methods. The results and conclusions are also illustrated by a mock-up prototype of a future system.

The papers of Fiſer & Lenardiĉ and Borin et al. are devoted to the enrichment of language resources in CLARIN on the basis of the already existing resources in a way focused on the identified users' needs. Fiſer & Lenardiĉ report on the results of an overview of the state of affairs in the area of language resources based on parliamentary records. The survey was supported by a dedicated CLARIN workshop and brought about an idea of a kind of virtual collection of such resources. Borin et al. describe the outcome of the first phase of the project on turning the linguistic material available in Grierson's classical Linguistic Survey of India (LSI) into a digital language resource, which will be managed and offered via CLARIN.

Next, we enter the domain of IPR (Intellectual Property Rights) issues which is often a stumbling block in SS&H, especially in relation to the use of corpora. Kelli et al. discuss the paradigm of Open Science from the point of view of its implementation in CLARIN. Calamai et al. present a very useful study on the management of IPR in the area of digitised analogue-born speech corpora, in which informants, record-makers, corpus creators and researchers responsible for digitisation, annotation and corpus publishing form a surprisingly complicated picture. However, some practical conclusions were drawn.

Finally, the volume is completed by two examples from the top CLARIN layer, i.e. the layer of research applications in which CLARIN tries to fulfil its strategy of lowering the technological and knowledge barriers, i.e. research application aimed at making users free from the necessity

of installing and setting-up software, as well as decreasing the required amount of technological knowledge possessed by users (e.g. from the area of language engineering). Maryl et al. describe a web-based application called LEM (Literary Exploration Machine) that offers several functions for the extraction of various statistics from text related to linguistic features of a text. It works according to a very simple scheme: upload the data and select one of several options with just one click. LEM was built in close co-operation with users and its every function is a response to a demand of a concrete user – a researcher. Piasecki et al. present a new multilingual version of WebSty – an open, web-based stylometric system, offering advanced, efficient language processing and a rich functionality for statistical processing. However, at the same time, WebSty allows for performing a stylometric analysis by simply selecting one of the few predefined ‘express’ options.

In addition, CLARIN published a rich set of materials related to the conference on the web:

1. the detailed list of topics, to be found in the call for papers:  
<https://www.clarin.eu/news/call-papers-clarin-annual-conference-2017>
2. the complete conference program and most of the slides presented:  
<https://www.clarin.eu/content/programme-clarin-annual-conference-2017>
3. the recordings of most talks, the two invited lectures and several other video materials are available on a dedicated channel of *VideoLectures*:  
[http://videolectures.net/clarinannualconference2017\\_budapest/](http://videolectures.net/clarinannualconference2017_budapest/).

### **Programme Committee for the CLARIN Annual Conference 2017**

- Catia Cucchiari, Dutch Language Union, The Netherlands/Flanders
- Lars Borin, University of Gothenburg, Sweden
- António Branco, University of Lisbon, Portugal
- Koenraad De Smedt, University of Bergen, Norway
- Tomaž Erjavec, Jožef Stefan Institute, Slovenia
- Eva Hajičová, Charles University Prague, Czech Republic
- Erhard Hinrichs, University of Tübingen, Germany
- Nicolas Larrousse, Huma-Num, France
- Krister Lindén, University of Helsinki, Finland
- Bente Maegaard, University of Copenhagen, Denmark
- Monica Monachini, Institute for Computational Linguistics «A. Zampolli», Italy
- Karlheinz Mörth, Austrian Academy of Sciences, Austria
- Jan Odiijk, Utrecht University, the Netherlands
- Maciej Piasecki, Wrocław University of Science and Technology, Poland (chair)
- Stelios Piperidis, ILSP, Athena Research Center, Greece
- Kiril Simov, IICT, Bulgarian Academy of Sciences, Bulgaria
- Inguna Skadiņa, University of Latvia, Latvia
- Jurgita Vaičėnienė, Vytautas Magnus University, Lithuania

- Tamás Váradi, Research Institute for Linguistics, Hungarian Academy of Sciences
- Kadri Vider, University of Tartu, Estonia
- Martin Wynne, University of Oxford, United Kingdom

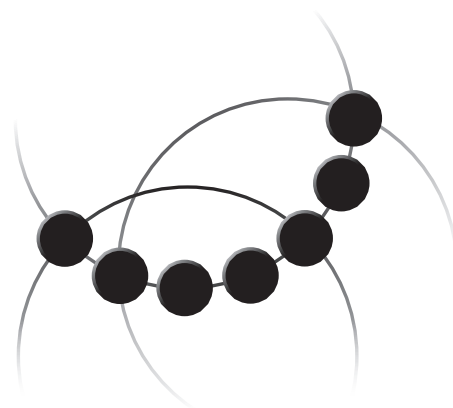
## Contents

Introduction	i
<i>Franciska de Jong and Maciej Piasecki</i>	
CLARIN-IT: State of Affairs, Challenges and Opportunities	1
<i>Lionel Nicolas, Alexander König, Monica Monachini, Riccardo Del Gratta, Silvia Calamai, Andrea Abel, Alessandro Enea, Francesca Biliotti, Valeria Quochi and Francesco Vincenzo Stella</i>	
CORLI: A linguistic consortium for corpus, language, and interaction	15
<i>Christophe Parisse, Céline Poudat, Ciara R. Wigham, Michel Jacobson and Loïc Liégeois</i>	
Something will be connected - Semantic mapping from CMDI to Parthenos Entities	25
<i>Matej Ďurčo, Matteo Lorenzini and Go Sugimoto</i>	
A Bridge from EUDAT's B2DROP cloud service to CLARIN's Language Resource Switchboard	36
<i>Claus Zinn</i>	
Examining Web User Flows and Behaviours in CLARIN Ecosystem	46
<i>Go Sugimoto</i>	
Digital Classics and CLARIN-IT: What Italian Scholars of Ancient Greek Expect from Digital Resources and Technology	61
<i>Monica Monachini, Anika Nicolosi, Alberto Stefanini</i>	
Parliamentary Corpora in the CLARIN infrastructure	75
<i>Darja Fišer and Jakob Lenardič</i>	
Many a Little Makes a Mickle – Infrastructure Component Reuse for a Massively Multilingual Linguistic Study	86
<i>Lars Borin, Shafqat Mumtaz Virk and Anju Saxena</i>	
Implementation of an Open Science Policy in the context of management of CLARIN102 language resources: a need for changes?	
<i>Aleksei Kelli, Krister Lindén, Kadri Vider, Penny Labropoulou, Erik Ketzan, Pawel Kamocki and Pavel Straňák</i>	
Authorship and copyright ownership in the digital oral archives domain: The Gra.fo digital archive in the CLARIN-IT repository	112
<i>Silvia Calamai, Chiara Kolletzek, Aleksei Kelli and Francesca Biliotti</i>	
Literary Exploration Machine A Web-Based Application for Textual Scholars	128
<i>Maciej Maryl, Maciej Piasecki and Tomasz Walkowiak</i>	
Open Stylometric System WebSty: Towards Multilingual and Multipurpose Workbench	145

*Maciej Piasecki, Tomasz Walkowiak and Maciej Eder*



# CLARIN



Linköping Electronic Conference Proceedings  
eISSN 1650-3740 (Online) • ISSN 1650-3686 (Print)  
ISBN 978-91-7685-273-6

147  
2017

**Front cover illustration:**  
picture composition by Marcin Oleksy • CLARIN ERIC

Licensed under Creative Commons Attribution 4.0 International: <https://creativecommons.org/licenses/by/4.0/>