



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



Linköping Electronic Conference Proceedings 147
eISSN 1650-3740 (Online) • ISSN 1650-3686 (Print) 2017
ISBN 978-91-7685-273-6

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/>

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 (Pariśe, 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 Fišer & Lenardič 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. Fišer & Lenardič 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/.

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 Odijk, 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 CLARIN language resources: a need for changes?	102
<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-IT: State of Affairs, Challenges and Opportunities

Lionel Nicolas

Eurac Research, Bolzano, Italy
lionel.nicolas@eurac.edu

Alexander König

Eurac Research, Bolzano, Italy
alexander.koenig@eurac.edu

Monica Monachini

ILC "A. Zampolli"
CNR, Pisa, Italy
monica.monachini@ilc.cnr.it

Riccardo Del Gratta

ILC "A. Zampolli"
CNR, Pisa, Italy
riccardo.delgratta@ilc.cnr.it

Silvia Calamai

Università di Siena, Italy
silvia.calamai@unisi.it

Andrea Abel

Eurac Research, Bolzano, Italy
andrea.abel@eurac.edu

Alessandro Enea

ILC "A. Zampolli"
CNR, Pisa, Italy
alessandro.enea@ilc.cnr.it

Francesca Biliotti

Università di Siena, Italy
francesca.biliotti@unisi.it

Valeria Quochi

ILC "A. Zampolli"
CNR, Pisa, Italy
valeria.quochi@ilc.cnr.it

Francesco Vincenzo Stella

Università di Siena, Italy
francesco.stella@unisi.it

Abstract

This paper gives an overview on the Italian national CLARIN consortium as it currently stands two years after its creation at the end of 2015. It thus discusses the current state of affairs of the consortium on several aspects, especially with regards to members. It also discusses the events and initiatives that have been undertaken, as well as the ones that are planned in the close future. It finally outlines the conclusions of a user survey performed to understand the expectations of a targeted user population and provides indications regarding the next steps planned.

1 Introduction

Among the research fields of interest for the CLARIN initiative as a whole, several have a long history of research efforts performed in Italy over the past decades and have identifiable associations organizing recurrent Italian events. For example, for Computational Linguistics and Language Technology Applications - particularly for Italian - there is the *Associazione Italiana di Linguistica Computazionale*¹ (AILC) that organizes, among other events, the yearly celebrated conference CLIC-IT² and the periodically held evaluation campaign of Natural Language Processing - Evalita³; for Speech Sciences there is the *Associazione Italiana di Scienze della Voce*⁴ (AISV), that also organizes a yearly celebrated conference, together with the Franco Ferrero prize, and for Digital Humanities there is the *Associazione per l'Informatica Umanistica e le Culture Digitali*⁵ (AIUCD) that also organizes, among other events, a yearly

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

¹<http://www.ai-lc.it/>

²<http://www.ai-lc.it/en/initiatives/clic-it/>

³<http://www.ai-lc.it/en/initiatives/evalita>

⁴<https://www.aisv.it/>

⁵<http://www.aiucd.it>

celebrated AIUCD conference⁶. Accordingly, it is only natural that ever since CLARIN started in 2008 with a preparatory phase, it has always been of great interest for several Italian institutions⁷.

When CLARIN ERIC was established in 2012 after the end of the preparatory phase in 2011, several efforts have been made to create a national consortium. On October 2015, the *Ministero dell'Istruzione, dell'Università e della Ricerca* (MIUR) signed the Memorandum of Understanding to become a full member and Italy finally joined CLARIN ERIC with the *Department of Social Sciences and Humanities* (DSU) of the *Consiglio Nazionale delle Ricerche* appointed as Representing Entity, the *Istituto di Linguistica Computazionale* (ILC) nominated as leading Italian participant and Monica Monachini nominated as National Coordinator.

This paper aims at providing a clear overview of the Italian national CLARIN consortium as it currently stands two years after its creation. In Section 2, it discusses the current state of affairs of the consortium, be it in terms of members, funding, technical infrastructure or role within the CLARIN federation. Section 3 then provides a number of information on the CLARIN-IT members, especially with regards to what they offer to CLARIN in terms of resources, services and expertise, and what CLARIN offers them to further their own research. Section 4 discusses the CLARIN-IT events organized in Italy so far and the events planned in the close future, especially with regards to the organization of the CLARIN 2018 conference. Finally, Section 5 outlines the conclusions of a user survey performed to understand the expectations of a targeted Italian user population while Section 6 provides indications regarding the next steps planned for the consortium as a whole and for each member individually.

2 Current State of Affairs

2.1 Members

As it stands at the moment, the CLARIN-IT consortium includes four institutions as full members with two other institutions in the process of formally joining it. The four current full members are:

1. the *Istituto di Linguistica Computazionale "A. Zampolli"* (ILC) of the *Consiglio Nazionale delle Ricerche* in Pisa,
2. the *Institute for Applied Linguistics* (IAL) of *Eurac Research* in Bolzano,
3. the *Dipartimento di Scienze della Formazione, Scienze Umane e della Comunicazione Interculturale* (DSFUCI) of the *Università di Siena*,
4. the *Dipartimento di Filologia e critica delle Letterature antiche e moderne* (DFCLAM) of the *Università di Siena*.

The two other institutions in the process of formally joining the consortium are the *Dipartimento di Discipline Umanistiche* of the *Università di Parma* and the *Dipartimento di Studi Umanistici* of the *Università "Ca' Foscari"* in Venezia.

Aside from these six institutions, a noticeable number of other Italian institutions from a wide range of disciplines have expressed their interest in participating. Among those, we can cite the *Fondazione Bruno Kessler* (Trento) the *Università Cattolica del Sacro Cuore* (Milano), the *Università "Tor Vergata"* (Roma), and the *Università di Pisa, Dipartimento di Linguistica* (Pisa).

2.2 Funding

One reason for the late arrival and the limited number of members (when compared to other CLARIN consortia) is due to the fact that negotiations regarding an Italian national funding of the CLARIN-IT consortium with the Research Ministry are still ongoing. Consequently, while other institutions have put on hold their membership until a viable context for their participation can be arranged, the current

⁶<http://www.aiucd.it/convegno-annuale/>

⁷One of them, the *Institute for Computational Linguistics "Antonio Zampolli"* (ILC) of the *Italian National Research Council*, was already a member of the consortium that carried out the preparatory phase under the FP7-INFRASTRUCTURES EC programme (GA:2007-212230).

members have either sought funding for personnel at regional or local level, have committed some of their own internal resources or are contributing on a purely voluntary basis.

2.3 CLARIN-IT committees

Following the best practices implemented within CLARIN ERIC and the CLARIN federation, CLARIN-IT has established the following committees: the *technical committee*, the *metadata and standards committee*, the *legal issues committee* and the *committee for the relations with users*.

The *technical committee* coordinates all CLARIN-IT type C and B centres and ensures the smooth functioning of all the technical services. It will be responsible of ensuring conformance to CLARIN ERIC technical requirements and of the prompt uptake of technological upgrades and new solutions developed and/or suggested by CLARIN ERIC. It also advises the National Coordinator on any critical issue regarding the quality of the services provided and on the possible measures to take.

The *metadata and standards committee* is responsible for the adoption of the metadata and data format standards supported by CLARIN ERIC. As such it selects and disseminates the existing supported standards relevant for its user communities; helps adapting the standards to the specific needs of the users and members, and contributes to the definition of metadata and concepts in the CLARIN Concept Registry, when needed. It also gives advice to the National Coordinator in matters of standards.

The *legal issues committee* deals with IPR and privacy protection issues. Its main task is to revise and adapt the policies and licenses devised and recommended by CLARIN ERIC to the needs of CLARIN-IT. The committee also helps and advises members on IPR critical matters about specific data resources, with the aim of maximising research data sharing within the CLARIN community.

The *committee for the relations with users* discusses and coordinates the national activities towards an active engagement of user communities. Its main responsibilities are to adapt and implement the guidelines and best practices promoted by CLARIN ERIC within CLARIN-IT, discuss new ideas for involving new users and research communities, receive feedback from the users, disseminate information about services, resources, projects and all relevant CLARIN-like activities in Italy and beyond.

In addition to those, CLARIN-IT is also creating an *advisory board* that will provide strategic advice on various matters such as, among others, quality, new initiatives or synergies with international and national related infrastructures and projects. The *advisory board* will be formed by high profile scholars that are not directly involved in CLARIN-IT activities, but who have access to relevant networks in the Social Sciences and Humanities (SSH).

2.4 Technical infrastructure

As regards the CLARIN centres and resources made available, the ILC currently hosts the ILC4CLARIN, CLARIN Type C Centre, and is in the active process of achieving a CLARIN-B certification⁸. ILC4CLARIN is the first CLARIN-IT technological node that links the Italian SSH community to the EU-wide CLARIN communities; it has set up a CLARIN DSpace repository which will soon offer deposit services to the Italian community. The IAL has successfully created its own CLARIN DSpace repository and is progressively making its resources available on it. It is also aiming at achieving a CLARIN C status as soon as possible. The DSFUCI and DFCLAM are actively working on making their resources available via the ILC4CLARIN repository.

On a different technical perspective, the CLARIN-IT consortium closely cooperates with the *Consortium GARR*, the Italian University and Research Network, in particular with the IDEM-GARR⁹ office that supports federated authentication in CLARIN. Thus, any member or participant of the IDEM-GARR federation already has access to services hosted at any CLARIN centre in Europe via their institutional credentials. The CLARIN-IT consortium is also in contact with the CLOUD-GARR¹⁰ office so as to allow members to safely and securely deposit data in the cloud.

⁸<https://www.clarin.eu/content/centre-requirements-revised-version>

⁹<https://www.idem.garr.it/en>

¹⁰<https://cloud.garr.it>

2.5 CLARIN-IT within the CLARIN federation

Regarding the participation in CLARIN events, CLARIN-IT members participated in the CLARIN Annual General Assembly 2016 and 2017, in the CLARIN Annual Conference 2015, 2016 and 2017 and in the CLARIN Centre meeting in 2016 and 2017. CLARIN-IT members also participated in the first, second and third CLARIN-PLUS Workshop on Oral History (Oxford, Utrecht, Arezzo)¹¹, the CLARIN-PLUS Workshop on User Involvement, the CLARIN-PLUS Workshop on Digital Collections of Newspapers, the CLARIN-PLUS Workshop "Sustainability and Governance", and in the CLARIN Workshop on "Interoperability of L2 resources and tools". Finally, CLARIN-IT was represented at the CLARIN booth at LREC 2016 and at the PARTHENOS WP3 Meeting (an initiative of which CLARIN is member of) in November 2016.

3 CLARIN-IT centres

A large networking initiative such as CLARIN allows institutions with their own agendas to devise efficient roadmaps to approach their common or inter-related challenges and achieve several added values such as, among many others, preventing the duplication of efforts, the sharing of resources or the creation of new initiatives resulting from productive encounters. Also, a common added value brought to all CLARIN-IT members comes from the opportunities in terms of sustainability, be it through the CLARIN-supported standards and tools or through the interaction with expert fellow stakeholders. More specifically, we can outline the following synergies between the overall CLARIN initiative and the CLARIN-IT centres.

3.1 Synergies between the ILC and CLARIN

3.1.1 The ILC in few words

The *Institute for Computational Linguistics "A. Zampolli"* is a reference centre in the field of Computational Linguistics at both national and international levels. Its various research lines (Digital Humanities, Representation Standards, Distributed Research Infrastructures and Knowledge Management) makes the ILC a unique institution. The Institute is part of the *Department of Social Science and Humanities, Cultural Heritage* (DSU) of the *Consiglio Nazionale delle Ricerche* (CNR). It was already an active participant in the CLARIN preparatory phase.

3.1.2 The ILC as an asset for CLARIN

ILC has for many years been active in the field of language resources and technologies for natural language processing. The group of Language Resource and Infrastructures¹² has been paying attention to the development of digital resources (corpora, computational lexicons) for Italian and English and is now creating new lexical resources for Greek and Latin according to the Linked Open Data (LOD) paradigm. ILC recognizes, indeed, that there is still a lack of lexical resources dealing with 'historical' languages, such as ancient Greek, Latin or Sanskrit, and this can be seen as a missed opportunity for the DH community. ILC is thus making available legacy, digitalized, print resources as LOD, as well as creating new resources by linking existing ones and distributing them with standard methods such as SPARQL end points and/or HTML browsing. ILC is an active member and covers leading roles within the ISO Committee TC/37 SC4, as well as in the W3C OntoLex working group, thus facilitating both the liaison and the coordination between CLARIN ERIC and the ISO Standard Committees. ILC is also involved in developing methods and digital technology for preservation of textual archives. Experts are dealing with text encoding and mark-up to provide the scientific community with digital data access, exchange and research on textual heritage of the literature held by ILC. ILC has set up a CLARIN C-Centre (aiming for type B certification in 2018), ILC4CLARIN¹³, along with a CLARIN DSpace repository, where the above-mentioned language resources are deposited and/or described according to the CMDI model

¹¹<http://oralhistory.eu/workshops>

¹²<http://lari.ilc.cnr.it>

¹³<https://ilc4clarin.ilc.cnr.it/en/>

(Broeder et al., 2012), which make them also visible and retrievable in the CLARIN Virtual Language Observatory¹⁴ (VLO) (Van Uytvanck et al., 2010; Goosen and Eckart, 2014).

Along with the digital resources made available through the repository, ILC4CLARIN provides a set of linguistic services¹⁵ such as systems for querying text corpora, natural language analysis and annotation tools, tools for extraction and acquisition of linguistic information, format converters and tools for lexicon creation or manipulation. Many of these tools are offered in the form of webservices; some of them are already available in Weblicht¹⁶ (Hinrichs et al., 2010) and the Language Resource Switchboard¹⁷ (Zinn, 2016) or are currently being integrated there (see Section 6 for the next steps).

Through its repository, ILC4CLARIN also makes available both web applications and lexical resources for Latin and Greek. The web application for lemmatizing short Latin texts¹⁸ offers also a REST web service¹⁹ which outputs the results of the lemmatization process in JSON; a search interface is available for browsing several wordnets in different languages including Italian, Ancient Greek, Latin, Croatian, and Arabic.²⁰ Together with these web applications, a revised portion of the Ancient Greek WordNet is also available.²¹

3.1.3 CLARIN as an asset for the ILC

Participating in CLARIN provides a number of opportunities in terms of sustainability, preservation, persistent identification, and visibility for the ILC's research outputs. Sustainability is a key aspect for the ILC's strategy as it kept on growing and conducting research over the years; preservation and persistent identification of research data and results is fundamental as well, since they provide to users and researchers the technologies to retrieve data and replicate experiments. CLARIN offers ILC frameworks and platforms where to promote and support the use of technology and text analysis tools. For example, Weblicht²² allows to combine web services so as to handle and exploit textual data. Finally, the VLO makes the resources produced and described in the ILC centre available to a wider audience in the DH community while the CMDI model ensures a high quality in terms of metadata.

3.2 Synergies between the IAL and CLARIN

3.2.1 The IAL in few words

The *Institute for Applied Linguistics*²³ (IAL) is part of *Eurac Research*, a private non-profit research centre located in Bolzano and composed of several research groups focussing their efforts on research subjects of particular importance for the South Tyrolean region where it is situated. The IAL in particular aims at addressing current issues of language and education policy as well as economic and social questions at the local and international level. It is an international research environment where around 25 Junior and Senior Researchers with heterogeneous backgrounds are performing research on a wide range of language-related subjects.

3.2.2 The IAL as an asset for CLARIN

With a majority of its workforce dedicated to linguistics-related or terminology-related research questions, the IAL is an active figure in several research fields and an active producer of manually crafted and curated high-quality datasets.

¹⁴<https://vlo.clarin.eu>

¹⁵<https://ilc4clarin.ilc.cnr.it/services/>

¹⁶<https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main\Page>

¹⁷In particular, the Italian tokenizer whose REST APIs are described at <http://ilc4clarin.ilc.cnr.it/services/ltfw/readme>, while the CMDI file used by Weblicht is available from <http://hdl.handle.net/20.500.11752/ILC-85@format=cmdi>.

¹⁸<http://hdl.handle.net/20.500.11752/ILC-59>

¹⁹<http://cophilab.ilc.cnr.it:8080/LatMorphWebApp/services/complete/<words>>

²⁰<http://hdl.handle.net/20.500.11752/ILC-55>

²¹<http://hdl.handle.net/20.500.11752/ILC-56>. Such format is compatible with the Global Wordnet initiative <http://globalwordnet.org/wordnets-in-the-world/>.

²²<https://weblicht.sfs.uni-tuebingen.de>

²³<http://www.eurac.edu/en/research/autonomies/commul/Pages/>

As regards linguistics-related questions, the IAL is a known figure in the research fields of Learner Corpora, Didactics and E-lexicography. Among the initiatives undertaken for the field of Learner Corpora, the IAL has created or participated in the creation of several Learner Corpora such as Kolipsi (Abel et al., 2012), KoKo (Abel et al., 2014), and Merlin (Wisniewski et al., 2013). It also has organized in 2017 the 4th Learner Corpus Research Conference²⁴. As regards to Didactics, the IAL has both strong connections with schools and policy makers in and outside South Tyrol and organizes a number of workshops²⁵ and training courses for teachers and pupils (Engel and Colombo, 2018). It also organized the international conference on language competences "Sprachkompetenzen erheben, beschreiben und fördern im Kontext von Schule und Mehrsprachigkeit"²⁶ celebrated in Bolzano in 2017. Finally (with regards to linguistics-related questions), the IAL is an active member of the COST Action "European Network for e-Lexicography" (ENeL), is currently leading the European Association for Lexicography (EURALEX) and has organized in 2014 the 16th edition of the EURALEX International Congress²⁷.

As regards terminology-related questions, the IAL is active in the field of Legal Terminology, for which it has produced and made available several terminological datasets such as the LexALP and Bistro Information Systems (Chiocchetti et al., 2013; Lyding et al., 2006; Streiter et al., 2004). The IAL is also part of the ISO Committee TC/37 for "Terminology and other language and content resources", is an active member of the RaDT²⁸, is part of the beta-test group for the SDL Multiterm and Trados Studio²⁹ and is acting on regular occasions and through several local projects as terminological consultant for the local South Tyrolean government.

With the rest of its workforce providing assistance on automatic processing for their colleagues, the IAL has also become over time an active figure in the domain of Language Technologies, especially with regards to the automatic processing of the South Tyrolean German Dialect. Among the efforts undertaken for this field, the IAL has developed expertise for non-standard written communication such as computer-mediated communication (CMC), with a special focus on social media, and webcorpora. In that research context, it has released the CMC corpus Didi (Frey et al., 2015) and the Webcorpus Paiza (Lyding et al., 2014). It also has organized in 2017 the 5th Conference on CMC and Social Media Corpora for the Humanities³⁰ and is an active contributor in a TEI Special Interest Group (TEI-CMC-SIG). Finally, as regards Language Technologies, the IAL also started a very CLARIN-alike local project named DI-ÖSS³¹ which aims at establishing a local digital infrastructure among the South Tyrolean language stakeholders allowing them to benefit from each others' expertise and services.

Except for specific cases, the IAL intends to integrate as many resources as possible into its CLARIN DSpace repository³². Because of its diversity in terms of research subjects and member profiles, the IAL relies on a varied set of workflows and can accordingly be an asset by providing a range of expertise of interest to a larger scope of stakeholders. Therefore, it also intends to be involved in several CLARIN initiatives and committees³³.

3.2.3 CLARIN as an asset for the IAL

The main added value from the IAL's participation to CLARIN is the number of opportunities it offers in terms of sustainability, an aspect that became key in the IAL's strategy as it kept on growing over the years. In that aspect, an initiative such as CLARIN DSpace greatly benefits the IAL which could not hope to develop such an advanced solution on its own.

²⁴<http://lcr2017.eurac.edu/>

²⁵Up to now, more than 3000 pupils (aged 8 to 18) took part in the offered didactic activities.

²⁶"Describe, nurture, and improve language competencies in the context of school and multilingualism".

²⁷<http://euralex2014.eurac.edu>

²⁸"Rat für Deutschsprachige Terminologie" (an expert panel including large institutions such as the UNESCO).

²⁹Leading professional solutions in language and content management services with a focus on terminology and translation.

³⁰<https://cmc-corpora2017.eurac.edu/>

³¹"Digitale Infrastruktur für das Ökosystem Südtiroler Sprachdaten und -dienste" (Digital infrastructure for the South Tyrolean ecosystem of language data and services).

³²<https://clarin.eurac.edu/>

³³Members of IAL already participate actively in the CMDI taskforce and the CLARIN DSpace initiative.

In a different but similar logic, as outlined earlier, the research profile of the IAL is rather varied and as such the IAL lacks often enough the tools (or uses suboptimal ones) to pursue some research opportunities, as it cannot afford developing and maintaining new ones. However, CLARIN as a whole is even more varied in terms of research profiles and a number of CLARIN-related initiatives, targeted at first to the needs of other institutions, directly address needs of the IAL. A good example is the Language Resource Switchboard which allows non-expert stakeholders to seamlessly use advanced natural language processing tools and can thus allow linguists and terminologists at the IAL to test and develop independently their own research ideas, while relying on their colleagues' expertise in language technologies for the later stages (e.g. for the fine tuning of the automatic tools). In that perspective, such technologies, despite having been developed independently of the IAL, directly tackle one of its needs³⁴.

Finally, CLARIN represents a great asset for the IAL in terms of visibility and dissemination. Indeed, because the IAL is an active producer of high-quality datasets, being able to reference such datasets on international catalogues such as the VLO is particularly interesting.

3.3 Synergies between the DSFUCI and CLARIN

3.3.1 The DSFUCI in few words

The *Dipartimento delle Scienze della Formazione, Scienze umane e della Comunicazione interculturale* (DSFUCI) is one of the 15 Departments of the *Università di Siena* and is located in Arezzo. The Arezzo campus brings together a community of scholars with a range of methodological approaches and research interests in various areas of education, languages, the humanities, and the social sciences. The Department coordinates and promotes theoretical and applied research projects aimed in particular at improving and changing life and work styles; strengthening cultural, linguistic and professional skills of adults and professionals; studying the role of languages, technologies and the media in today's world and in the historical development of groups and social communities; providing services to public and private organizations, administrations and professional associations in the realm of human resources development (educators, language experts, school teachers, cultural managers, trainers and middle management).

3.3.2 The DSFUCI as an asset for CLARIN

The DSFUCI carried out together with the *Scuola Normale Superiore di Pisa* (Pier Marco Bertinetto, p.i.) the *Grammo-foni* (Gra.fo) project (Calamai and Frontini, 2016), a co-founded project³⁵ devoted to the building of a digitization and cataloguing system with the aim of creating a regional network for the management of speech and oral archives of the past (Calamai et al., 2013). The preservation of analogue archives, that have so far remained unknown to the large public, entailed their detection as a first step, and then the digitisation (including restoration, when necessary) and cataloguing of the recordings contained in them. The oral documents preserved are disseminated via a web portal³⁶ that allows registered users to access the audio files and the corresponding cataloguing records, together with the relative transcriptions and accompanying material (when available). A subsequent project, *Voci da ascoltare*³⁷, was devoted to the dissemination of oral archives to high school students and also to the building of cultural trail via Mobile APPs (Pozzebon and Calamai, 2015; Pozzebon et al., 2016). Therefore, with respect to the speech and oral archives domain, the participation of DSFUCI and the Gra.fo archive in CLARIN would give several advantages. With over 3,000 hours of digitized recordings and the incredibly vast range of type of documents and topics covered, the Gra.fo archive is a unique and exemplary accomplishment in the Italian panorama. Having preserved such a significant collection of oral documents, Gra.fo not only constitutes a precious repository of Tuscan memory and provides a first-hand documentation of Tuscan language varieties from the early 1960s, but also represents a model for other research groups or institutions dealing with oral archives. Gra.fo covered the entire workflow with respect to the managing of oral archives: from digitization to long-term preservation, cataloguing

³⁴The interest in being involved in several CLARIN initiatives and committees is also motivated by the possibility to detect, influence and contribute to other useful initiatives such as the Switchboard.

³⁵Regione Toscana PAR FAS 2007-13

³⁶<https://grafo.sns.it>

³⁷*Università di Siena and Unicoop Firenze*, 2016-2017.

and description, ethical and privacy issues managing, and dissemination, also in terms of public history and general public involvement (Calamai et al., 2016).

Nevertheless, the DSFUCI's commitment to speech and oral archives does not confine itself to the *Grafo* experience. We succeeded in discovering and locating the first oral archive related to an Italian psychiatric hospital – which was located in the same buildings as the department in Arezzo, also where the historical archive of the Arezzo psychiatric hospital is hosted. The oral archive of *Anna Maria Bruzzone*, an analogue archive (made of 36 compact cassettes) contained the testimonies (life stories) of more than thirty former patients. It represents the documental basis of the book *"Ci chiamavano matti. Voci da un ospedale psichiatrico"* (Bruzzone, 1979). The author wrote it after a two-month stay in Arezzo, when she spent almost every day in the hospital. The book testifies to the patients' miserable lives inside and outside the hospital and sheds light on the atrocity of their everyday condition by *letting them speak for themselves*. Yet, what the book contains is not their actual voice: their voice is contained in the tapes that Bruzzone recorded during her research, when she witnessed the lives of the inpatients, in a continuous dialogue of which only a part is collected in the published interviews. The tapes were donated by the heirs and we are currently working on their digitisation and on metadata description.

3.3.3 CLARIN as an asset for the DSFUCI

Being part of CLARIN would benefit the speech sciences and oral history research communities in at least three main aspects: (1) the possibility to use a shared and internationally consistent metadata standard (e.g., the OralHistory profile in the CLARIN component registry³⁸); (2) the possibility to ensure the long-term preservation of the original speech data (both preservation and access copy) and of the metadata according to the FAIR principles (Wilkinson et al., 2016); (3) the possibility to offer a proper reuse of research data (license agreement, ethical and legal issues). As for (3), the inclusion of a member of DSFUCI in the CLARIN Legal Issues Committee³⁹ may be considered as the first step towards a more conscious involvement of the Italian research communities in the ethical and legal issues associated to the web dissemination and re-use of speech and oral archives.

At present, another crucial issue is represented by Automatic Speech Recognition tools. One of the aims of the Oral History research group inside CLARIN was to provide full Speech Recognition for different languages in order to perform one of the main "steps" envisaged in the OH transcription chain⁴⁰, enabling the researchers to go from a "recorded interview" to a findable, accessible and viewable digital AV-document with relevant metadata on the Internet. Italian language is devoid of a web-based ASR, which would be of a great benefit for both communities of oral historians and linguists.

3.4 Synergies between the DFCLAM and CLARIN

3.4.1 The DFCLAM in few words

The *Dipartimento di Filologia e Critica delle Letterature Antiche e Moderne* (DFCLAM) of the *Università di Siena*, ranked in 2018 as one of the national excellence departments by the Italian Government (MIUR), focusses on the philological, literary and anthropological competences that lie at the very heart of the study of literary texts, from the ancient world to modernity and for each literary genre. The interaction between philology and literature is central in the long-standing European humanistic tradition and the history of the Department includes significant names of the Italian literature (Antonio Tabucchi, Franco Fortini, Alessandro Fo, etc.) and some forerunners of the application of anthropological methods to literature (M. Bettini, S. Ronchey). In particular, its strongest points of engagement concern the anthropology of the ancient world (Centre AMA), the Italian contemporary literature (Centro Fortini) and the study of medieval literatures (Latin and Romance) through digital methods and tools. The Department includes some research centres such as the Centre for Comparative Studies « I Deug-Su » which is strongly engaged also in research on digital humanities and three laboratories on digital humanities funded by a development project newly approved by the MIUR.

³⁸https://catalog.clarin.eu/ds/ComponentRegistry#/?itemId=clarin.eu\%3Acr1\%3Ap_1369752611610®istrySpace=public

³⁹<https://www.clarin.eu/governance/legal-issues-committee>

⁴⁰<http://oralhistory.eu/workshops/transcription-chain\#transcriptions>

3.4.2 The DFCLAM as an asset for CLARIN

The DFCLAM committed itself to offering data and free online access to some digital archives of literary and historical texts: among them the ALIM (the Archive of the Italian Latinity of the Middle Ages), the largest digital library (with textual analysis tools and a medieval-Latin lemmatizer), which includes Latin texts and documents, encoded in XML-TEI from philologically checked sources or firstly edited from manuscripts, produced in Italy during the Middle Ages. Strategies for importing the metadata of ALIM in the CLARIN-ILC repository through a shared TEI-header are under study, as well as procedures for delivering dedicated tools for textual and linguistic analysis through the CLARIN channels. This would allow meta-queries and cross-queries on semantic items which could connect Latin and modern European languages derived from Latin and allow to develop semantic trees and networks of lexical derivations at the very heart of the European shared lexicon.

4 CLARIN-IT Events & promotion

4.1 Lectures

Four lectures were given to introduce CLARIN-IT to the next generation of collaborators.

A contribution called *"Language Resources and Infrastructures for Digital Humanities"* was presented at the *Curso de Verano 2016 "New trends in quantitative and computational linguistics"*, organized by the Universidad de Castilla-La Mancha in Ciudad Real, Spain.

A keynote lecture on *"Humanities: advantages, opportunities and benefits of the CLARIN Research Infrastructure and the CLARIN-IT national node for the Italian community."* was performed at the final ceremony of the Master Digital Humanities (2016-2017), held in Venice at the *Università di Ca' Foscari*. Subsequently, a lecture on *"Digital Humanities and Research Infrastructures: CLARIN"* was given during the Course *"Digital Humanities: Web Resources, Tools and Infrastructures"* of the third edition (2017-2018) of the Master in Digital Humanities⁴¹.

Finally, during the first and second part of the Workshop *"Digital Humanities and Greek Philology: resources and research infrastructures applied to the study of ancient Greek"* organized at the *Università di Parma* in November and December 2017, two lectures were given. The first one was entitled *"New technologies and new investigations: CLARIN-IT and some examples of application to the study of ancient Greek"* whereas the second one was entitled *"Infrastructures of Research and Classical Studies. CLARIN-IT: opportunities and perspectives"*. This event was devoted to the discussion of the opportunities and research perspectives offered by the collaboration between Research Infrastructures and the Digital Classics community. Different approaches to a traditional discipline are expected to offer, in perspective, new study habits that, based on the good practices inherited from the previous tradition, allow the development of new research methodologies and teaching practices.

4.2 Participation in Italian events

In order to raise awareness among the Italian research communities and extend the consortium, CLARIN-IT members have been participating in four relevant Italian events.

A keynote *"CLARIN-IT, l'Infrastruttura di Ricerca per le Scienze Umane e Sociali"*⁴² was presented at the 5th *Annual Conference of the Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD)* held in Venezia in September 2016 and in which numerous Italian researchers in Digital Humanities were taking part. During the conference a survey on CLARIN-IT aiming at raising awareness about CLARIN and collecting needs, requirements and expectations was launched (see Section 5.).

CLARIN-IT was presented at the Workshop *"Utilizzo e diffusione di metodi, strumenti e tecnologie digitali per gli studi filologici: l'applicazione della filologia digitale al greco antico"*⁴³ and at the Seminar *"Le risorse informatiche applicate alle discipline umanistiche: strumenti e metodi con esempi*

⁴¹CLARIN-IT, in collaboration with the AIUCD, patronizes the third Master in Digital Humanities (a.a. 2017-2018).

⁴²"CLARIN-IT: A research infrastructure for the Social Sciences and Humanities"

⁴³"Utilisation and dissemination of methods, instruments and digital technologies for the philologies: application of digital philology to ancient Greek"

sull'utilizzo didattico nelle discipline classiche"⁴⁴ held in October 2016. Such events were organized by the *Dipartimento di Discipline Umanistiche* of the *Università di Parma*, which is about to become a member. The contribution was called *"Infrastrutture di ricerca nel settore umanistico"*⁴⁵.

CLARIN-IT was also presented at the GARR 2016 Conference *"The CreActive NETwork: uno spazio per creare e condividere nuova conoscenza"*⁴⁶, held in November 2016 and organized by the *Gruppo per l'Armonizzazione delle Reti della Ricerca* (GARR), an Italian network aiming at providing high-performance connectivity and developing innovative services for the daily activities of teachers, researchers and students and with which CLARIN-IT is actively collaborating on technical questions. The presentation was entitled *"Corpora digitali: dall'obsolescenza tecnologica, alla salvaguardia e alla condivisione"*⁴⁷ (Sassolini et al., 2016).

4.3 Organization of CLARIN & CLARIN-IT events

A presentation of CLARIN-IT, the ILC and the ILC4CLARIN repository was held at the *Consiglio Nazionale delle Ricerche* in Pisa in March 2017. This presentation was the occasion to raise awareness among colleagues about the aim and functioning of CLARIN, its potential and its benefits.

A first result of CLARIN's interest towards the Tuscan speech and oral archives can be found in the CLARIN Oral History workshop (Arezzo, May, 10-12 2017; Henk van den Heuvel p.i.), whose aim is the finalization of the setup of a transcription chain for OH-interviews⁴⁸. An implementation plan for an OH transcription chain that can be integrated into the CLARIN infrastructure has been set up during the Arezzo workshop. As for the Italian community, the meeting brought together the CLARIN-IT executive committee (ILC) and representatives from the *Italian Speech Sciences Association* (AISV) and the *Italian Oral History Association* (AISO). The workshop undertook the challenging task of putting together different kinds of expertise (from Linguistics to Oral History, to Language and Speech Technology, to Infrastructure Analysis and Implementation).

In June 2017, an application to organize and host the 7th edition of the CLARIN conference was submitted and selected, thus acknowledging the efforts and capacities of the CLARIN-IT consortium, its contribution as a full member of the federation, as well as demonstrating the interest in supporting its growth. The CLARIN Annual Conference is an important scientific event where the wider Humanities and Social Sciences communities can meet in order to exchange ideas and experiences with the CLARIN infrastructure. This includes the design, construction and operation of the CLARIN infrastructure, the data, tools 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. The Special Thematic session for this edition will be in the areas of multimedia, multimodality and speech, including the collection, annotation, processing and study of audio, visual or multimedia data with language as an important part of the content. The conference will be held in October 8-10, 2018 in Pisa. It is expected to last 3 days and receive around 200 participants.

On October 4th, IAL organized a CLARIN User Involvement event titled *"How to use TEI for the annotation of CMC and social media resources: a practical introduction"*. The event was held in conjunction with the 5th Conference on CMC and Social Media for the Humanities (cmccorpora17)⁴⁹.

5 Survey

5.1 Motivation

In CLARIN, users are recognized as a central part of the infrastructure and of any service design process, but saying that our audience are Humanists is not enough. We have a wide range of scholars working within the Academy or research institutions who have different needs.

⁴⁴"Computational resources applied to the humanities: tools and methods with examples on their didactic use in the classical disciplines"

⁴⁵"Research infrastructures in the Humanities"

⁴⁶"The Creative Network: a space for creating and sharing new knowledge"

⁴⁷"Digital corpora: from technological obsolescence towards preservation and sharing"

⁴⁸<http://oralhistory.eu/workshops/arezzo>

⁴⁹<https://cmc-corpora2017.eurac.edu/>

While there is a soaring interest for the use of digital resources and related tools in the broader context of Humanities, some specific scientific communities are still reluctant to adopt them. We performed a survey to ascertain the current interest in digital methods, the practice and the related needs within the community of classical philologists; in particular, it was performed on a restricted sample of Italian digital humanists with Ancient Greek Philology as the main focus of interest.

Other surveys were sporadically carried out during the last decade, aiming at collecting input from sectors that, although not strictly within the realm of Digital Classics, may have similar requirements and arrive at similar conclusions as far as resource design is concerned.

A point worth remarking is that the preceding studies concern a wide spectrum of scientific interests within the Digital Humanities realm and involve mostly English native speakers scholars; our study, instead, focused on the specific scientific community dealing with Digital Classics and aimed at evaluating the impact of digital techniques in their practice.

5.2 Context

The CLARIN-IT survey collects the points of view of a restricted sample of Italian digital humanists, with a focus of interest on ancient Greek philology. This area of study is a relatively small field but it retains great interest in Italy, where it also looks back to a great tradition. It also includes university students and schoolteachers. Moreover, Italian Scholars of Ancient Greek are an active part of a large international community (especially in Europe, North and South America). The perspective of the study was enhanced by the fact that its spectrum involves also Latin, Ancient History and Philosophy, and Classics in general. Finally, it is important to remember that Ancient Greek studies are an essential part of our Occidental Cultural Heritage, and it is crucial that the highest number of people knows these texts and their contents. For all these reasons, the Italian Ancient Greek community is an excellent field to test new opportunities about knowledge and quality in transmission of ancient texts.

The questionnaire was sent to a selected group of Italian researchers whose main focus of study was Ancient Greek language, although their interests span over a broader area, encompassing Greek and Latin literature. The sample was numerically consistent with the survey target (about 10% with respect to the potential target population of about 130 people). The survey focused on the digital resources and tools needed to support an excellent and usable digital edition of an ancient text. For this reason, first, the applicants were asked to evaluate the tools they use and know. They were then asked to indicate their expectations towards technologies and to rank a set of four functionalities in priority order. Finally, they were asked to rate, on a 1-5 scale, the set of functionalities considered as crucial.

5.3 Main Outcomes and Action Plan

Consistently to the preceding surveys, the key outcome was that most of the available resources do not respond to users' requirements⁵⁰. Many respondents pointed out that important research needs in the field are models and software for authoring, editing, indexing and presenting a digital edition, how to link it to the available resources and improve them. All of them insisted on the need to develop and/or make tools more reliable and usable, thus lamenting the absence of tools integrating textual data and bibliography links, or hypertext links with other texts or resources available.

Based on the outcomes of this survey, CLARIN-IT could address a set of R&D priorities that may be the base for establishing a research and innovation action plan for Digital Classics. As it currently stands, the plan foresees a workbench in which to insert text in a simple and intuitive way and visualize its encoding with specific TEI transcription; provide apparatus, literature and translation, link together primary sources and lexica, provide textual (and metrical) analysis and commentary, and offer search tools. We are developing a sample prototype to submit for evaluation by end-users.

At a larger scale, the work represented one of the first attempts undertaken within the context of CLARIN-IT to contribute to the wider impact of CLARIN on the specific Italian community of users interested in the application of Digital Humanities to the field of Classics and to ancient world studies.

⁵⁰For an extensive analysis, see Monachini et al. (2018).

6 Next steps

6.1 For the CLARIN-IT consortium as a whole

For the consortium as a whole, the next step is to include more members which will depend on whether or not a national funding for personnel can be secured. The plan is to agree with the Ministry on a national project aiming at strengthening the infrastructural activities in Italy to foster the use of digital technologies in the Humanities, through the collection of needs and requirements of the community and the development of case studies. CLARIN-IT will enhance the use of language resources and technology through the Italian infrastructure and, at the same time, will encourage innovation in research paradigms and methodologies of the sectors. The consolidation of the consortium will respond to representativeness criteria. While a scientific criterion is aimed at covering research sectors related to the study of language and gather language resource producers, linguists, computational linguists and language engineers, a geographical criterion will also be considered to ensure territorial coverage. Participation in the national consortium of all the most important research centers in the language technologies will allow to achieve the goal of coverage, ensuring the long-term preservation of the great wealth of digital resources and their easier access to the scientific community. The CLARIN-IT consortium aims to attract the scientific communities of the various fields: classical, modern and contemporary history, literary studies, political science, communication science, sociology, theology, philosophy, social anthropology and ethnography, linguistics and philology. Furthermore, CLARIN-IT is also aimed at attracting disciplines that make use, albeit less massively, of text resources and technologies, such as law, education, archeology, artistic disciplines and entertainment, design, architecture, music, demography, human geography, economics, social and political studies, the history of science and medicine. The CLARIN-IT consortium is also deeply involved in one of the aspects on which ERIC insists, namely the training sector, with the launch of master's or doctoral theses and university courses in line with the objectives of CLARIN.

The Digital Classics survey, now publicly available on the CLARIN-IT channels⁵¹, may further help the Italian Consortium in fostering new and sustaining existing knowledge in Digital Classics (DC). CLARIN-IT will play an important role in disseminating the results to the relevant academic, cultural, industrial communities and the interested public.

Furthermore, our plans are to extend the survey to other CLARIN consortia, thus helping to identify gaps and drive the development of new technologies for ancient studies at large. This will contribute to the general CLARIN mission to grow its infrastructure so as to serve in a better way the international community of scholars from any disciplines dealing with language and help them to boost their studies.

Last but not least, since each consortium is unique but none is fully different from the others, the CLARIN federation constitutes an important source of inspiration as regards to the next efforts and initiatives to undertake. Therefore, we will keep on observing the past and on-going initiatives undertaken by other national consortia and, whenever relevant and possible in practice, undertake similar ones.

6.2 For each CLARIN centre

Regarding the ILC4CLARIN, the next steps are to complete the set of linguistic resources freely accessible through its online portal and achieve a CLARIN-B certification in 2018, which is under way.

A high priority task of the near future is the integration of the web services developed within previous funded projects (and thus already available) into the CLARIN federated services Language Resource Switchboard and Weblicht. As mentioned in Del Gratta (2018), these are mainly basic NLP services that may serve various purposes and can thus be included in useful analysis chains for textual research.

For the IAL, the next steps are to get recognized as a CLARIN-C Centre as soon as possible. During the course of 2018, IAL will start integrating all its language resources into its recently-established CLARIN DSpace repository, and will undertake the additional steps needed to achieve a B Centre certification within 2018 or 2019. Finally, through its (CLARIN-like) local DI-ÖSS project, the IAL intends to organize a number of events with the South Tyrolean language stakeholders to raise awareness around

⁵¹<http://www.clarin-it.it/it/content/sondaggio-current-practice-digital-classics-tools>

digital infrastructures, the DI-ÖSS project itself, the CLARIN-IT consortium and the overall CLARIN initiative as a whole.

As pertaining to the DSFUCI, the next steps are to make the Gra.fo digital archives accessible via CLARIN DSpace (Calamai et al., 2017) and ensure their long-term preservation, to describe new digital archives according to CLARIN metadata profiles (e.g. BAS-COALA service) and to update the Registry of Oral History Collections in Italy, which is made accessible and maintained by CLARIN ERIC. Finally, another future objective is to strengthen the collaboration among linguists and oral historians in the speech and oral archives domain.

Finally, regarding the DFCLAM, the next steps are to make the ALIM digital archive accessible via ILC4CLARIN and ensure its long-term preservation.

7 Conclusion

This paper presented the current Italian CLARIN consortium and discussed its current state of affairs. This paper also provided a number of information on its current members, especially with regards to what they offer to CLARIN in terms of resources, services and expertise, and what CLARIN offers them to further their own research, as well as information on the institutions that are expected to join in the close future. The events and initiatives undertaken at the Italian level have also been discussed together with one planned in a close future, namely the 2018 edition of the CLARIN conference. This paper finally outlined the conclusions of a user survey performed to understand the expectations of a targeted user population and provided indications regarding the next steps planned.

As one can observe from the efforts undertaken and the results achieved, CLARIN-IT has a lot to offer to CLARIN and vice-versa. Despite limited means, CLARIN-IT is slowly but surely taking its place in the CLARIN landscape. The consortium has yet to grow larger and address several questions. Nonetheless, its steady growth and its widening participation in the CLARIN federation are positive indications as regards to the challenges to come.

References

- Andrea Abel, Chiara Vettori, and Katrin Wisniewski. 2012. *Gli studenti altoatesini e la seconda lingua: indagine linguistica e psicosociale Die Südtiroler SchülerInnen und die Zweitsprache: eine linguistische und sozialpsychologische Untersuchung*. Eurac Research.
- Andrea Abel, Aivars Glaznieks, Lionel Nicolas, and Egon Stemle. 2014. KoKo: an L1 Learner Corpus for German. In *Proceedings of the LREC Conference*.
- Daan Broeder, Menzo Windhouwer, Dieter Van Uytvanck, Twan Goosen, and Thorsten Trippel. 2012. Cmdi: a component metadata infrastructure. In *Describing LR with metadata: towards flexibility and interoperability in the documentation of LR workshop programme*, volume 1.
- Anna M. Bruzzone. 1979. *Ci chiamavano matti. Voci da un ospedale psichiatrico*. Einaudi.
- Silvia Calamai and Francesca Frontini. 2016. Not quite your usual kind of resource. Gra.fo and the documentation of Oral Archives in CLARIN. In *Proceedings of the 5th CLARIN Annual Conference (CAC)*.
- Silvia Calamai, Pier Marco Bertinetto, Chiara Bertini, Francesca Biliotti, Irene Ricci, and Gianfranco Scuotri. 2013. Architecture, methods and purpose of the Gra. fo sound archive. In *Digital Heritage International Congress (DigitalHeritage)*, volume 2, pages 439–439. IEEE.
- Silvia Calamai, Veronique Ginouvès, and Pier Marco Bertinetto, 2016. *Sound Archives Accessibility*, pages 37–54. Springer International Publishing, Cham.
- Silvia Calamai, Francesca Biliotti, and Aleksei Kelli. 2017. Authorship and ownership in the digital oral archives domain: The Gra.fo digital archive in the CLARIN-IT repository. In *Proceedings of the 6th CLARIN Annual Conference (CAC)*.
- Elena Chiocchetti, Barbara Heinisch-Obermoser, Georg Löckinger, Vesna Lušicky, Natascia Ralli, Isabella Stanizzi, and Tanja Wissik. 2013. Guidelines for collaborative legal/administrative terminology work. *EURAC*.

- Riccardo Del Gratta. 2018. (Re)Using OpeNER and PANACEA Web Services in the CLARIN Research Infrastructure. In *Digital Infrastructures for Research 2017*.
- Dana Engel and Sabrina Colombo. 2018. Strategien in der Förderung von Multilingual Awareness im Rahmen der Südtiroler Wanderausstellung „Sprachenvielfalt – in der Welt und vor unserer Haustür“. *Sprachen lehren, Sprachen lernen*.
- Jennifer-Carmen Frey, Aivars Glaznieks, and Egon W Stemle. 2015. The DiDi Corpus of South Tyrolean CMC Data. In *Proceedings of the 2nd Workshop of the Natural Language Processing for Computer-Mediated Communication/Social Media*.
- Twan Goosen and Thomas Eckart. 2014. Virtual language observatory 3.0: What’s new. In *CLARIN annual conference*.
- Erhard Hinrichs, Marie Hinrichs, and Thomas Zastrow. 2010. WebLicht: Web-based LRT services for German. In *Proceedings of the ACL 2010 System Demonstrations*, pages 25–29. Association for Computational Linguistics.
- Verena Lyding, Elena Chiochetti, Gilles Sérasset, and Francis Brunet-Manquat. 2006. The LexALP information system: Term bank and corpus for multilingual legal terminology consolidated. In *Proceedings of the workshop on multilingual language resources and interoperability*. Association for Computational Linguistics (ACL).
- Verena Lyding, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell’Orletta, Henrik Dittmann, Alessandro Lenci, and Vito Pirrelli. 2014. The PAISA corpus of Italian web texts. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*.
- Monica Monachini, Anika Nicolosi, and Alberto Stefanini. 2018. Digital Classics: A Survey on the Needs of Ancient Greek Scholars in Italy. In *Proceedings of the CLARIN 2017 Conference*. Linköping University Electronic Press.
- Alessandro Pozzebon and Silvia Calamai. 2015. Smart devices for Intangible Cultural Heritage fruition. In *Digital Heritage, 2015*, volume 1, pages 333–336. IEEE.
- Alessandro Pozzebon, Francesca Biliotti, and Silvia Calamai. 2016. Places Speaking with Their Own Voices. A Case Study from the Gra. fo Archives. In *EUROMED 2016*, pages 232–239. Springer.
- Eva Sassolini, Sebastiana Cucurullo, and Alessandra Cinini. 2016. I corpora digitali: dall’obsolescenza tecnologica, alla salvaguardia e alla condivisione. In *GARR Conference Proceedings*.
- Oliver Streiter, Natascia Ralli, Isabella Ties, and Leonhard Voltmer. 2004. BISTRO: the online platform for terminology management. Structuring terminology without entry structures. *Linguistica Antverpiensia, New Series–Themes in Translation Studies*, (3).
- Dieter Van Uytvanck, Claus Zinn, Daan Broeder, Peter Wittenburg, and Mariano Gardelleni. 2010. Virtual language observatory: The portal to the language resources and technology universe. In *Seventh conference on International Language Resources and Evaluation [LREC 2010]*, pages 900–903. European Language Resources Association (ELRA).
- Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3:160018.
- Katrin Wisniewski, Karin Schöne, Lionel Nicolas, Chiara Vettori, Adriane Boyd, Detmar Meurers, Andrea Abel, and Jirka Hana. 2013. MERLIN: An online trilingual learner corpus empirically grounding the European reference levels in authentic learner data. In *6th edition of the ICT for Language Learning Conference (ICT4LL)*, Florence, Italy.
- Claus Zinn. 2016. The CLARIN Language Resource Switchboard. In *Proceedings of the 5th CLARIN Annual Conference (CAC)*.

CORLI: A linguistic consortium for corpus, language, and interaction

Christophe Parisse

Modyco, Inserm,
University of Nanterre,
France
cparisse@parisnantes.fr

Céline Poudat

Université Côte d’Azur,
CNRS, BCL, France
poudat@unice.fr

Ciara R. Wigham

Laboratoire de Recherche
sur le Langage
Université Clermont Au-
vergne, France
ciara.wigham@uca.fr

Michel Jacobson

LLL Université d’Orléans et
Tours, France
michel.jacobson@gmail.com

Loïc Liégeois

Department (optional)
CLILLAC-ARP (EA 3967)
& LLF
Université Paris Diderot,
France
loic.liegeois@univ-paris-diderot.fr

Abstract

CORLI is a consortium of Huma-Num, an organization that helps to develop digital humanities in France and provide services for this. CORLI is a consortium dedicated to linguistics and includes all aspects of linguistic research and development.

CORLI has a key role in corpus linguistics in France, and it can act as an interface or a facilitator between CLARIN and the scientific community of linguists. As France just joined CLARIN as an observer, the role of the consortium CORLI is very important in organizing the relationship between CLARIN and the French community.

The goal of CORLI is to help linguists create, use, and disseminate linguistic corpora and digital tools. CORLI has always maintained a policy of providing funding and technological help to finalize and publish corpora issued from a wide range of institutional or personal research projects. CORLI is also involved in recommending and broadcasting guidelines related to research and technical practices, especially about linguistic corpora. Finally, CORLI organises workgroups whose goal is to create and moderate networks that target tools and practices in linguistics. These workgroups are organised thematically around topics including metadata, formats, tools, and practices for corpus exploration, archiving systems, multimodal practices, and annotations. Their goal is to help showcase innovative work and trends undertaken in research labs and to finalize and disseminate current methods and practices in digital humanities research.

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

1 Introduction

1.1 What is CORLI?

CORLI (*Corpus, Langues et Interactions*: Corpus, Languages, and Interaction¹) is a French consortium of Linguistics laboratories, gathering people involved in linguistic research and teaching. It is one of several such consortia involved in digital humanities overseen by Huma-Num². Huma-Num, which stands for *Humanités Numériques* (Digital Humanities), was created to help specialists in the humanities use new digital material and services.

CORLI is steered by a board of people and laboratories representing the linguistic research community in France. Although the consortium's specific goals and the set of people and laboratories involved in CORLI may change from one year to another, the general goal is to promote the creation and the use of corpora in linguistics, and to represent the whole research community.

CORLI, as a consortium, focuses on knowledge information. It provides help to researchers through information support. It is also involved in facilitating corpus creation and dissemination, and designing tools or formats to handle linguistic data. The technical support for linguistic data is handled by the French CLARIN centres (ORTOLANG³, Cocoon⁴, SLDR⁵) or in some cases by the Huma-Num technical support.

1.2 Huma-Num

The goal of Huma-Num is to help and promote the use of digital technology for all humanities. For this purpose, they developed both technological and human responses to the queries from researchers and users in the humanities. Technological responses are means to store, process, broadcast, disseminate, search, and archive data. Human responses are consortia that target specific fields of study (for example linguistics, music, ethnology, etc.), or specific digital material used in corpora and databases (for example, maps, pictures, 3D images, etc.). The goal of a consortium might vary from one to another, but their general purpose is to develop and increment the digital data available, and to provide information and requirement about good practices for digital information. In a certain way, Huma-Num reproduces, for a larger set of scientific fields but at a smaller scale, what can be found in the CLARIN centres. On the one hand, some centres provide technical support, while on the other hand some centres provide knowledge information. Huma-Num, with the help of the French B and C CLARIN centres, provides technical support. CORLI provides knowledge information.

1.3 History of CORLI

CORLI was established in January 2016 and it is foreseen that the structure will run for another four years. It is built on previous consortia for linguistics, that ran from 2012 to 2015. The first one was *Corpus-écrits* (Research Infrastructure for Written corpora), which was specialized in corpora based on written material and the second one was *IRCOM* (Research Infrastructure for Oral and Multimodal Corpora) which was specialized in oral or multimodal corpora. The goal of the initial consortium projects, as defined by Huma-Num, was to help for the creation and deposition of corpora. At that time, the focus was on making previously existing corpus projects available that, previously, had never been made public, either due to the lack of access to a repository or due to a lack of technical information/expertise. Both consortia decided that they had, on the one hand, to provide technical or financial help to corpora that were not yet disseminated, and on the other hand, to provide good practices about norms, formats, rights, and dissemination.

This was decided after consulting the community thanks to the organisation of several general assemblies that were open to all colleagues who wanted to be involved in the creation or use of corpora. More recent decisions have been taken by the steering committee (see below) or after consultation with the workgroups. The fusion of the two previous consortia into a unique consortium has not changed the overall organisation and purpose of the consortium.

¹ <https://corli.huma-num.fr/>

² <http://www.huma-num.fr>

³ <https://www.ortolang.fr>

⁴ <https://cocoon.huma-num.fr>

⁵ <http://sldr.org>

2 Organisation

CORLI (see figure 1) has a *Comité de Pilotage* (CP: steering committee) which is responsible for deciding which annual goals CORLI should set itself and for handling its relationship with the parent organisation Huma-Num. Financial responsibility and management is under control from the *Institut de Linguistique Française* (ILF: Institute of French Linguistics, which is part of the national research council (CNRS) structure – see <http://www.ilf.cnrs.fr/>). The head of ILF is the official head of CORLI.

The CP is composed of specialists in the field of linguistics who are involved in corpus linguistics. The members of the CP are also representatives for the local research laboratories to which they belong. The number of CP members is not set, but is around twenty. This makes them very representative of the field. CP membership can easily be changed, according to the needs or contingencies of the CP members.

CORLI is also organized in *Groupes de Travail* (GT: thematic workgroups). GT membership is open to anyone who is involved in linguistics and all meetings are public. Members of a GT can be active, in the sense that they work on organizing scientific events, producing documents, or handling people that might be hired on specific projects. However, they may also be observers, in the sense that they participate in the discussions or provide their own experience to other members. This allows the consortium's work to be based on the real-life, current needs or knowledge of the larger scientific community that it represents.

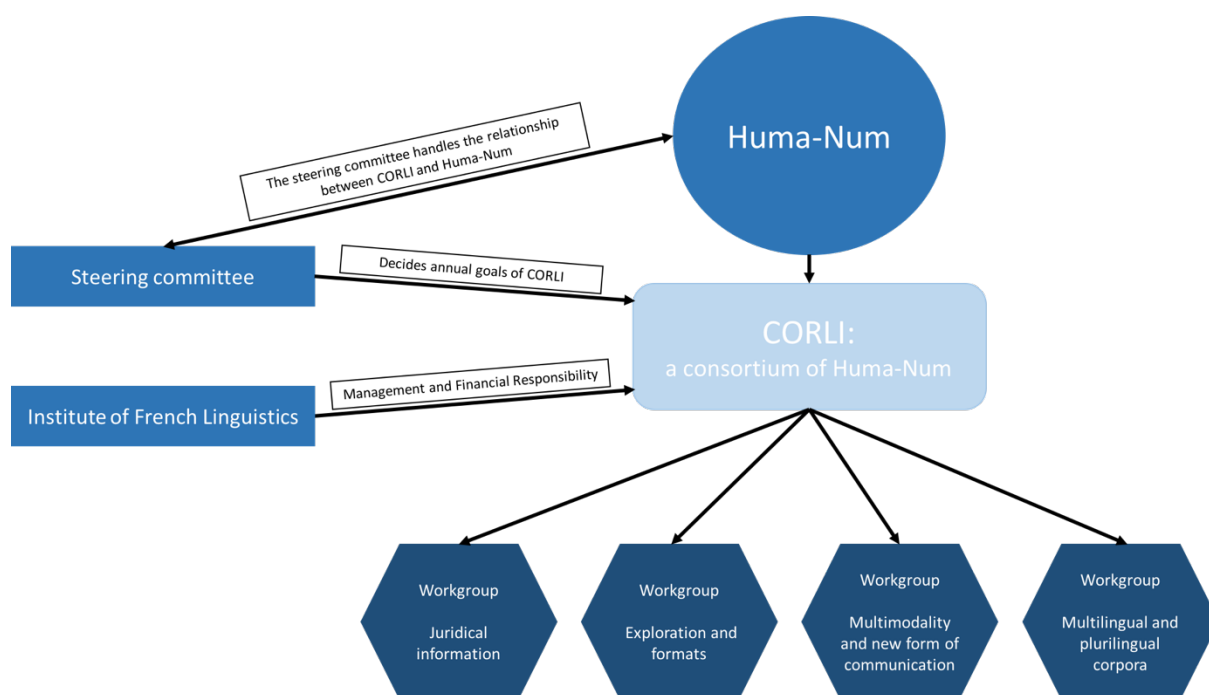


Figure 1: Organisation of CORLI

3 Goals

The goals of CORLI are defined each year by the CP. They are validated by the scientific committee of Huma-Num who sets the general goal of the consortia and is responsible for validating the project. The goals for 2017 are mainly follow-up tasks based on work that was accomplished in the previous years. The goals of CORLI are divided in two types of actions. General actions are under the direct control of the CP. They usually concern the whole community of linguistic research. The other types of action are the workgroups. They have a more specific purpose that either concerns only a part of the community or targets very specific tasks.

4 General actions

4.1 Describing resources

Although official repositories and archives are very efficient means to disseminate and preserve data about linguistic research or applied linguistics, they cannot and do not have to cover all uses and all types of corpora. Local research laboratories and projects are developing tools, creating data, managing working groups, and they cover a much larger range of formats, tools, and purposes than the official centres. All this material represents actual ongoing research. The nature of this material makes it difficult to be harvested using traditional open access inventories (OAI) because it is not yet normalized or made into standards. The purpose of this action is to provide a means to centralize information about these resources.

We are developing a portal that will make it possible for the people responsible for this type of resources to document themselves their product. It is necessary in this case to reverse the OAI mechanism, as it is only the people themselves who can provide information about their own resources, because this information is non-standard by nature (due to the on-going nature of the projects). The portal will make it possible to document the availability of digital linguistic resources, their formats and description, as well as how they can be accessed. This concerns a variety of data types including corpora, texts, lexicons, thesauri, dictionaries, etc. as well as tools such as software, libraries, scripts, stylesheets, query portals, etc.

4.2 Evaluation of resources

The large development of the use of corpora in linguistic research has led to an important investment of researchers and laboratories in the creation of corpora, but also digital tools, resources, and data. This takes a lot of time and resources, and is very important for the visibility of the people who worked for the projects. In a time where evaluation of research and justification of the means (financial or human) is very important, it is necessary to take into account the production of researchers and laboratories, not only in terms of paper or book publications, but also in terms of corpora or digital data.

This new development of the evaluation of research means that it is important in the evaluation to take into account the existence of digital material, but also the quality of this material, which is yet not clearly defined. The members of CORLI, as representatives of the research community, feel that this it is necessary to offer guidelines for corpus evaluation. The goal of this action is to help to define possible evaluation criteria for linguistic corpora, taking into account not only the size of corpora, but also their availability, the format, and the future uses by the community. The criteria would then be used for the evaluation of scientific work. The system of peer evaluation for corpora is in keeping with the French tradition of evaluation of scientific research.

4.3 Technical courses and information

The use and creation of corpora and digital data is possible only with the use of tools and methods that are rapidly evolving and developing. It is not possible for everyone to know which technology is the best or to learn to optimally use it by themselves. However, it is very important for senior or junior researchers to be up to date because it is necessary for their work and for teaching appropriate techniques and tools.

To answer this need, CORLI has been organizing annual technical courses. The nature of the courses is determined by the workgroups (see below) according to the actual needs of the community. For example, we have also included courses about data management, e.g. use of metadata or depositing corpora in repositories. The courses are divided into four main categories, as described below. Each category has several sub-categories. All courses are not presented each year, as this will depend on the actual requirements of the users.

Corpus annotation tools

The creation of oral or multimodal corpus linked with one or several media files requires the use of specialized software which is not easily mastered, especially at an advanced level. Several courses are

organized for different levels such as beginners, advanced, or experts for tools including CLAN⁶, ELAN⁷, Praat⁸, SPPAS⁹.

Corpus exploration tools

Existing corpora need to be exploited in the best way, including searching for data, classifying and categorizing, but also using advanced statistical tools and techniques. Some of these tools and methods have been developed in France. They are now frequently used in France by linguists and other researchers interested in text and corpus exploration. This explains why there is a large demand for training courses that cover these tools and methods. The tools that have been presented are, among others, TXM¹⁰ (textometric exploration, R queries, interface with CQP), Iramuteq¹¹ (interface from texts to many R libraries), Unitex¹² (graphic interface for building text parsers), Le Trameur¹³ (textometric exploration), Lexico5¹⁴ (textometric exploration), DTM-VIC¹⁵ (data and text mining), and Hyperbase web¹⁶ (textometric exploration).

Video and sound recording

Recording audio or video data for corpus creation is a very time consuming activity. It is also expensive and cannot easily be done again and again if some hiccups appear during the data collection. For this reason, it is very important to acquire data of the best quality with minimal risk of failure. This technical course has been organized for several consecutive years with great success. Information was also provided about the format for best and most useful data compression, and also with regards to data mixing or movie/audio editing.

Metadata and corpus dissemination

There is a very large consensus about the use of metadata in order to deposit and to find corpora or any data in the most efficient way. However, creating correct and useful metadata is not an easy task. This is basically the work of specialists such as librarians or information specialists. Moreover, the access to such people is not always possible in small laboratories or for small projects. Also, metadata for objects such as linguistic corpora are different to metadata for other type of material. Thus, specialists cannot always help people in linguistics because metadata for linguistics is still a fairly new domain. So, it is important that people are able to produce, by themselves, the most adequate metadata, and that they are able to use this information to find data for themselves. This is why CORLI has offered information sessions on metadata use and creation.

Another requirement today, which is complementary to the creation of correct metadata, is to deposit corpora and tools in repositories such as CLARIN centres. This operation is not always so easily done because this does not simply correspond to copying the data that people have on their computers. The data is not always easily reused by other people and long term archiving is not always possible if the data is not in an open source format as is required by institutes like CINES¹⁷ (long term digital archiving centre for France). CORLI has organized sessions to help people deposit their data, for example working interactively with their own data under supervision of a specialist in the field.

4.4 Finalization of corpora

Many corpora already exist but are not yet ready to be deposited in a corpus repository for dissemination, or this was never done due to lack of information or sometimes simply because of a lack of time and/or the opportunity to do this. CORLI, as well as the previous consortia *Corpus Ecrits* and *IRCOM*, but also projects such as ORTOLANG¹⁸ (CLARIN B Centre candidate in France), have organized calls for

⁶ <http://alpha.talkbank.org/clan/>

⁷ <https://tla.mpi.nl/tools/tla-tools/elan/>

⁸ <http://www.fon.hum.uva.nl/praat/>

⁹ <http://www.sppas.org/>

¹⁰ <http://textometrie.ens-lyon.fr/>

¹¹ <http://www.iramuteq.org/>

¹² <http://unitexgramlab.org/fr>

¹³ <http://www.tal.univ-paris3.fr/trameur/>

¹⁴ <http://lexi-co.com/>

¹⁵ <http://www.dtmvic.com/>

¹⁶ <http://hyperbase.unice.fr/>

¹⁷ <https://www.cines.fr/>

¹⁸ <https://www.ortolang.fr/>

finalizing corpora since 2013. The process was not exactly the same over the years, but most of the time this meant offering people the possibility to have some small financial or technical help to deposit their corpus in an official repository such the CLARIN B (candidate: ORTOLANG: Pierrel, 2014) and C (Cocoon¹⁹: Jacobson, Badin and Guillaume, 2015; SLDR²⁰: Bel and Gasquet-Cyrus, 2011) centres in France. Some conditions need to be fulfilled to ask for this support:

- ✓ The corpus should be already advanced;
- ✓ The corpus should complement the already available corpora;
- ✓ The corpus should be open access for research;

For the year 2017, the maximal financial help for individual projects was 7000 €. There were 25 submissions to the 2017 call, which represented much more than the possible budget of 40 000 €. After the scientific evaluation of the different projects, 13 projects were accepted for 2017. The previous years had roughly the same amount of selected projects. The type of help requested in the submissions is always quite diverse, and can include for simple cases only information, or financial help for actual cleaning and depositing of the data, and in more complex cases coding of the data or finalisation of data collection so as to make it possible to deposit the data in its final form.

5 Workgroups

The most important creative work done in CORLI is the product of the workgroups. Workgroups target thematic subjects, which means that each of them brings together specialists in the domain. The principle that underlies most of the workgroups is that they are open groups for any person who has an interest in the theme of the workgroup. This can be people working on fundamental or applied research who are specialists of the field, or people that are not specialists but need to work on the subject. This way it is possible for a workgroup to know the specific needs of the people working effectively on the subject, and to have or build adequate responses thanks to the guidance of actual specialists with practical experience on the subject. The product of the workgroups can take the form of recommendations, of reference documentation, of norms or formats, or in some case of software development for small size projects.

5.1 Workgroup 1: Exploration and formats

The goal of this workgroup is to advertise the most useful and efficient tools for creating and exploring all types of linguistic corpora. Another goal is to showcase good practices in the use of metadata and formats. When necessary, this workgroup participates in the creation of tools dedicated to conversion formats or metadata handling and also in the definition of corpora formats and metadata. The workgroup works hand-in-hand with both linguists, users, and tool developers.

This workgroup was formed by merging two workgroups from the previous consortia; one that was working on written corpus exploration, and one that was working on designing a common format and methods that allow users to aggregate oral corpora. Initially, the existence of two workgroups was a consequence of the current state of research tools and corpora. For written data, corpora are less difficult to build, and so large corpora have existed for a long time, which called for the development of tools that were adapted to corpus exploration and statistical analysis. For oral data, corpora are difficult and expensive to build, and tools were first developed so that transcription and linking was easily done. Some tools exist for exploration of sound properties, but the tools that explore large oral language corpus are not in an advanced stage of development, if only because oral corpora (with included original media) are often small. Designing ways to use the same format for oral corpora, makes it easier now to build a large corpus. So this means that the tools made for written language become interesting to use with oral language corpora, which explains why both groups currently work together. Several actions are in progress in the workgroup. Results, whenever it was possible, have been presented at corpus linguistic conferences.

Exploration: Methods, tools and visualisations for analysing and processing corpora

The goal of this action is to find out what methods and tools exist for analysing and processing corpora, which formats they use, and how data can be prepared for this purpose. The format used by the tools will be taken into account by the other actions (see below) so that conversion between formats and

¹⁹ <https://cocoon.huma-num.fr>

²⁰ <http://sldr.org>

description of metadata can directly target the tools that researchers use. The discussion about the actual use of the tools in the laboratories, which means how much it is used and how well mastered it is, provides information that is used to decide which technical courses and information (see above) is the most interesting to organize. This workgroup has led in the previous years to the publication of a book (Poudat and Landragin, 2017).

Formats and Metadata

Sharing corpora is highly dependent on two conditions: 1) using a common scheme for transcription and metadata format; 2) the quality of the information available in the metadata.

The workgroup has worked, for quite a few years, on using TEI as a support for oral language transcription and sharing. The work is based on the TEI Oral ISO format (International Organization for Standardization, 2016).

Good quality metadata must make it possible to describe the method used for the creation of the data and the content of the data. The basic level of metadata (Dublin-Core) used in the corpus repositories is often insufficient for fine-grained scientific purposes. This is not a question of format, or of the quality of the existing metadata. This is just because, in the linguistic data, a higher level of semantic content is required for research purposes.

Some projects do present more complex metadata, but when metadata go beyond Dublin-Core level, then the content might be different from one repository or one project to another. To avoid this and to encourage people to create fine-grained metadata, the workgroup has described a set of metadata for the analysis of oral language corpora that is considered as “minimal” in the sense that it contains enough information to create metadata of very good quality. The same work is planned to be done for written corpus metadata.

Tools for format and metadata

To make it possible for different users to use and produce the same metadata, it seems important for non-specialist users to have a tool available that is easy to use and that produces a format that can be automatically processed. We chose to develop a specific tool for this purpose whose settings can be customised and that is also easily accessible on the Internet. The tool produces a web interface in a web browser. This interface is easily changed with a configuration file. The result file is an XML file, which format is described in the configuration file. The tool is already available in its first version²¹, but is still in the testing phase. The tool edits only the specific nodes that have metadata information in an XML file and leaves other data unchanged. It can, therefore, be used as a complement to other software programmes that edit XML files. The present format of the metadata is based on the TEI. Conversions to other metadata format such as CMDI will be done automatically.

Tools for format conversion

The existence of a common format for the transcription of oral language is interesting if it is easy to produce data in this format. For this purpose, a conversion tool has been developed. It allows an easy conversion from the TEI structure to the major oral language transcription formats used in France (CLAN, ELAN, Transcriber, Praat). The development of this tool was shared with ORTOLANG. It also allows users to convert back from the TEI to the other formats. No data is lost in the conversion to TEI and the conversion back to the same format. Some information can be lost when using the tool to convert between CLAN, ELAN, Transcriber and Praat formats, as these formats have different limitations in the nature of the data that they can store. We tried to keep the data lost in conversion between application formats as minimal as possible. The software is open source and freely available²². The use of the software to aggregate multiple corpora was presented by Parisse et al. (2017).

5.2 Workgroup 2: Multimodality and new form of communication

As for the previous workgroup, the multimodality workgroup brings together colleagues who were involved in the previous consortia. One consortium was working on written data and was taking into account new communication modes and especially computer-mediated communication (CMC), and one consortium was working on multimodal oral communication, including domains such as gestures (co-verbal gestures and sign languages). The new workgroup wishes to extend its target domain outside of

²¹ <http://ct3.ortolang.fr/teimeta/readme.php>

²² <http://ct3.ortolang.fr/tei-corpo/>

the field of linguistics, for example to domains such as education or sport sciences. The goal of the workgroup is to find common points and specificities of domains that link verbal and non-verbal data, and to propose solutions that are both useful and as generic as possible. Mixing communities that work on CMC data and sign language is one of means to reach this difficult goal.

This workgroup is dedicated to the development of cutting-edge practices, either in the human interaction domain (including gesture, visual languages, co-verbal communication), or concerning computer-mediated communication and social media corpora. In the human interaction domain, one goal is to integrate representations of new data types into corpora, such as motion capture, eye-tracking, EEG. Also, for sign language studies, a goal is to find representation systems that do not need the exclusive use of gloss in another language, but can represent movement of the hand and the body, for example. These new practices call for the organisation of dedicated training sessions that will be organized in the future.

The group will also follow up on the work on representation of CMC, network communication, and all type of hybrid communication. This will mix oral and written language representation, and comments which are found in a lot of collaborative systems (e.g. collaborative edition, wiki, online press, video, etc.).

All this work calls for harmonisation of the structure of the data that is used to create and deposit corpora. This is especially true because this type of data is new and changing a lot and very rapidly. Annotation proposals exist already for multimodal (TEI proposal of Oral data, TEI for Linguist SIG²³) and for CMC (TEI for CMC SIG²⁴). However, these proposals still need to be improved to take new developments into account. This will also represent a follow-up workshop with a similar theme organized previously by the consortium *Corpus-écrits*. The current actions include mostly the production of good practice manuals for multimodal annotation and corpus creation. The workgroup is involved with current European research on CMC (see Beisswenger and Wigham, 2017), including participating in the annual CMC and Social Media Corpora for the Humanities conference series (cmc-corpora.org) and organising training sessions on structuring CMC corpora in TEI in association with CLARIN.

5.3 Workgroup 3: Multilingual and plurilingual corpora

This workgroup brings together researchers working on oral or written corpora of culture with a written tradition and researchers working on oral corpora of culture with oral tradition only. One goal of the workgroup is to share experience between the two communities, especially about the tools used for research and the theoretical aspects of the work. More specifically, the subjects under study are:

- ✓ Creation of oral or written corpora for language used by a whole country or a large community as opposed to creation of a corpus of a new language spoken by a small community: which are the best tools to be used, who are the best annotators, how can research be prioritized?
- ✓ Quantitative use of massive corpora of frequently studied languages vs. quantitative use of small corpus of unfrequently studied languages: which statistical models and methods can be used, which theoretical questions can be targeted?

The workgroup will organize training sessions for multilingual and plurilingual data, as well as training sessions for statistical processing. The group plans to promote the creation of multi-plurilingual corpora and to organize workshops on this subject. In 2017, a first panel was organized in Villejuif, France. Plans for the following years include the organisation of large size colloquium and the production of white papers on the subject. Also, the use of collaborative annotation with specialists of different languages is a promising option that needs to be developed.

5.4 Workgroup 4: Juridical information

Awareness and adherence to juridical regulations is very important for data that are subject to property rights and that might contain private or sensitive information. This workgroup has already produced white papers on the subject. These papers are freely available in the previous IRCOM website²⁵ and were produced in collaboration with other consortia. New development will be needed in 2018 to follow up new regulations, especially European regulations.

²³ <http://www.tei-c.org/Activities/SIG/CMC/>

²⁴ <http://www.tei-c.org/Activities/SIG/CMC/>

²⁵ <http://ircom.huma-num.fr/site/p.php?p=groupetravail5>

6 Relationship with CLARIN

In 2017, France joined CLARIN ERIC with an observer status. This was considered as a good opportunity to integrate the European research effort in making linguistic data freely available for everyone. French linguistics, tools, and data, would certainly benefit from being included in CLARIN - and CLARIN could also benefit from the French expertise. This would open opportunities for European collaboration and help researchers who are already involved in international projects.

A large part of the work already completed within CORLI is highly compatible with the type of work achieved in CLARIN. First of all, CORLI's main objective is to make all linguistic corpora in France available in one of the repositories that are already CLARIN centres, or are on the verge of becoming a CLARIN centre. The existence of CLARIN centres in France is not surprising because joining CLARIN was an old objective in France. So, although France is a recent CLARIN member and only an observer, many of the CLARIN principles were already effective in France.

The oldest centres are Cocoon and SLDR. Cocoon (<https://cocoon.huma-num.fr>) is a digital resource centre from and for the humanities and social sciences communities. The resources managed by this centre are speech recordings (audio or video) that are potentially accompanied by annotations and documentation. The services offered by the centre include storage, long-term preservation, integrity management, identification, description, curation and access to resources. The centre is based in Paris and is hosted by the CNRS/Huma-Num infrastructure. SLDR (<http://sldr.org>) is a centre that manages resources for spoken language and multimodal data. The centre covers storage, long-term preservation, and permanent identifiers. It is now integrated into the technical infrastructure of ORTOLANG, which is hosted at INIST²⁶. ORTOLANG is a new French centre that aims to preserve and extend the work completed at SLDR and CNRTL²⁷. Thus, the goal, as is the case for all French centres, is to cover storage, long-term preservation, and permanent identifiers. The data handled at ORTOLANG includes spoken and written language, as well as tools, lexicon, and terminologies. Any material that concerns language can potentially be preserved at ORTOLANG. The centre is based in Nancy, Aix-en-Provence, and Nanterre. The technical infrastructure is hosted by INIST in Nancy. All centres use OAI-PMH protocols, with metadata in OLAC and CMDI formats. ORTOLANG is harvested by the VLO of CLARIN.

Secondly, CORLI shares with other French initiative such as Isidore²⁸ from Huma-Num the belief that the quality of metadata is vital to the dissemination and use of language corpora. So CORLI has stressed many times how important metadata are and is working with the objective of improving the metadata.

Thirdly, good practices and sharing information are key to sharing and reusing data. This is why CORLI builds upon previous work that emphasizes the importance of sharing good practice guides, using well-known tools and metadata, and sharing information. The use of well-known formats (TEI, CMDI) is strongly encouraged.

Lastly, CORLI and a large part of the French community believe very strongly in open and free data, whenever this is possible. In most cases, people who deposit data use a CC-BY-NC licence, or another free access licence. This will make the data that is available in France also available to foreign partners.

CORLI has already established working relationships with foreign partners, for example regarding the work on oral transcription format and on CMC corpora. CORLI has a close working relationship with the already existing CLARIN centres in France (C-centres: Cocoon and SLDR; candidate B-centre: ORTOLANG). This relationship is strong through the calls for corpora finalization. For projects financed through these calls, all data must be deposited in the centres. It is also strong through the use of standard formats and metadata. For example, the metadata from the French centres is already harvested by the VLO of CLARIN. Whenever it is possible, the metadata format is CMDI. When it is not the case, a conversion to CMDI format could be organized.

7 Conclusion

The CORLI initiative has goals that are very much aligned with the objectives of CLARIN. The consortium is currently assessing the benefits of a full integration of France into CLARIN. Now, our short-

²⁶ <http://www.inist.fr/>

²⁷ <http://www.cnrtl.fr/>

²⁸ <https://www.rechercheisidore.fr/>

term aim is to explain, as best as possible, to French researchers and users of language data how the integration into CLARIN could offer opportunities to them and their research laboratories and projects, but also explain to CLARIN users in other countries what kind of material they might find in the French data that is currently available.

This has been done already twice. Once in a whole day session organized in Paris in September 2017 where researchers from other CLARIN members (Norway, Denmark, Germany, Italy) presented how CLARIN worked in their country and the implications for their own research. A second presentation was held in Montpellier in November 2017, with a conference and a poster presentation of CLARIN.

Our aim now is to organize ourselves in such a way as to be able to apply to become a CLARIN K Centre. It seems to us that the work we are currently accomplishing is very close to what a K Centre should do, and we feel like this application could be one of the ways to push France into becoming a full member of CLARIN. It would also help us to make the best use of the tools and services provided by CLARIN, and to ensure a productive dialog with the CLARIN community.

References

- [Beisswenger Michael, Wigham Ciara. R., et al. 2017] Michael Beisswenger, Ciara. R. Wigham, et al. 2017. Connecting Resources: Which Issues Have to be Solved to Integrate CMC Corpora from Heterogeneous Sources and for Different Languages? In Stemle, E. and Wigham, C.R. (2017). (eds). *Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities* (cmccorpora17). 3-4 October 2017, 52-55.
- [Bel and Gasquet 2011] Bernard Bel, Médéric Gasquet-Cyrus. 2011. Interdisciplinarity and the sharing of oral data open new perspectives to field linguistics. *Colloque de l'AFLS : Regards nouveaux sur les liens entre théories, méthodes et données en linguistique française*, Sep 2011, Nancy, France.
- [International Organization for Standardization 2016] International Organization for Standardization. 2016. *Language resource management - Transcription of spoken language* (ISO/DIS Standard 24624) - <https://www.iso.org/obp/ui/#iso:std:37338:en>
- [Jacobson, Badin and Guillaume 2015] Michel Jacobson, Flora Badin, Séverine Guillaume. 2015. Cocoon une plateforme pour la conservation et la diffusion de ressources orales en sciences humaines et sociales. *8es Journées Internationales de Linguistique de Corpus*, Sep 2015, Orléans, France. 2015, <<http://jlc2015.sciences-conf.org/>>.
- [Parisse, Benzitoun, Etienne, Liégeois 2017] Christophe Parisse, Christophe Benzitoun, Carole Etienne, Loïc Liégeois. 2017. Agrégation automatisée de corpus de français parlé, *Journées de Linguistique de Corpus*, Grenoble, Juillet.
- [Pierrel 2014] Jean-Marie Pierrel. 2014. ORTOLANG : une infrastructure de mutualisation de ressources linguistiques écrites et orales. *Actes de TALN 2014*, Marseille, France <http://talnarchives.atala.org/TALN/TALN-2014/taln-2014-demo-001.pdf>.
- [Poudat and Landragin 2017] Céline Poudat, and Frédéric Landragin. 2017. *Explorer un corpus textuel : Méthodes - pratiques - outils*. De Boeck Supérieur, 240pp.

Something will be connected

- Semantic mapping from CMDI to Parthenos Entities

Matej Ďurčo
ACDH-OEAW
Vienna, Austria
matej.durco
@oeaw.ac.at

Matteo Lorenzini
ACDH-OEAW
matteo.lorenzini
@oeaw.ac.at

Go Sugimoto
ACDH-OEAW
go.sugimoto
@oeaw.ac.at

Abstract

The Parthenos project aims at pooling resources from existing infrastructures of the broad cultural heritage and humanities cluster. Central to this effort is the common semantic framework - Parthenos Entities - that shall serve as a target data model for mapping of metadata about resources from participating infrastructures. Acting as a representative of the linguistic domain, CLARIN will deliver metadata about language resources. Within the Parthenos project, separate provisions are foreseen for the mapping task. However, given the complexity of CLARIN's underlying metadata model (CMDI), traditional one-to-one schema mapping is not applicable and an alternative conceptual and technical approach is required. This paper presents the current mapping solution and points out a number of issues identified during the process partly perpetuated from the ongoing metadata quality discussion within CLARIN.

1 Introduction

Parthenos¹ (Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies) is a project funded by the European Commission's Horizon 2020 framework programme that started May 2015 and runs for four years. The project empowers digital research in the fields of history, language studies, cultural heritage, archaeology, and related fields across the (digital) humanities. It brings together several existing research infrastructures to make it easier to find, use and combine information about main entities involved in the research process from different domains, such as datasets, services or actors. The project aims to establish interoperability in humanities domain, building a bridge between the existing European Research Infrastructure Consortia including CLARIN², DARIAH³, EHRI⁴, ARIADNE⁵, CENDARI⁶, CHARISMA⁷, and IPERION-CH⁸. One of the biggest challenges is the aggregation of heterogeneous data coming from such different research infrastructures into a common semantic framework called Parthenos Entities model (PE).

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

¹ <http://www.parthenos-project.eu/>

² <https://www.clarin.eu/>

³ <http://www.dariah.eu/>

⁴ <https://ehri-project.eu/>

⁵ <http://www.ariadne-infrastructure.eu/>

⁶ <http://www.cendari.eu/>

⁷ <http://www.charismaproject.eu/>

⁸ <http://www.iperionch.eu/home>

CLARIN is a major partner in Parthenos with regard to language resources and language studies in general. It has been operating one of the biggest catalogues of language resources in Europe, Virtual Language Observatory (VLO)⁹, since 2010 (Van Uytvanck et al., 2012; Eckart et al., 2015). It aggregates the metadata about the resources from over 60 data providers, containing more than 1.6 million records. The backbone of the VLO is CMDI¹⁰ (Component Metadata Infrastructure) (Broeder et al., 2011; Goosen et al., 2014) which offers a flexible standardised framework to facilitate formalised descriptions for a wide range of resources, aimed at fostering resource discovery within the linguistic domain and beyond. In order to deliver the information about CLARIN resources to Parthenos, it is required to map the metadata schemas defined in CMDI to PE. This paper presents an approach adopted for this mapping and highlights the encountered problems.

2 Underlying standards and components

In the following, we introduce the standards and components that play a role in the mapping task.

2.1 Component Metadata Infrastructures (CMDI)

CMDI provides a framework for creating and (re)using self-defined metadata schemas in order to meet various needs of data providers, and yet to set a mechanism to aggregate and unify heterogeneous metadata of language resources. It relies on a modular model of reusable components, which are assembled together to define profiles serving as a blueprint for custom schemas to be used for new metadata creation. The CMDI Component Registry¹¹ (Broeder et al., 2010) was created as a central online environment for the creation and discovery of metadata components and profiles to promote their reuse and sharing. The registry contains all CMD components and profiles used to describe metadata in VLO, currently holding around 1.000 components and around 200 profiles. To enable semantic interoperability between the various profiles, fields in the components are linked to well-defined concepts, primarily drawn from the CLARIN Concept Registry (CCR¹²) (Schoorman et al., 2015), a separate module of CMDI, which allows to openly specify stable definitions of semantic concepts.

2.2 Common Semantic Model – Parthenos Entities Model (PE)

Parthenos proposes a common ontological model, Parthenos Entities, to be able to describe, in a generic manner, basic characteristics of all main entities involved in the knowledge generation process as encountered in the source metadata records, irrespective of the peculiarities of individual source formats. The model is composed of four main entities:

- *PE18 Dataset*: defined in PE model, sets or collections of data, records or information (provided by participating infrastructures) that constitute distinct units of information in the knowledge generation process.
- *E39 Actor*: defined in main CIDOC CRM ontology, is an individual or a group that exercises agency in the knowledge generation process, for which they are responsible.
- *PE 1 Service*: defined in PE model, represents the ability and willingness of an actor to execute on demand by a client certain activities of specific benefit to the client. The service includes all auxiliary abilities of the same actor to execute the respective activities, but not services provided by third parties in the course of their service provisioning.
- *D14 Software*: defined in CRMdig extension, represents an artefact that can be executed on a computer to perform specific operations. In particular, software is the necessary information to process datasets algorithmically and to integrate datasets in a collaborative infrastructure.

⁹ <https://vlo.clarin.eu>

¹⁰ <https://www.clarin.eu/content/component-metadata>

¹¹ <https://catalog.clarin.eu/ds/ComponentRegistry/>

¹² <https://www.clarin.eu/ccr>

The categorical description of these entities is defined by a minimal metadata set. The minimal metadata set is not meant to represent all the information present in the source metadata, but solely to establish an identity for any entity mapped from the graph, i.e. if it is the same or different from another aggregated entity. Thus the mappings and transformations to the PE are lossy by design. The PE model is not intended to capture all the structure and semantics of CMDI, let alone to replace CMDI or any other of the source formats. The goal of Parthenos is merely harmonisation of basic information about resources aggregated from different research infrastructures to enable resource discovery in a unified manner.

The PE model is formalised based on CIDOC CRM and its extension CRMdig. The former serves to capture the information about cultural heritage and the latter to describe the provenance of the information and digitisation process.

The CIDOC CRM, which became an ISO standard in 2006, is an ontology comprising 86 classes and 138 properties which provides definitions and a formal structure for describing the implicit and explicit concepts and relationship used in cultural heritage documentation. It is also intended to be used as a top-level ontology to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information (ICOM/CIDOC CRM Special Interest Group 2017). It does this by defining very general concepts like space, time, object, event, activity, etc., which are independent of a particular problem or domain, while providing also cultural heritage specific properties such as “curated”, “used specific technique” and “has current keeper”. The CRMdig¹³, developed as compatible extension of CIDOC CRM, is an ontology and a RDF Schema for encoding metadata about the steps and methods of production (“provenance”) of digitization products and synthetic digital representations such as 2D, 3D or even animated models. The PE model additionally defines 33 classes and 37 properties as specialisations of entities defined in the base ontology, though in the target model both the additional entities and selected entities and properties from CIDOC CRM and CRMdig are used. The adoption of CIDOC CRM and CRMdig as a baseline of the PE enables us to maximise the data interoperability and thus support resource discovery across different cultural heritage and humanities domains.

2.3 Parthenos infrastructure components

Within Parthenos, the 3M mapping tool (Minadakis et al. 2015) is employed to collaboratively define mappings from different data models encountered in the participating research infrastructures into one common model, the PE. 3M is an online open source tool for managing the mapping definition files expressed in X3ML¹⁴, an XML-based schema for describing schema mappings from XML to RDF (see Listing 1 for a sample). 3M assists the users during the mapping definition process with a human-friendly user interface that suggests and validates the user input against the source and target schemas. The structure of an X3ML file consists of: 1) a header with basic provenance information such as the date of creation and the author of the mapping file; 2) a series of mappings, each containing a domain and a number of ‘link’ elements with a ‘path’ and a ‘range’ to map the source values to. Each link describes the relation (path) of the domain entity to the corresponding range entity.

Listing 1. Sample mapping in X3ML format

```
<mapping>
  <domain>
    <source_node>/cmd:CMD/cmd:Resources/cmd:ResourceProxyList/cmd:Re-
sourceProxy/cmd:ResourceRef</source_node>
    <target_node>
      <entity>
        <type>crmpe:PE29_Access_Point</type>
        <instance_generator name="UUID"/>
      </entity>
    </target_node>
  </domain>
  <link>
    <path>
      <source_relation>
        <relation>/cmd:ResourceType</relation>
      </source_relation>
```

¹³ http://www.ics.forth.gr/isl/index_main.php?l=e&c=656

¹⁴ <https://github.com/delving/x3ml>


```

    <target_relation>
      <relationship>crm:P28_has_type</relationship>
    </target_relation>
  </path>
  <range>
    <source_node>/cmd:ResourceType</source_node>
    <target_node>
      <entity>
        <type>crm:E55_Type</type>
        <instance_generator name="ConceptURI_2step"> ...
      </entity>
    </target_node>
  </range>
</link>
</mapping>

```

These mappings serve as input for the customisable aggregation infrastructure, D-Net¹⁵, which allows to select and configure the needed services and easily combine them to form complex automated data processing workflows. Its scalability and reliability are proven as it powers a number of aggregation platform, for example, the huge research publication portal OpenAire¹⁶. For the Parthenos project, the 3M engine has been integrated into D-Net infrastructure to support the aggregation of metadata records from the source research infrastructures based on mappings expressed in X3ML language. D-Net itself is integrated into the hybrid data infrastructure d4science¹⁷, Parthenos' central content and service provisioning infrastructure based on the software solution gCube¹⁸. It forms the Parthenos Content Cloud Framework (CCF), the component responsible for the whole aggregation, transformation, storage and indexing workflow. In this framework aggregated and transformed metadata records are transformed into different formats and ingested into multiple storage and indexing components which serve as end-points for resource discovery applications: a) as RDF adhering to PE model into a Virtuoso¹⁹ triple store, allowing full-fledged complex SPARQL²⁰ queries on the whole RDF graph, b) flattened into indices of an Apache Solr instance for full-text search systems and c) as serialized RDF available via an OAI-PMH²¹ endpoint. Figure 1 depicts the whole metadata aggregation and provisioning infrastructure employed in Parthenos.

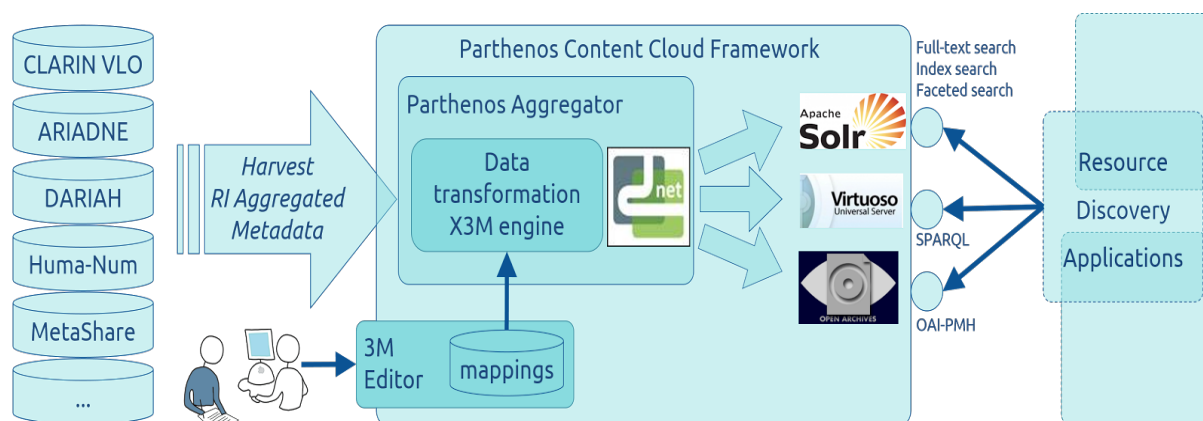


Figure 1. Diagram of the Parthenos metadata aggregation and provisioning infrastructure

¹⁵ <http://d-net.research-infrastructures.eu/>

¹⁶ <https://www.openaire.eu/search/find>

¹⁷ <https://www.d4science.org/>

¹⁸ <http://gcube-system.org/>

¹⁹ <https://virtuoso.openlinksw.com/>

²⁰ <https://www.w3.org/TR/rdf-sparql-query/>

²¹ <https://www.openarchives.org/pmh/>

SOURCE ↔		TARGET ↔	
D	../cmd:CMD	PE22_Persistent_Dataset	
P	../cmd:MdSelfLink	P1_is_identified_by	
R	../cmd:MdSelfLink	E42_Identifier	
P	../cmd:MdCreator	P94i_was_created_by	
R	../cmd:MdCreator	E65_Creation [create1]	
		P14_carried_out_by	
		E39_Actor	
P	../cmd:MdCreationDate	P94i_was_created_by	
		E65_Creation [create1]	
		P4_has_time-span	
		E52_Time-Span	
		P82_at_some_time_within	
R	../cmd:MdCreationDate	rdf-schema#Literal	
P	../cmd:MdCollectionDisplayName	PP23i_is_dataset_part_of	
R	../cmd:MdCollectionDisplayName	PE24_Volatile_Dataset	
P	../cmd:MdSelfLink	P129_is_about	
R	../cmd:MdSelfLink	E73_Information_Object	
P	../cmdp:TextCorpusProfile	PP39_is_metadata_for	
R	../cmdp:TextCorpusProfile	PE24_Volatile_Dataset [data1]	
P	cmd:Resources	PP39_is_metadata_for	
		PE24_Volatile_Dataset [data1]	
		PP8i_is_dataset_hosted_by	
R	cmd:Resources	PE15_Data_E-Service	

SOURCE ↔		TARGET ↔	
D	../cmd:Resources	PE15_Data_E-Service	
P	../cmd:ResourceProxy	PP28_has_designated_access_point	
R	../cmd:ResourceProxy	PE29_Access_Point	

SOURCE ↔		TARGET ↔	
D	../cmd:ResourceProxy	PE29_Access_Point	
P	cmd:ResourceType	P2_has_type	
R	cmd:ResourceType	E55_Type	

Figure 2. Screenshot of the 3M mapping tool

3 Mapping

3.1 Method

The default mapping approach in the Parthenos project is a 1:1 crosswalk between a “local” source schema specific to individual research infrastructure and the target schema (PE). However, as outlined in the previous section, CMDI is not just one schema but a framework for creating and reusing schemas. In fact, currently more than 200 different schemas have been defined. It is, therefore, not feasible to define the mapping in this traditional way. Instead we take advantage of the mechanism already employed in the VLO, which is a mapping relying on the built-in semantic interoperability layer, that is, the semantic binding of the structural elements of CMD profiles to well-defined concepts. The developed mapping solution aims to identify PE properties which are (near) equivalent to the concepts of CCR (Figure 3. Mapping Definition Phase), to derive XPath²² patterns for any profile by matching concepts in the corresponding XML schema (Figure 3. Profile Pre-processing Phase), and finally to use the XPaths to extract values from actual CMD instances (records) to generate a corresponding entity description adhering to the PE model (Figure 3. Aggregation Phase).

While the basic mechanism is similar to the one applied for populating the VLO, the specific context is quite different, requiring a new custom solution. In particular, the question is how to integrate the automatic mapping step, i.e. the resolution of concepts to appropriate XPaths, into the foreseen aggregation pipeline, aimed at extracting values from source metadata and generating the target structured records. Following scenarios were considered: a) the VLO software component responsible for data

²² <https://www.w3.org/TR/xpath/>

transformation and ingestion can become a part of the D-Net aggregation infrastructure (with some adjustments), b) custom XSL stylesheets (natively supported by D-Net) can be generated, or c) the generated mapping is converted to a format required by X3ML, pushing all processing logic to the Parthenos side. We chose the third option and developed a simple java application²³ that does not do the actual transformation of the records, but only generates the specific X3ML-mapping files, based on a mapping file template containing multiple concepts and fall-back XPathS (as is the case in the concepts to facets file serving as input for VLO-importer) in specific locations to be resolved against a given individual CMD profile. The entire procedure is depicted in Figure 3.

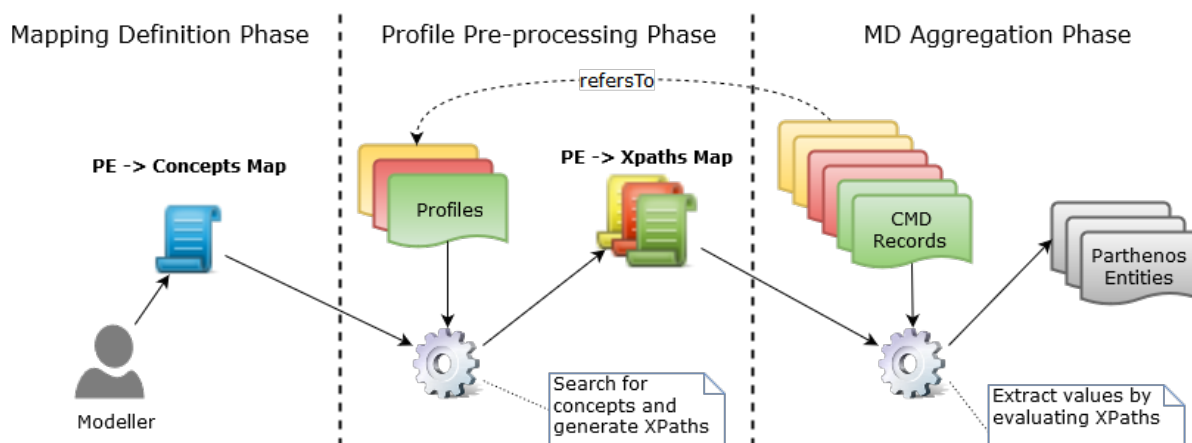


Figure 3. CMDI to Parthenos Entities mapping generator algorithm

3.2 Mapping decisions

There is a broad leeway in how the source data can be expressed in the target model (PE). To ensure conceptually sound mappings as well as a harmonized approach among the infrastructures, a number of modelling decisions were taken. We present some of them below in Table 1.

Based on these general modelling conventions, we defined mappings from CMD schemas to PE in an iterative collaborative process. Following the general model of the CMDI framework, we distinguish between global mappings of the generic CMD envelope applicable to all CMD records (selected examples in Table 2) and ‘local’ mappings custom to individual CMD profiles (Table 3).

²³ https://github.com/acdh-oeaw/parthenos_mapping

Type of information	PE	Note
Values	E40_Legal_Body → P3_has_note → rdf-schema#Literal, E35_Title → P1_is_identified_by → E41_Appellation → rdfs:label	If the entity refers to a literal value, the referred data is mapped as rdf:literal. If the entity refers to a value string, the referred data is mapped as rdfs:label
Publication date	PE24_Volatile_Dataset → crm:P94i_was_created_by → crm:E65_Creation → crm:P4_has_time-span → crm:E52_Time-Span → crm:P82_at_some_time_within → http://www.w3.org/2000/01/rdf-schema#Literal	interpreted as the creation date of the resource PE24_Volatile_Dataset or as the date on which curaton of the dataset begins
Title	crmpe:PE24_Volatile_Dataset → crm:P1_is_identified_by → crm:E41_Appellation → rdfs:label	
Email, phone	E74_Group/crm:E40_Legal_Body/crm:E21_Person → crm:P76_has_contact_point → crm:E51_Contact_Point [crm:E55_Type = "parthenos-type:email"]	Further specify type of contact point with E55_Type
URL, handle	crmpe:PE22_Persistent_Dataset → crm:P1_is_identified_by → crm:E42_Identifier	URL to encode is typed as crm:E42_Identifier

Table 1. Selected general modelling decisions

CMDI XPath	PE	Note
/cmd:CMD	crmpe:PE22_Persistent_Dataset	Metadata record itself also represented as first-class citizen
./cmd:Header	PE22 → crmdig:L11i_was_output_for → D7_Digital_Machine_Event	Creation of the record as an event
./cmd:Header	crmdig:D7_Digital_Machine_Event	
cmd:MdCreationDate	D7 → crm:P4_has_time_span → crm:E52_Time_Span → crm:P82_at_some_time_within → rdf-schema#Literal	When did the creation event happen
cmd:MdCreator	D7 → crmdig:L23_used_software_... → crmpe:PE21_Persistent_Software	Field cmd:MdCreator is very heterogeneous containing references to persons, institutions, projects as well as software

		applied for generation. Proposed mapping reflects the last variant.
cmd:MdProfile	D7 → crmdig:L23_used_software_... → crmpe:PE38_Schema	CMD schema as the “software” used in the creation event
//cmd:Components /cmdp:*	PE22 → crmpe:pp39_is_metadata_for → crmpe:PE24_Volatile Dataset	Explicit aboutness-relation between record and resource
→ cmd:ResourceProxy	→ crmpe:pp39_is_metadata_for → crmpe:PE24_Volatile Dataset → crmpe:PP8i_is_dataset_hosted_by → crmpe:PE15_Data_E-Service	Relation between the one CMD record to potentially many described resources
→ cmd:Header/ cmd:MdCollectionDisplayName	crmpe:PE24_Volatile_Dataset [resource!] → crmpe:PP23i_is_dataset_part_of → crmpe:PE24_Volatile_Dataset → crm:P1_is_identified_by → crm:E41_Appellation	Part of relation between the resource (not the metadata record!) and a collection.

Table 2. Selected global mappings

CMDI	PE
../cmdp:TextCorpusProfile	crmpe:PE24_Volatile_Dataset
→ cmdp:Name	→ crm:P1_is_identified_by → crm:E41_Appellation
→ cmdp:Title	→ crm:P1_is_identified_by → crm:E35_Title
→ cmdp:Owner	→ crm:P105_right_held_by → crm:E40_Legal_Body
→ cmdp:Description	→ crm:P3_has_note → rdfs-schema#Literal
→ cmdp:Project	→ crm:P94i_was_created_by → crm:E65_Creation → crmpe:PP43i_is_project_activity_supported_by → crmpe:PE35_Project
→ cmdp:Availability	→ crm:P129i_is_subject_of → crm:E30_Right → crm:P3_has_note → rdf-schema#Literal
../cmdp:Access	crmpe:PE15_Data_E-Service
→ cmdp:Contact	crmpe:PP2_provided_by → crm:E40_Legal_Body

Table 3. Examples of local mappings

3.3 Current status

Based on the experience we gathered while manually defining mappings in X3ML for 3 sample profiles (teiHeader²⁴, TextCorpusProfile²⁵, and OLAC-DcmiTerms²⁶), we derived two templates as expected by the mapping generator, one for datasets, the other for services, and furnished these with the most frequently referred concepts to be resolved against the individual schemas. These manually defined mappings were applied on a sample collection of roughly 3.000 CMD records, which were processed through the Content Cloud Framework and made available as PE conformant RDF.

In a next step, we identified all CMD profiles with records in a recent VLO data dump, and based on the template files we automatically generated maps for all these currently employed profiles.

The initial transformation of the small sample dataset is an important milestone demonstrating the feasibility of the approach and established connectivity. However, it also revealed many issues on various levels of the aggregation process, prompting a feedback loop to fine-tune the individual steps of the transformation workflow: a) the generation of profile-specific mappings, b) mapping from CMDI to PE; c) normalisation, harmonisation of values; d) transformation and ingest from PE to a flat index-search engine (Solr). Finally, there is also a possibility that the problem already lies in the source data (CMD records) as delivered by the original service providers (cf. section 4 Issues and challenges).

4 Issues and Challenges

During the initial mapping process, we encountered several issues which will have adverse effect on the discovery and exploitation of the aggregated data. A major issue arising in the mapping task is the oftentimes ambiguous or underspecified semantics of numerous structures/expressions used in CMDI. The foremost example is `cmd:ResourceProxy`. One metadata record can contain a number of `ResourceProxies` (`cmd:ResourceProxyList{1}/cmd:ResourceProxy{1...n}`) expressing three different semantics:

1. Different access points for the same resource. This case is covered by a specific mapping of the `cmd:ResourceRef` elements as typed `PE29_Access_Point` entities.
2. The record represents a collection and all `ResourceProxies` point to other metadata records describing the items of the collection. In this case, the relation between the collection and its members can be expressed using `crmpe:PP23i_is_dataset_part_of`.
3. The record represents a number of distinct resources. In this case the `id`-attribute can be referenced from the corresponding XML-elements in the `cmd:Components` mapping block. This case is not yet fully covered by the mapping provisions.

This setup is by design and is algorithmically distinguishable, but it requires specific provisions in the mapping task, i.e. injection of procedural processing in the mapping process beyond declarative cross-walk definitions. An evaluation on a sample recent VLO data dump with 879.497 CMD records yielded that there are 685.832 records of case 1, 1.421 records of case 2 and 193.662 records of case 3.

Another substantial shortcoming in CMDI semantics is unclear statements about the persistent nature of the described resource (i.e. can the resource change, or is it immutable?), and the mingling of information about a provided web service and the underlying software.

PE makes a clear distinction between Software and Service (D14 vs. PE1 or PE8 for E-Service), but it is partly impossible to derive the difference from CMD records. The PE also distinguishes between a Volatile and a Persistent Dataset (PE24 vs. PE22). While the former is defined as “dataset that are changed without notice or archiving of intermediate states but maintained by an instance of PE12 Data curating Service.” and “are typically whole databases or mash-ups with active data feeds”, the scope of the latter is “datasets that contain collections of data, records or information kept as a persistent unit of information in the knowledge generation process from primary records up to any level of aggregation

²⁴ clarin.eu:cr1:p_1381926654438

²⁵ clarin.eu:cr1:p_1271859438164

²⁶ clarin.eu:cr1:p_1288172614026

or integration”. Also in this case, it is sometime impossible to decide to which class given resource belongs, as the original metadata was often created without concerning such difference.

An example of problematic semantics on the instance level is the different values in the `cmd:MdCreator` element with a mix of around 300 distinct person names, projects, collections, software solutions and scripts involved in the creation of the records²⁷.

In addition, we have to deal with implicit entities. For example, although there is a lot of information about actors encoded within CMD records (e.g. publisher, organisation responsible for creation of the resource, rights holder etc.), it needs to be extracted to generate the corresponding Actor entities. Here, we are confronted with a long standing issue in CMDI metadata - the variability of descriptors. It is caused inter alia by not using identifiers, but rather just string values to denote entities, like organisations. As a consequence, we are not able to fully identify the same entities described in different variations of vocabularies (e.g. “Max Planck Institut” and “MPI” may or may not be the same entity). The normalisation of values is on-going process within the CLARIN’s metadata curation taskforce.

In addition, we encountered information gaps. Even if a record contains information about the corresponding actor, in most cases it is not sufficient to build a full description, such as the hierarchy of organisations. It needs to be either collected from other sources, or curated manually. Nonetheless, in the specific case of organisations, we can build on the work done in the CLAVAS project²⁸, where organisations from the VLO were extracted, manually curated, and published as a vocabulary.

Moreover, the well-known problems of metadata quality under discussion in the context of CLARIN resurface in the mapping task. Of note among these are, in particular the (facet) coverage (King et al., 2015), i.e. missing values for specific aspects of a resource description, and the variability of values, especially those denoting entities like organisations (Ostojic et al., 2016). Both issues have strong influence on the quality of the resulting harmonized metadata and dramatically hamper the recall. The latter is especially problematic given the goal of the overall Parthenos mapping task to establish identities for main entities, and make also actors (e.g. organisations and persons) first-class citizen in the CIDOC-PE data space.

5 Conclusion

In this paper, we describe the ongoing work on mapping CMDI metadata to Parthenos Entities model. The mapping strategy relies on semantic interoperability mechanisms established in the CLARIN infrastructure. We introduced an intermediate processing step, in which a hand-crafted template file furnished with CCR concepts is expanded by a dedicated small utility Java application into a valid X3ML mapping file with XPaths corresponding to given concepts relative to a specific CMD schema. These generated mapping files are used by the integrated transformation framework D-Net to extract values from CMD instances and to generate an entity description in PE model. After the crafting of the template file based on two profiles, mapping files for all profiles encountered in the VLO were generated.

During our mapping effort, several problems were recognised. One of the major issues was the semantic ambiguity and lack of explicit statements regarding crucial aspects of the described resources in numerous structures in the CMD records, for instance concerning the distinction between a software and a service or between a volatile and a persistent dataset. In addition, well-known metadata quality issues such as missing values and variability of values cause mapping errors.

We strongly believe, that the task of mapping the CLARIN metadata to the PE model is not only an academic exercise and a one-way contribution, but also that CLARIN’s metadata infrastructure and community can benefit greatly from expressing the information about the resources in a well-established high-level conceptual model like CIDOC CRM. Conversely, the process of mapping the complex CMDI metadata also allows us to identify potential omissions in the PE model and has proven useful for the modelling work.

²⁷ https://github.com/acdh-oeaw/parthenos_mapping/blob/master/cmd_utils/mdCreator_values.txt

²⁸ <https://openskos.meertens.knaw.nl/clavas/>

The mappings between PE and other schemas of different infrastructures are in the final phase. When our mapping is completed, Parthenos will be able to harvest and aggregate metadata from all the participating infrastructures, offering the users access to a comprehensive aggregation of datasets and tools in cultural heritage for their research.

References

- [Broeder et al. 2010] D. Broeder, M. Kemps-Snijders, D. Van Uytvanck, M. Windhouwer, P. Withers, P. Wittenburg, and C. Zinn. 2010. A data category registry-and component-based metadata framework. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation*. Pp. 43–47.
- [Broeder et al. 2011] D. Broeder, O. Schonefeld, T. Trippel, D. Van Uytvanck, and A. Witt. 2011. A pragmatic approach to XML interoperability—the Component Metadata Infrastructure (CMDI). In *Proceedings of Balisage: The Markup Conference 2011*. Balisage Series on Markup Technologies, volume 7.
- [Doerr 2003] M. Doerr. 2003. The CIDOC Conceptual Reference Module: an ontological approach to semantic interoperability of metadata. In *AI magazine*, 24(3):75.
- [Eckart et al. 2015] T. Eckart, A. Helwig, and T. Goosen. 2015. Influence of Interface Design on User Behaviour in the VLO. In *CLARIN Annual Conference 2015 Book of Abstracts*. <https://www.clarin.eu/sites/default/files/book%20of%20abstracts%202015.pdf>
- [FORTH-ICS 2017] PARTHENOS Entities: Research Infrastructure Model V2.0.
- [Goosen et al.2014] T. Goosen, M. Windhouwer, O. Ohren, A. Herold, T. Eckart, M. Ďurčo and O. Schonefeld. 2014. CMDI 1.2: Improvements in the CLARIN Component Metadata Infrastructure. In *Selected Papers from the CLARIN 2014 Conference*. Pp. 36–53.
- [ICOM/CIDOC CRM Special Interest Group 2017] Definition of the CIDOC Conceptual Reference Model Version 6.2. 3 October 2017.
http://www.CIDOC CRM.org/sites/default/files/2017-12-30%23CIDOC%20CRM_v6.2.3_esIP.pdf
- [King et al. 2016] M. King, D. Ostojic, M. Ďurčo, and G. Sugimoto. 2016. Variability of the Facet Values in the VLO—a Case for Metadata Curation. In *Selected Papers from the CLARIN Annual Conference 2015*, October 14–16, 2015, Wrocław, Poland (pp. 25–44) Linköping University Electronic Press. <http://www.ep.liu.se/ecp/123/003/ecp15123003.pdf>
- [Minadakis et al. 2015] N. Minadakis, Y. Marketakis, H. Kondylakis, G. Flouris, M. Theodoridou, M. Doerr, and G. de Jong. 2015. X3ML framework: an effective suite for supporting data mappings. In: *Workshop for Extending, Mapping and Focusing the CRM—co-located with TPD’2015*
- [Ostojic et al. 2016] D. Ostojic, G. Sugimoto, and M. Ďurčo. 2016. Curation module in action - preliminary findings on VLO metadata quality. Retrieved from https://www.clarin.eu/sites/default/files/ostojic-et-al-CLARIN2016_paper_22.pdf
- [Schoorman et al. 2015] I. Schoorman, M. Windhouwer, O. Ohren, and D. Zeman. 2015. CLARIN concept registry: the new semantic registry replacing ISOcat. In *CLARIN Annual Conference 2015*. Pp. 80–83.
- [Van Uytvanck 2012] D. Van Uytvanck, H. Stehouwer, and L. Lampen. 2012. Semantic metadata mapping in practice: The Virtual Language Observatory. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*. Pp. 1029–1034.

A Bridge from EUDAT's B2DROP cloud service to CLARIN's Language Resource Switchboard

Claus Zinn

Seminar für Sprachwissenschaft

Universität Tübingen

`claus.zinn@uni-tuebingen.de`

Abstract

The Language Resource Switchboard is becoming a central pillar in the CLARIN infrastructure as it helps researchers to connect resources with tools that can process them in one way or another. Languages resources can be found in different places, and ideally, the switchboard is available nearby. Resources located at users' desktop computers can simply be uploaded to the switchboard, and resources found in CLARIN's Virtual Language Observatory can simply be sent to the switchboard by a simple click. Until now, the switchboard was only indirectly accessible for resources stored in the cloud. Here, users had to download a resource from their cloud storage to their desktop device before uploading it again to the switchboard to find applicable tools, which is tedious. In this paper, we describe how we linked EUDAT's B2DROP cloud service to the switchboard, giving users the capability to directly launch the switchboard with a resource from their B2DROP account. Also, we describe the usage of B2DROP to support the switchboard's back-end for intermediate file storage. The reported work makes a link to another infrastructure, and hence, facilitates and promotes the provision of complementary services to CLARIN members. We believe the cooperation between CLARIN and EUDAT to be of mutual benefit. On the one hand, our bridge makes the use of the generic cloud storage service from EUDAT more attractive to CLARIN members so that they are encouraged to use B2DROP rather than another cloud provider. On the other hand, it encourages EUDAT users to try out and profit from the CLARIN tool space, which in turn will challenge the tool providers to cope with an increased demand, and potentially new user requirements.

1 Introduction

The CLARIN Language Resource Switchboard (LRS) aims at bridging the gap between language-related resources and tools that can deal with these resources in one way or another. For a given resource, the LRS identifies all tools that can process the resource; users can then select and invoke the tool of their choosing. By invoking the tool, all relevant information about the resource is passed onto the tool, and the tool opens with most information gathered by the switchboard. This makes it easy for users to identify the right tools for their resource, but also to use the chosen tool in the most effective way possible.

The EUDAT Collaborative Data Infrastructure aims at providing services that seek to address the full life-cycle of research data. EUDAT's services include, among others, B2DROP (sync and exchange of research data), B2SHARE (store and share research data), B2FIND (find research data), and B2HANDLE (register your research data). B2DROP is directed at scientists to store and exchange data easily and to facilitate data synchronisation between cloud storage and desktop computers. EUDAT services are designed, built and implemented based on user community requirements. The CLARIN consortium contributes to EUDAT as one of the main communities in the Social Sciences and Humanities.

In this paper, we describe the use of B2DROP in the CLARIN Language Resource Switchboard. In the main use case, we anticipate an individual researcher or a small team of researchers to use B2DROP as cloud storage for language-related resources. The researcher(s) will want to work with and analyse the

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

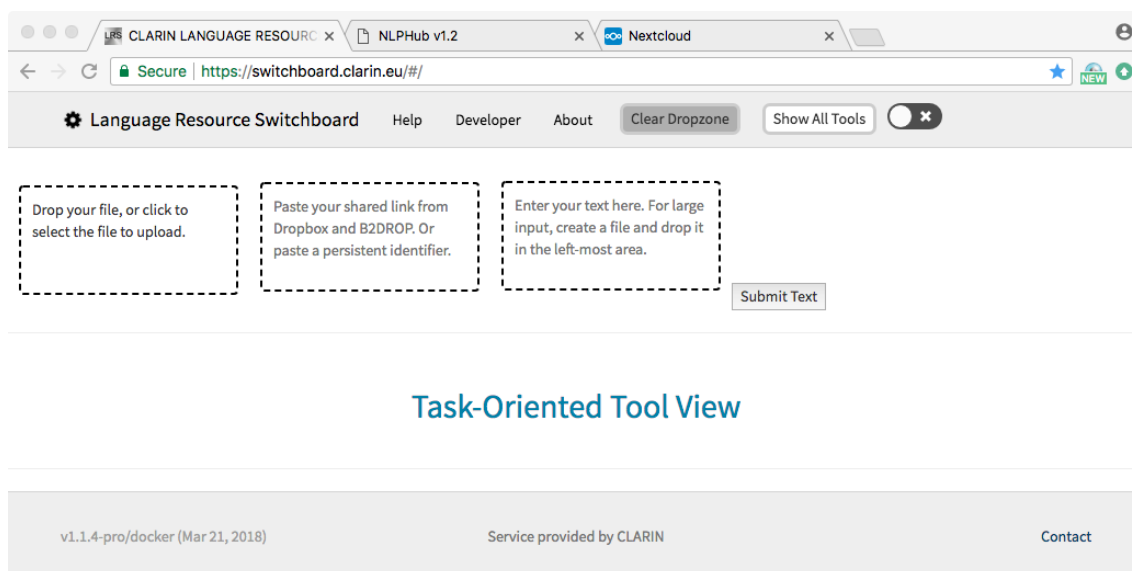


Figure 1: The CLARIN Switchboard’s Main Page

resources using community-specific tools of the CLARIN tool space. From the B2DROP user interface, the researcher(s) will want to easily transfer a given resource to the LRS, which in turn suggests tools to process the resource. In a second use case, we describe the use of B2DROP as a technical vehicle for intermediate cloud storage, supporting a crucial aspect of the LRS’ back-end implementation.

2 Background

2.1 The Language Resource Switchboard

The LR Switchboard (LRS) has been developed within the CLARIN-PLUS project (Zinn, 2016). The development of the LRS started as a browser-based stand-alone version.¹ Here, users simply upload their resource from their desktop machine to the browser, which is then temporarily stored on a file server at the Max Planck Computing and Data Facility (MPCDF)². With the help of the Apache Tika library³, the LRS then detects the resource’s language and media type, and it uses this information to identify all tools registered with the LRS that can process the resource. The list of applicable tools is sorted along typical processing tasks (*e.g.*, tokenization, dependency parsing, named entity recognition) and shown to the user. When the user selects a tool from the list, the LRS constructs a URL that points to the tool’s web location and also encodes the tool’s parameters such as a reference to the storage location of the resource as well as the resource’s language or an analysis id. The LRS then directs the browser to open the URL in a new browser tab. For a tool to be connected to the LRS, it must be reachable under the given base URL and capable of interpreting and processing all URL-encoded parameters passed during tool invocation. In particular, the tool will need to download the resource from the storage location that is encoded in the URL. The tool is then updating its internal model and graphical view accordingly.⁴ Many tools leave users with no configuration options; here users simply press the start button to invoke the tool. Other tools have a richer interface where users can choose from many options before starting the tool.

The LRS has also been connected to the Virtual Language Observatory⁵ (VLO), the main CLARIN site for searching language-related resources via CMDI-based metadata (Uytvanck et al., 2012). When users find a resource of interest in the VLO, they can start the LRS directly from VLO’s resource viewer.

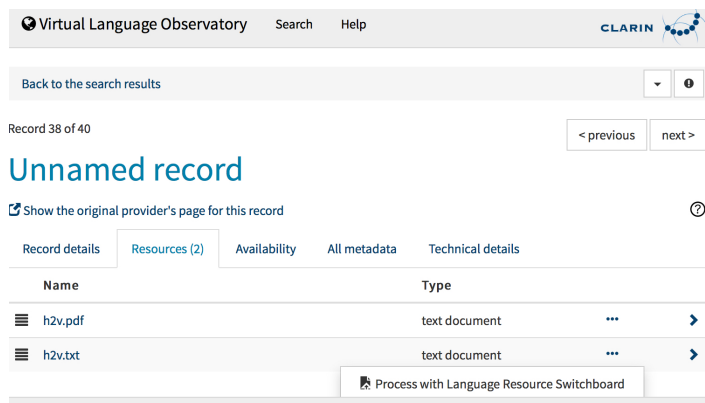
¹See <http://switchboard.clarin.eu>.

²See <http://www.mpcdf.mpg.de>.

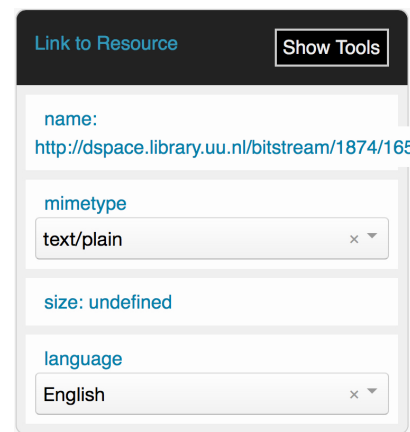
³See <https://tika.apache.org>.

⁴Many tools are capable of displaying the resource’s content in a text-area, which reassures users that the resource has been successfully passed to the tool.

⁵See <http://vlo.clarin.eu>.



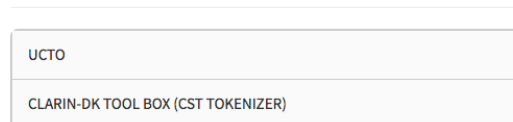
(a) The VLO – LRS Interface.



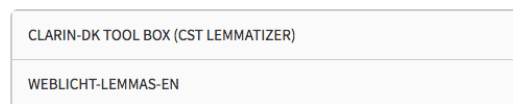
(b) The LRS Resource Pane.

Task-Oriented Tool View

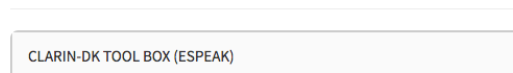
Tokenisation



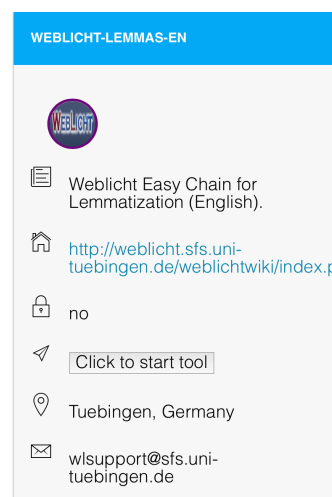
Lemmatization



Voice Synthesis



(c) The LRS Task Oriented View.



(d) The LRS Tool Detail View.

Figure 2: The LRS in Action.

Here, the VLO passes to the LRS data that is read from the CMDI metadata record of the resource: a URL pointing to the resource as well as information about the resource's language and media type. Given these pieces of information, there is no need for the switchboard itself do access the resource; the switchboard trusts the information given by the VLO, and does not derive media type and language itself.

Of course, all tools connected to the switchboard need access to the resource in order to process it. Here, resource providers must trust tool providers to handle their data with care. Switchboard users should be aware of data privacy issues, and when in doubt, they should not share sensitive data with the switchboard and its associated tools. Fig. 1 displays the main user interface of the CLARIN switchboard when it is called in stand-alone mode. At the centre, users have access to three input devices. A file drag & drop mechanism (left), a link drop mechanism (middle), and a text area, where users can enter or paste

textual data (right). In the first and third input method, the input is uploaded to a file storage server so that the tools connected to the LRS can access the data from this server. In the second approach, the data is already located in (Internet-accessible) cloud storage, so tools access the data from there.

As we have said, the switchboard can also be invoked from another software application, the CLARIN Virtual Language Observatory⁶. Consider the scenario where a linguist uses the VLO to find an English text which she then would like to investigate further. On the VLO search results page, the user can now click on the ... area to invoke the LRS with this resource, see Fig. 2(a).⁷ In a new browser tab, the LRS opens and shows a resource pane that depicts all relevant information about the resource, see Fig. 2(b). The user is free to correct this metadata⁸, before clicking on 'Show Tools' to get to the task-oriented view, shown in Fig. 2(c). If the user, say is interested in the lemmatization task, she may wish to get more information about the two tools offered, in which case more detailed information about the chosen tool is given, see Fig. 2(d). When the user then clicks on 'Click to start tool', the chosen tool, here WebLicht, opens in a new browser tab. WebLicht obtains from the LRS a reference to the resource, the resource's mimetype and language as well as the chosen task. WebLicht opens with the predefined easy chain for lemmatization, loads itself the resource, and sets all relevant parameters so that the user is left to click on WebLicht's RUN command to start the processing chain. No further user action is required to parameterize WebLicht for this.

Status. At the time of writing (January 2018), a total of 60 browser-based applications and a dozen of web services have been connected to the switchboard. The tools are sorted along the tasks they achieve. Tools include: a chunker for Polish, constituent parsers for English and German, dependency parsers for Polish, German, Dutch, English, Slovenian, Croatian, and Serbian, named entity recognizers for German, English, Polish and Slovenian, shallow parsers for Polish, and tools for word sense disambiguation and sentiment analysis. There are also web services for the analysis of audio data such as runASR for the transcription of speech signals, and runMinni for the segmentation of speech data into phonetic segments. So far there has been an emphasis on tools for the processing of German, English and Polish texts but we strive to integrate tools that offer NLP tasks for other European languages. Note however, that the switchboard has a rather generic nature: given the media type and language of the resource, it suggests applicable tools that can process the resource. Once the language characteristics is set to generic, only the media type become a discriminating factor for tool selection.

2.2 The EUDAT service B2Drop

B2DROP is one of the main data services offered by the EUDAT Collaborative Data Infrastructure. The service is advertised as "a secure and trusted data exchange service for researchers and scientists to keep their research data synchronized and up-to-date and to exchange with other researchers"⁹ (van de Sanden et al., 2015). B2DROP's base functionality competes with commercial services such as Dropbox¹⁰, OneDrive¹¹, Google Drive¹², and many others. Standard functionality includes some free amount of cloud storage, cross-platform synchronization support, file versioning, and the ability to share files with other users. B2DROP's added value stems from its embedding in the EUDAT infrastructure. B2DROP is targeted at European researchers and guarantees that all research data stays on European servers.

Fig. 3 shows the role of B2DROP in the context of the other EUDAT Services. While B2DROP is meant to help researchers managing volatile research data (e.g., draft research papers, experimental setups), it offers a bridge to B2SHARE¹³ to publish such data once it has reached a final state. For this,

⁶It is planned to link the switchboard to CLARIN's Virtual Collection registry (<https://clarin.ids-mannheim.de/vcr> and Federated Content Search (<https://spraakbanken.gu.se/ws/fcs/2.0/aggregator>).

⁷For this, the VLO constructs a URL that points the switchboard, and which encodes (i) a reference to the resource (often a handle), the resource's media type and the resource's language.

⁸The Apache Tika Library usually yields good results, but sometimes the detection of a resource's media type or language is incorrect, for example, when the resource contains too little or obscure data.

⁹See <https://eudat.eu/services/b2drop>.

¹⁰See <https://www.dropbox.com>.

¹¹See <https://onedrive.live.com/about/en-us/>.

¹²See <https://www.google.com/drive/>.

¹³See <https://b2share.eudat.eu>.

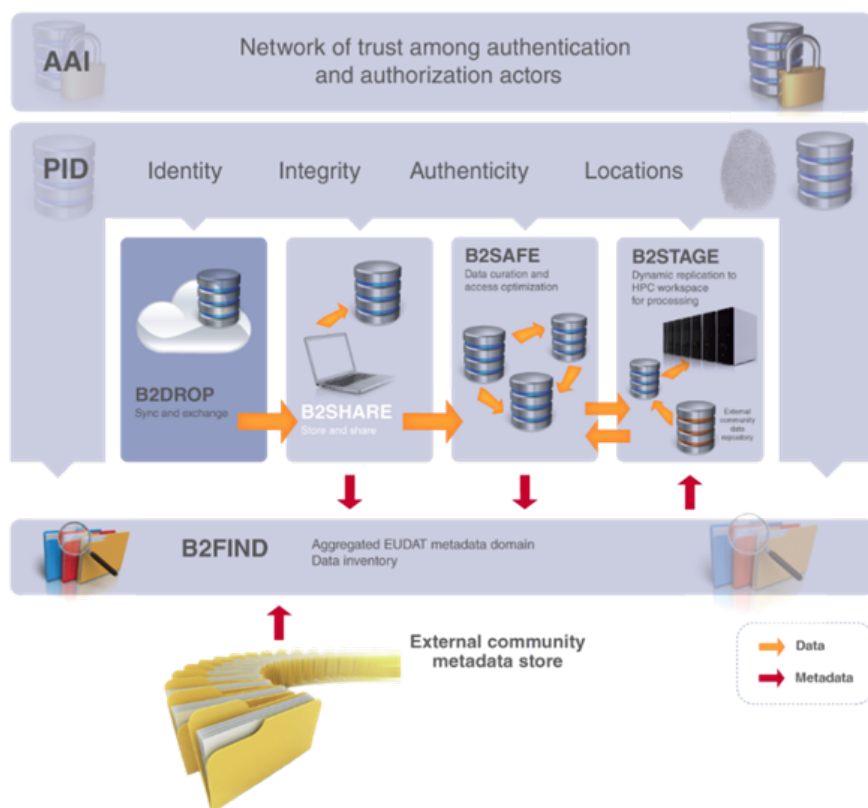


Figure 3: B2DROP in the B2 Service Suite

research data needs to be described with metadata and propagated to B2SHARE, where it also receives a persistent identifier. For public research data, the B2FIND service¹⁴ can be used to search for the data using its metadata descriptions. EUDAT's services constitute a network of trust where users can be authenticated via B2ACCESS; here users can log in with an identity from a research organization they work for, or alternatively with their social identity such as their Google or ORCID¹⁵ account.

Now, reconsider B2DROP in more detail. B2DROP allows individual users to store 20G of research data in the cloud, and to exchange such data with selected colleagues, over a given amount of time.¹⁶ B2DROP is built upon Nextcloud, a fork of ownCloud¹⁷, which is written in the PHP programming language.¹⁸ B2DROP's major contribution to Nextcloud is the provision of a common EUDAT look-and-feel of the cloud's interface. Also, EUDAT developers have provided a Nextcloud plug-in that helps researchers to transfer resources from their personal B2DROP account to B2SHARE, where research data can be stored and preserved for the longer term. The official B2DROP service at <https://b2drop.eudat.eu> is hosted by the Forschungszentrum Juelich. With Nextcloud's software and B2DROP's extension being open-source, it is however possible to easily install, configure and operate a B2DROP server at a local host. For the following use cases, we have set up such a local B2DROP instance using a departmental server.

3 Integration Use Cases

We will discuss two scenarios where the use of B2DROP is beneficial for the LRS and its users. In the stand-alone version of the LRS, we propose replacing the existing file storage server with B2DROP. We

¹⁴See <http://b2find.eudat.eu>.

¹⁵See <https://orcid.org>.

¹⁶See <https://eudat.eu/services/userdoc/b2drop#UserDocumentation-B2DROPUsage-Documentdata>.

¹⁷See <https://nextcloud.com/> and <https://owncloud.org>.

¹⁸See <https://owncloud.org/blog/owncloud-and-php/>.

also suggest complementing the existing usage of the LRS (its use in stand-alone mode or via invocation from the VLO) with a cloud-based usage. While the first integration is of a purely technical nature (it changes the switchboard under the hood), the second integration offers a more visible usability benefit for B2DROP and switchboard users. We have implemented prototypes for both scenarios.

3.1 Using B2DROP as Alternative to the MPCDF server

When users of the stand-alone version of the LRS upload a resource, this is temporarily stored at an external file storage server at MPCDF. This is necessary as all tools connected to the switchboard need web-based access to the resource. The existing server has two drawbacks: the amount of available disk space is limited, and there is little access control in place permitting users aware of the server address to view and access all uploads. To address privacy concerns, it is necessary to better restrict access to file uploads. For this, we have replaced LRS' usage of the MPCDF file storage server with B2DROP:

1. an instance of B2DROP has been installed on a departmental server at the University of Tübingen;
2. a designated B2DROP user 'switchboard' has been registered;
3. when a user uploads a resource to the LRS, the resource is transferred to the B2DROP account of the designated user;
4. using B2DROP's API, the 'switchboard' user creates a shared link for the resource with a link expiration date set to 24 hours;
5. any tool invoked from the switchboard is given access to the shared link to access the resource.

Note that the entire content of the switchboard's B2DROP account is only visible to the 'switchboard' user. A shared link gives only access to the resource that is associated with the link; moreover, the link expires within a short time frame. This is a vast improvement with regard to the MPCDF solution.

A future version of the LRS may allow users with an existing account at <https://b2drop.eudat.eu> to use their own B2DROP cloud storage rather than the generic designated 'switchboard' account. In the meantime, we have also developed an input facility (see middle box in Fig. 1) where users can paste their shared links from their B2DROP or Dropbox account into the LRS.

3.2 Creating a Bridge between B2DROP and the Language Resource Switchboard

We have also created a GUI-based bridge from B2DROP to the Language Resource Switchboard. The inverse direction aims at supporting researchers who manage (part of) their language resources in the cloud, in part, because they need to easily share resources with other researchers (using, for instance, shared links). Here, using the switchboard's drag & drop mechanism would feel rather clumsy: users would need to copy the resource from the cloud to their local desktop, and then open the file explorer to drop the resource into the switchboard (the left-most dotted area in Fig. 1). To improve the usability aspect, we have built a switchboard plugin for B2DROP, which is depicted in Fig. 4. The "Files" view (see circled 1) shows all the files (including directories) stored by the user. Files can be shared with other researchers in which case a "Shared" tag is associated with the resource, see (2), together with a URL pointing to the resource, *e.g.*, <http://weblicht.sfs.uni-tuebingen.de/nextcloud/s/0qeeLnfsj3urgik>. Researchers can give this URL to other researchers so that they get access to the resource as well. Note that each file or directory is associated with a triple dot icon, see (3). When users click on the dots, a menu with actions connected to the resource opens, see (4). This menu has been extended with the action "Switchboard". When users select this option for a resource, the LR switchboard opens in a new browser tab, capable of processing the shared link created by the user.

Implementation and Installation Details

The developers of Nextcloud praise its open architecture; Nextcloud's functionality is extensible via a simple but powerful API for applications and plug-ins ("apps"). One such app is "b2sharebridge", which allows B2DROP users to share their resources via EUDAT's B2SHARE service. We have taken the

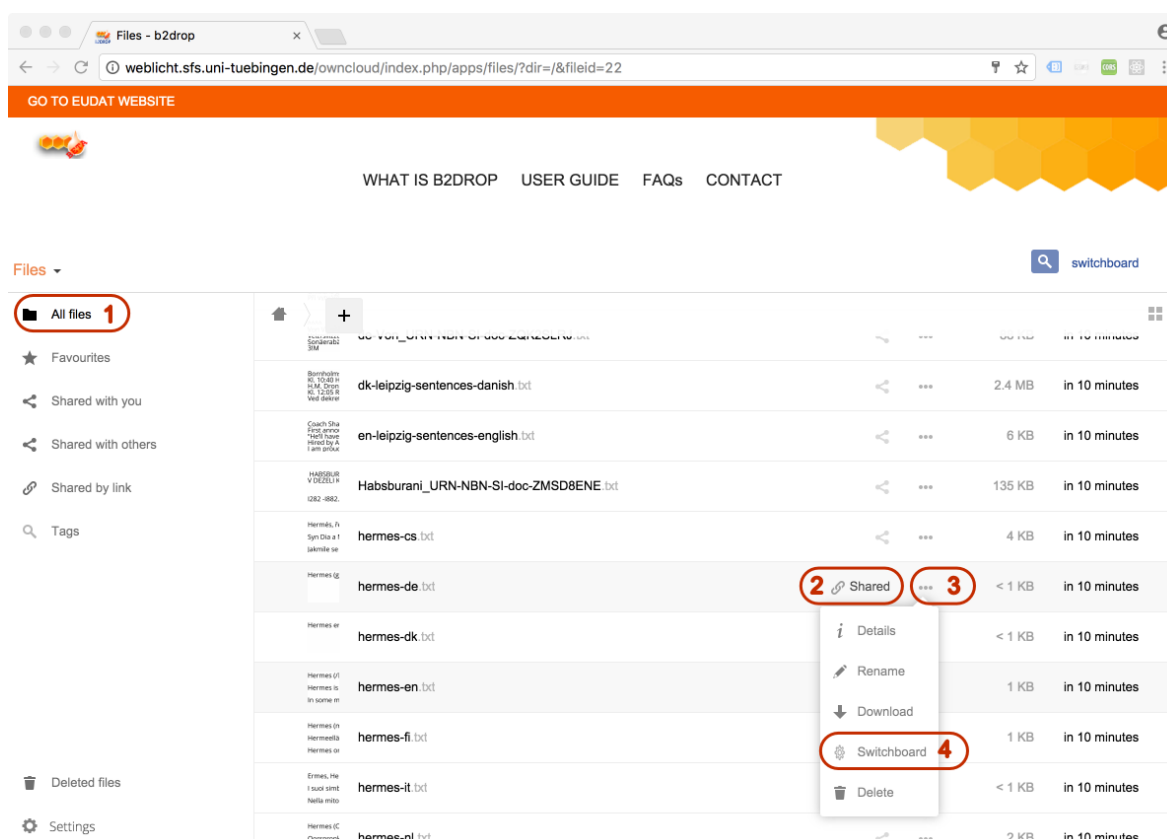


Figure 4: Bridge between B2DROP and the LRS

“b2sharebridge” code as example for the “lrswitchboardBridge” and followed the Nextcloud developer manual.¹⁹ Most of the work required the coding of Javascript code that (i) adds the new item “Switchboard” to the pop-up menu that associates file actions with a given resource (e.g., “Details”, “Rename”); and (ii) implements an action handler for the new action item. The handler creates a new XML HTTP request; here, a URL is constructed that encodes the web location of the LR switchboard, information about its caller, and the shared link to the resource in question. The plug-in then opens the URL²⁰ in a new browser tab. On the LRS side, we have added code that detects from the invocation URL the caller (“b2drop”), downloads the resource from the shared link, and determines the media type and language of the resource. Subsequently, the LRS proposes applicable tools to process the resource. The tool selected for invocation downloads the resource using B2DROP’s shared link.

The installation of the “lrswitchboardBridge” plug-in must be performed by the administrators of the B2DROP/Nextcloud server, following the standard procedure for plug-in installs.

4 Related Work

CLARIN has started to make use of EUDAT’s infrastructure in several ways.²¹ To give CLARIN users easy access to the EUDAT infrastructure, the CLARIN identity provider (IdP) has been integrated with B2ACCESS. Now, users can use their CLARIN account to access EUDAT services. On the repository level, there is an increasing uptake of B2SAFE, EUDAT’s infrastructure for data replication and backup. At the time of writing, the repository systems of five CLARIN centres (CLARIN-AT, Meertens, Tübingen, TLA, and Språkbanken) have been directly integrated with B2SAFE. The integration is ongoing at three other CLARIN centres. So far, more than 100TB of data is managed with B2SAFE.

¹⁹See https://doc.owncloud.org/server/9.0/developer_manual/app/.

²⁰For example, <https://switchboard.clarin.eu/#/b2drop/http://weblicht.sfs.uni-tuebingen.de/nextcloud/s/0qeeLnfsj3urgik/download>.

²¹For a full account, see the CLARIN-PLUS Deliverable D4.2 (Zinn et al., 2017).

B2STAGE is advertised as “a reliable, efficient, light-weight and easy-to-use service to transfer research data sets between EUDAT storage resources and high-performance computing (HPC) workspaces. To the author’s knowledge, the functionality has not yet been taken up by the CLARIN community. There has been considerable progress, however, in integrating the WebLicht workflow engine (Hinrichs et al., 2010) with EUDAT’s experimental Generic Execution Framework service (Dima et al., 2015). The integration allows users to bring the language processing tools integrated into WebLicht to an execution environment that also hosts the data, hence allowing language processing close to the data. The new development enables researchers to use WebLicht for both sensitive and big data (Zinn et al., 2018).

There is ample potential for CLARIN to profit from EUDAT and its generic cloud storage and computing services. An increasing number of users make use of cloud computing²², and many members of the CLARIN community use Dropbox, Microsoft’s OneDrive, Google Drive, or another commercial provider to manage their research data. B2DROP offers a non-commercial alternative: it is based on non-proprietary, open-source software, all data is stored on European-based servers, and EUDAT’s Terms of Use are user-friendlier than those of the commercial providers.

The sharing of services across infrastructures is of mutual benefit for both EUDAT and CLARIN. EUDAT increases its user base as any new “customer” strengthens the role of EUDAT as central infrastructure service provider. Also, CLARIN avoids to duplicate and maintain infrastructure that is available elsewhere. There is, however, a natural tension between generic infrastructure providers such as EUDAT and community-specific infrastructure providers such as CLARIN. For this, reconsider the switchboard plugin, which is currently being deployed and tested on CLARIN-hosted development servers, and which EUDAT has started testing on a EUDAT-hosted development server. Upon successful testing (and pending an agreement with CLARIN to ensure a long-term software support), EUDAT is likely to offer the plugin for all its B2DROP users. In this respect, note that the switchboard’s basic functionality is quite generic: it helps users to connect their resources with tools that can process them. So it might well be that the switchboard’s current tool range expands beyond language-related tools, especially when B2DROP users from other communities expect their community-specific tools to be connected to the switchboard.

From the CLARIN perspective, CLARIN-based B2DROP users would profit from a number of community-specific adaptations. The easiest change would be the inclusion of graphical elements of CLARIN’s corporate identity in B2DROP’s branding (its Nextcloud theme). This would give B2DROP’s GUI a more CLARIN-like look & feel. Also, EUDAT should support the connection of B2DROP with OpenCloudMesh²³, a framework for federated cloud sharing. This would enable users to seamlessly share files with each other, no matter whether they reside on the same (within B2DROP), or on a different cloud server.

A more complex issue is the provision of user delegation services for (trusted) CLARIN applications. Here, we believe that tools connected to the switchboard should be allowed to read from and write to a user’s B2DROP cloud space (with the permission of an authenticated user). In fact, CLARIN-D has tested a prototype implementation based on the UnityIDM authentication service²⁴ and its SAML/OAuth2 bridge with success, see (Blumtritt et al., 2014).

In any case, switchboard users should be aware that data travels through the network. Tools invoked via the switchboard need access to the resources to process them. Including the EUDAT network into CLARIN’s network of trust would certainly help the case. Another step is to make all tools connected to the switchboard part of the trusted network, but this requires considerable future work, especially with regard to user delegation issues.

²²A eurostat report claims that 21 % of EU enterprises used cloud computing in 2016, mostly for hosting their e-mail systems and storing files in electronic form (Giannakouris and Smihily, 2016). In a related statistics, it is reported that in 2014, one in five EU citizens aged 16-74 saved files on internet storage space. Most cloud users appreciated the ease of accessing files from several devices or locations (Seybert and Reinecke, 2014). Quite likely, such numbers will be higher for academic institutions and individual researchers.

²³See <https://oc.owncloud.com/opencloudmesh.html>.

²⁴See <http://www.unity-idm.eu>.

5 Discussion and Conclusion

In this paper, we have sketched two uses of the EUDAT infrastructure service B2DROP for the CLARIN Language Resource Switchboard. The first use of B2DROP improves the back-end of the LRS with the provision of a file storage server that strengthens the privacy aspect of file uploads. File uploads are only accessible for users with access to the shared link, and such links expire after a short time frame.

We consider the second use case more important. So far, the services of the LRS have been at the users' fingertips for personal resources (the stand-alone version of the LRS with file uploads) and for resources advertised in the CLARIN Virtual Language Observatory. With the latest addition, the LRS is now easily accessible for teams of researchers sharing a cloud storage. Resources are uploaded to a Nextcloud-based server, and when a resource is marked as shared, a user can invoke the switchboard with a single click. Once directed at the LRS, users then invoke the tool of their choice also with a single click. We believe that the Nextcloud-based access to the LRS is a feature many users will want to have.

The author is in contact with the administrators of the B2DROP service at <https://b2drop.eudat.eu> to get the switchboard plug-in installed for all B2DROP users. While the technical installation is itself simple, issues regarding the long-term support for the plugin need to be addressed (*e.g.*, who updates the plugin when B2DROP is updated the next version of Nextcloud?). Here, some kind of formal agreement between CLARIN and EUDAT needs to be drawn. B2DROP has been very forthcoming so far, given that the CLARIN community is only one of many communities that take part in the EUDAT project. Having the "IrswitchboardBridge" plug-in enabled by the official B2DROP administrators would bring the CLARIN and EUDAT communities closer together and contribute to service compatibility across digital research infrastructures. With the new bridging service, CLARIN researchers would get the incentive to use B2DROP (and hence, associated EUDAT services such as B2SHARE). As a consequence, commercial services with no such benefits would lose their attractiveness. If the "IrswitchboardBridge" were supported by EUDAT, then European researchers using B2DROP would get easy access to the CLARIN tool space via the LRS. This would significantly increase the usage of many tools across communities, which in turn would challenge tool developers to cope with the new demand, and probably, with new user requirements.

From a wider perspective, the CLARIN community needs to reconsider and potentially adapt its infrastructure pillars. Clearly, many researchers will want to use cloud computing to store research data. Here, EUDAT's B2DROP service helps those researchers to manage and share their data in the cloud. Although B2DROP's adaptation is minimal (it rebrands Nextcloud with the provision of a GUI theme), costs for hosting the service and for user support need to be taken into account. In this paper, we hinted at a tighter integration of EUDAT services within CLARIN. We believe that cloud computing is becoming increasingly important to the CLARIN community. Rather than re-implementing and providing such services within CLARIN, it might be less expensive to use existing generic services of other infrastructures, and have them adapted to community-specific needs whenever possible. The use of components from other infrastructures provides a good opportunity to revisit the overall design rationale of the CLARIN infrastructure and to reconsider the appropriateness of certain components. This assumes, of course, that EUDAT's follow-up project, the EOSC-Hub project²⁵ continues developing, maintaining, and supporting B2DROP and related services.

Acknowledgments We would like to thank the anonymous referees for their comments.

References

- [Blumtritt et al. 2014] Jonathan Blumtritt, Willem Elbers, Twan Goosen, Marie Hinrichs, Wei Qiu, Mischa Sall, and Menzo Windhouwer. 2014. User Delegation in the CLARIN Infrastructure. *Linköping Electronic Press*, (116):14–24.
- [Dima et al. 2015] Emanuel Dima, Christian Pagé, and Reinhard Budich. 2015. D7.5.2: Technology Adaptation and Development Framework (final). Technical report, EUDAT deliverable. Available at

²⁵See <https://www.egi.eu/about/newsletters/introducing-the-eosc-hub-project/>.

https://b2share.eudat.eu/api/files/4cc8cf0e-99a2-4b6b-981a-0ffcd870af19/EUDAT-DEL-WP7-D7%205%202-Technology_adaptation_and_development_framework-2.pdf.

- [Giannakouris and Smihily 2016] Konstantinos Giannakouris and Maria Smihily. 2016. Cloud computing - statistics on the use by enterprises. Technical report, eurostat - Statistics Explained. ISSN 2443-8219, available at http://ec.europa.eu/eurostat/statistics-explained/index.php/Cloud_computing_-_statistics_on_the_use_by_enterprises.
- [Hinrichs et al. 2010] Erhard Hinrichs, Marie Hinrichs, and Thomas Zastrow:. 2010. Weblicht: Web-Based LRT Services for German. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (System Demonstrations)*.
- [Seybert and Reinecke 2014] Heidi Seybert and Petronela Reinecke. 2014. Internet and cloud services - statistics on the use by individuals. Technical report, eurostat - Statistics in focus 16/2014. SSN:2314-9647, available at http://ec.europa.eu/eurostat/statistics-explained/index.php?title=Internet_and_cloud_services_-_statistics_on_the_use_by_individuals.
- [Uytvanck et al. 2012] Dieter Van Uytvanck, Herman Stehouwer, and Lari Lampen. 2012. Semantic metadata mapping in practice: the virtual language observatory. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 1029–1034. European Language Resources Association (ELRA).
- [van de Sanden et al. 2015] Marie van de Sanden, Christine Staiger, Claudio Cacciari, Roberto Mucci, Carl Johan Hakansson, Adil Hasan, Stephane Coutin, Hannes Thiemann, Benedikt von St. Vieth, and Jens Jensen. 2015. D5.3: Final Report on EUDAT Services. Technical report, EUDAT. Available at <http://hdl.handle.net/11304/2433d23a-6079-49a6-9010-ca534f6e348d>.
- [Zinn et al. 2017] Claus Zinn, Twan Goosen, Marie Hinrichs, Emanuel Dima, Willem Elbers, Dieter Van Uytvanck, Dirk Goldhahn, Thorsten Trippel, and Josef Misutka. 2017. Joint infrastructure services. Technical report, CLARIN-PLUS Deliverable D4.2. Available at: https://office.clarin.eu/v/CE-2017-0985-CLARINPLUS-D4_2.pdf.
- [Zinn et al. 2018] Claus Zinn, Wei Qui, Marie Hinrichs, Emanuel Dima, and Alexandr Chernov. 2018. Handling big data and sensitive data using EUDAT's Generic Execution Framework and the WebLicht workflow engine. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).
- [Zinn 2016] Claus Zinn. 2016. The CLARIN language resource switchboard. In *Proceedings of the CLARIN Annual Conference*. CLARIN ERIC. Available at <https://office.clarin.eu/v/CE-2016-0917-Proceedings-CAC-2016.pdf>.

Examining Web User Flows and Behaviours in CLARIN Ecosystem

Go Sugimoto

ACDH-ÖAW

Vienna, Austria

Go.Sugimoto@oeaw.ac.at

Abstract

This article attempts to draw a map of the user flows and behaviours in the multi-layered CLARIN's web structure by cross-examining the dynamic movements of different types of users within (and outside of) the CLARIN domain. In particular, the user traffic of several websites is analysed including the main website, various CLARIN web applications, and the partner websites, as well as the use of single sign-on. Consequently, this project is able to uncover the user interactions in the context of the large web ecosystem rather than those of an individual website. The evolution of the web traffic over a year reveals a comprehensive overview of the characteristics of the end-users and provides a clue for the next strategic decisions over the CLARIN's user-oriented services and business sustainability. This preliminary research also proves the potential of web analytics for Business Intelligence for measuring the impact of the aggregation services and research infrastructures in cultural heritage and digital humanities.

1 Background – the CLARIN ecosystem

One of the strategies of CLARIN is to create and maintain an infrastructure which is financially, technically and organisationally sustainable in the long-term¹. It is, therefore, essential to collect and analyse data about its performance and implement objective evaluation which would determine the course of its sustainability. In particular, as CLARIN's core activities are technically-oriented, offering a number of web-based services to the research community, critical evaluation of end users is necessary to check its performance in the long term and to make sensible decisions for the operation of CLARIN. This area of research is generally called Business Intelligence (BI). According to Chugh and Grandhi (2013), the BI is the process of applying tools and techniques to gather and analyse data from multiple sources, to create knowledge that helps in decision-making.

Several evaluations have been conducted for CLARIN in this respect. For example, Eckart et al. (2015) examined the statistics of the Virtual Language Observatory (VLO)², attempting to explain the user behaviours. Being a part of their technical development of the VLO, this analysis concentrated on the impact of the change of its design and functionality. Two survey periods were defined to examine the consequence of the interface improvement which took place between the survey periods. Subsequently they observed interesting phenomena relating to the user requests especially on full-text and facets searches. Sugimoto (2017) instead provides more comprehensive research on this topic. He conducted a detailed analysis of web traffic on the VLO from 2014 to 2016, taking into account the number of visitors, visit duration, and frequency, to search keywords, social networks, and downloads, as well as segmenting different user groups such as country. It covers most of the default Piwik³ analysis views. Although there are challenges to dealing openly with sensitive information about the performance of a

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

¹ <https://www.clarin.eu/content/mission-and-strategy> (Accessed on 2018-03-19)

² <https://vlo.clarin.eu/> (Accessed on 2018-03-19)

³ <https://piwik.org/> (Accessed on 2018-03-19)

community website, it unlocked a potential (or need) of Open Evaluation for publicly-funded research infrastructure.

However, there are two major aspects lacking in those contributions. First of all, they are limited to a single website. Secondly, they focus on a frequently-discussed technical service of the research infrastructure. With regards to the former, we should remember that CLARIN offers many services and websites for different purposes. Therefore, it is not sufficient to study a single website in order to evaluate the technical infrastructure as a whole. Indeed, a similar approach was taken by Culture24 in the UK (Finnis et al. 2012). They recognised that cultural heritage websites and their visitors can be more adequately assessed and understood by knowing the web use within the entire sector. Thus, major museums and cultural institutions including the British Museum, the National Gallery, Tate, and Kew agreed to share some basic statistics of their websites. The interesting initiative made it possible to standardise the datasets of web access across British heritage institutions and analyse the landscape of their web users. It may have been the first time that an overview of the web traffic within a larger sector was revealed, which massively contributed to the understanding of the bigger picture of emerging museum and heritage business on the Internet. As for the latter, CLARIN's success indicator should not be determined only by technical web applications, but by many other social and organisational services around. In particular, the main website of the infrastructure (CLARIN ERIC: European Research Infrastructure Consortium) should be included.

For those reasons, this paper (re-)evaluates the CLARIN services from a different angle. It takes a holistic approach to capture the traffic of end-users across various websites and applications as well as national centre websites in an attempt to better understand more global aspects of the “customers” of CLARIN. To this end, let us first analyse the CLARIN's web environment.

Although the individual websites of CLARIN are relatively simple, the whole web structure is multi-layered with regard to user movements (Figure 1). The most obvious website is clarin.eu. It is often an entry point for the existing and new users, mainly serving as a communication and dissemination website. It does not only offer the basic information (the missions, people, participating institutions etc.) and updates news and events, but also links to useful websites and services inside and outside the CLARIN. In addition to the main website, there are many web applications developed by the CLARIN developers such as, the VLO, Content Search Aggregator⁴, and WebLicht.⁵ They are useful research tools and are deployed either in the subdomains of CLARIN or its partners domains, therefore, truly making CLARIN a distributed infrastructure. The users jump from the main website, or directly go to, those services to start their research. Although more limited, the users also navigate between the CLARIN services and the partner websites. Many CLARIN national consortia have websites dedicated to providing domestic information, including CLARIN DK⁶ and LT⁷. Moreover, CLARIN centres may have their own websites often placing the CLARIN logo to suggest their connection, for example, the Center for Sprogteknologi⁸ in Denmark and the CLARIN Text Laboratory Centre⁹ in Norway.

As such, there are at least three major entry points to the CLARIN websites (the main website¹⁰, the CLARIN applications, and the national consortia/centres) and the movements of the users among those websites are complex. The author gives the name, “CLARIN ecosystem”, to refer to the full picture of those websites within the CLARIN community. In the sense that we analyse the web access and user flows within the CLARIN community, our approach is different from Culture24, which focuses on completely independent museum websites.

Among the web applications, VLO is probably the most typical case of the CLARIN ecosystem. Therefore, it deserves the name of “VLO ecosystem” on its own. It is a resource discovery portal service to search and locate the linguistic data and tools that the CLARIN consortium members hold, hence it merely collects metadata as an aggregation service provider. Van Uytvanck et al. (2010) describes that

⁴ <https://spraakbanken.gu.se/ws/fcs/2.0/aggregator/> (Accessed on 2018-03-19)

⁵ https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page (Accessed on 2018-03-19)

⁶ <http://info.clarin.dk> (Accessed on 2018-03-19)

⁷ <http://clarin-lt.lt> (Accessed on 2018-03-19)

⁸ <http://cst.ku.dk/> (Accessed on 2018-03-19)

⁹ <http://tekstlab.uio.no/clarino/> (Accessed on 2018-03-19)

¹⁰ It should be noted that there has been no detailed research on the web statistics of the main website, except some general facts and numbers demonstrated, for example, in CLARIN Annual Conferences as well as usability studies.

it tries to give a consistent online overview of the data that is available at a variety of computing centres. Using VLO, the users are directed to the repository of a data provider where the resources they find in the VLO search engine are stored.

Alongside such user streams, the CLARIN's single sign-on services will be examined in order to check the user behaviours by different types of the users including anonymous, the CLARIN registered, and academic users. The objective of this paper is, therefore, to unveil the interactions of various types of users in the large ecosystem which could not be recognised by the previous research based on the observation of a single website.

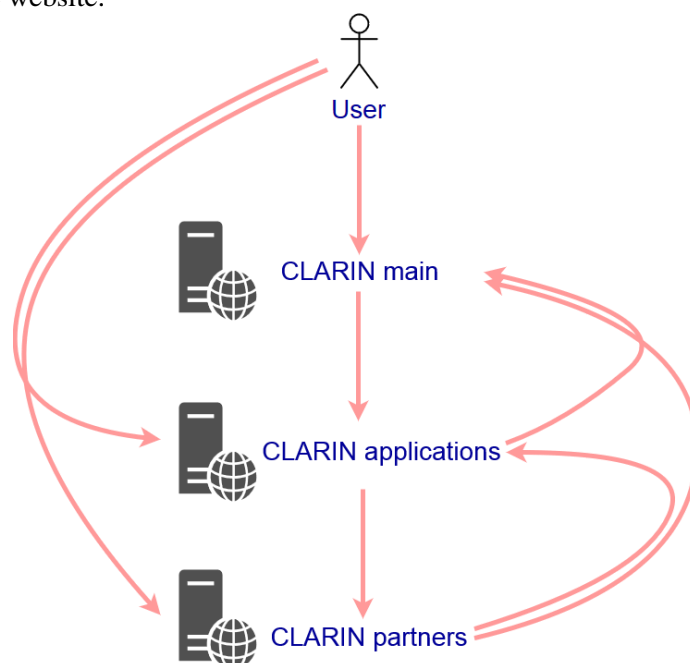


Figure 1. Multi-layered CLARIN web structure (“ecosystem”) and user access

2 Methodologies

The data range of this project is between February 1st 2016 and January 31st 2017, taking into account technical limitations and comparative studies (Figure 2). While Google Analytics is also used to record the traffic of the CLARIN main website, inevitably, Piwik was our choice to analyse the data, as it is the only GUI tool which keeps tracks of all the CLARIN websites that concern us. However, Piwik has been collecting the statistics of different websites since varying points in time. As the main website only started to use Piwik in 2016, we set the beginning of its recording more or less as the beginning of our analysis period. The first half of the period corresponds to the last quarter of the survey by Sugimoto (2017), which might be also useful, if the need of cross-analysis emerges in the future.

In order to reconcile the broad spectrum of the CLARIN's web structure, the author inspects the following websites: the main web-site, VLO, WebLicht, the Content Search Aggregation, the Discovery Service, and the Identity Provider. Although this does not include all of the CLARIN websites, it is assumed that it covers most critical ones, representing what CLARIN offers on the web.

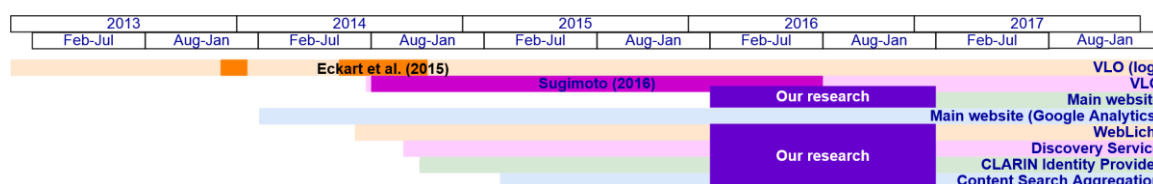


Figure 2. Research period coverage (Available data periods are represented in light colours and re-search periods in dark colours)

The transition view of Piwik is primarily used to analyse the user flow, in combination with other statistical data (Figure 3). It is a powerful tool, able to visualise from where users come to a certain web page and to where they move. As a single website comprises many webpages, there would be potentially

hundreds of views to check the user movements in this way. To avoid interpreting such a large amount of data, it is decided to select some hubs or junction points of user flows. The transition view also allows us to distinguish internal flows (inside the domain) from external flows (outside the domain), thus is very useful to understand the user behaviours. In addition, the classifications of traffic enables us to divide users into separate groups such as the users visiting by search engine or direct entry, as well as the users who downloaded a file or quit the web page. Such footprints of users would provide interesting information for improving CLARIN services.

3 Analysis on the CLARIN (especially VLO) ecosystem

3.1 At the main website -entry gate to CLARIN/VLO ecosystem

First of all, the entry points of the CLARIN ecosystem are examined. Figure 3 illustrates the user flows of the main website at its home page (i.e. clarin.eu). 21,945 page views are recorded in the period, in which 23% are from internal webpages, 18% from search engines, 10% from web referrers, and 36% from direct entries. Within the search engine flow, keywords like “clarin”, “clarin eric”, “clarin eu” and “https://www.clarin.eu” are extremely prominent with 85.6% in total. This implies that most users already knew CLARIN by name, or even the URL, and did not find it by coincidence, for example, when searching for linguistic information. As for the outbound paths, 51% of the users remain on the main website, of which 12% are through to Services, 11% to Events, 8.8% to Participating Consortia, 5.5% to Clarin-in-a-nutshell, and 5% to Users. In addition, 2.8% visited another website, whereas 40% exited (i.e. no more actions by the user). The statistics proved the importance of the VLO as one of the CLARIN’s primary services, as it gained 30% of the Outlinks of the visitors. The CLARIN Germany (3.8% for clarin-d.de) seems to be successful in attracting users from other countries.

It is possible to try to estimate the existing users discussed above more in detail. Firstly, the external access to the website should be the amount subtracting reload and internal pages ($21945 - 5089 - 2039 = 14,817$). The total amount of possibly existing users would be the sum of the search engine access with keywords related to CLARIN and direct access ($135 + 9 + 5 + 3 + 7876 = 8028$). The external access divided by existing users is, therefore, 54.2%. This would be the minimum amount as other channels of access can be observed. This figure can be compared to the more conventional statistics of repeating visits. Piwik recorded 10,000 visits for the same page views as Figure 3, when filtering visits more than once, which is 45.6% of the amount without filtering (21,945). It is not easy to explain the gap. Although Sugimoto (2017) interestingly investigated the black box aspect of Piwik (which would be applicable to any other Web Analytics), both the simple methodology of estimation here and the access handling and recording mechanism of such software are the factor of discrepancy and error. Still, this quick experiment seems to be the only way forward to try to understand the nature of Web Analytics and to adequately and systematically evaluate the web traffic.

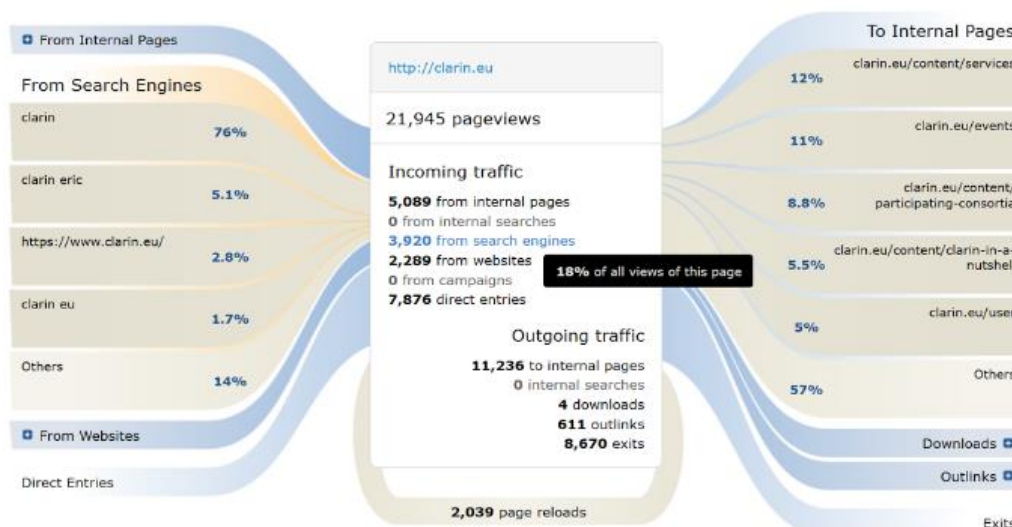


Figure 3. Transition view of the home page of the main website.
(It shows the web page of analytical interest (<http://clarin.eu>) in the centre, where the user was before on the left, and to where the user went afterwards on the right.)

Another point of interest is the Participating Consortia page.¹¹ It lists national consortia of European countries and observers. As described in section one, each consortium may have their own website, so that we can check what consortium receives visitors from this page.

According to Figure 4, outlinks are most represented by CLARIN Germany (12%), Austria (10%), Italy (9.6%), the UK (6.7%), and Latvia (6.1%). In contrast, although the total volume of traffic is 4 times less (i.e. 130) than outgoing traffic (i.e. 522), the incoming traffic from external websites originates from other CLARIN consortium domains. They are CLARIN Slovenia (50 and 25%) and Greece (13%) alongside Wikipedia Germany (13%). When we look at big announcements of national consortia joining CLARIN, there are three relevant countries in the survey period: Latvia (1st of June 2016)¹², Hungary (1st of August 2016)¹³, and France (1st of February 2017)¹⁴. Access to the Latvian website may be explained from this data, whereas the reasons for traffic to other popular consortia are unclear at this stage, as is the absence of Hungary and France. As mentioned above, the participating consortia page is one of the most visited web pages from the main page, so that it would be wise to provide useful and informative content about who the members of CLARIN are besides promoting the national websites. With regard to the internal web pages, nearly half of the visitors (48%) comes from the CLARIN home page, which is naturally expected.

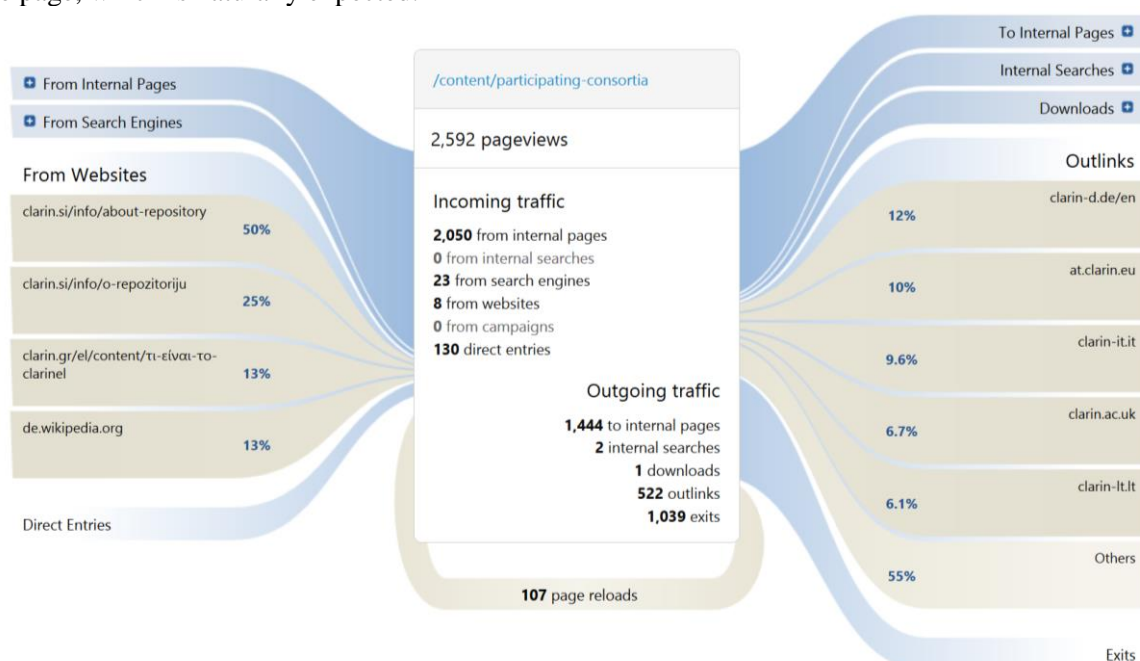


Figure 4. Participating consortia page user flow

The services section of the main website is also a decision-making point¹⁵. This introduction page contains several links to specific services and applications, therefore, allows us to identify the trend of user interests in those services. Figure 5 shows an extremely high percentage of inbound and outbound traffic for the internal website – 92% (3533) and 89% (3402) respectively. Also, given the introductory nature of the content, the data confirm that it is a typical walk-through page of the main website. The decision-making of which links to follow comes into play in our analysis. Within the outgoing flow, the main web page (clarin.eu/portal and clarin.eu) is prominent, but the VLO (9.8%) and the Language Resource Inventory (7%) are also visible among the top 5. Whilst the former is anticipated (see also below), the latter suggests that the users are interested in the LINDAT service on which the Language Resource Inventory is based. Although the incoming flow from external websites is highly limited, there are interesting facts that a few websites have a direct link to the service section page such as the University of Münster and Academic IT Research Support team of the University of Oxford. This is a case

¹¹ <https://www.clarin.eu/content/participating-consortia> (Accessed on 2018-03-19)

¹² <https://www.clarin.eu/news/latvia-joins-clarin-eric> (Accessed on 2018-03-19)

¹³ <https://www.clarin.eu/news/hungary-joins-clarin-eric> (Accessed on 2018-03-19)

¹⁴ <https://www.clarin.eu/news/france-joins-clarin-eric> (Accessed on 2018-03-19)

¹⁵ <https://www.clarin.eu/content/services> (Accessed on 2018-03-19)

of a small fraction of user flow, but Piwik has proven useful to analysing what referrals exist and how the users enter the CLARIN ecosystem.

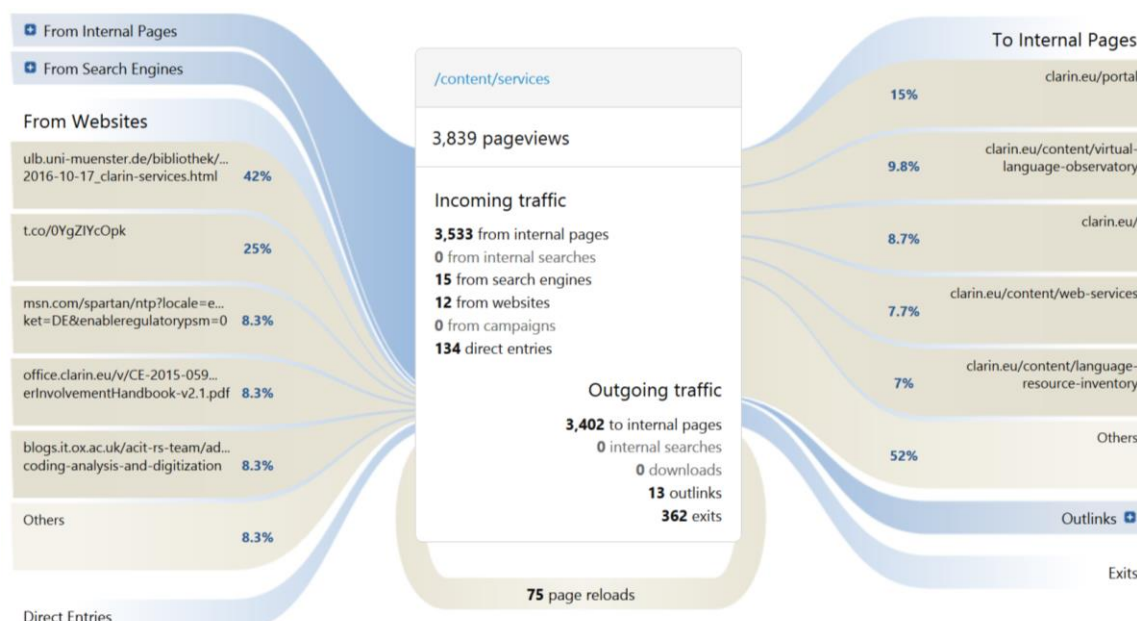


Figure 5. User flow at the service section of the main website

There is a VLO introduction page on the main website which would be one of the main gates to the VLO (Figure 6). 77% of all the visits went to the VLO, so that most users pass this connection point to arrive at the VLO. 44% of the users find the web page from the service section of the website, while the other routes are rather limited (internal search 0.1%, website 4.7%, direct entry 12%). The relatively high number of entries via search engines (21%) suggests that the users know the VLO, because their search keywords include specific terms referring to the VLO or CLARIN. The user flows from the VLO to the CLARIN centres are much more complex and the examination is in progress. Although understandable, it is a pity that we have no access to the statistics of the CLARIN centres. If the access permission is somehow granted, it is possible to examine the complex VLO ecosystem in a similar way that Culture24 was able to do. What we suggest is to share a subset of the whole data in the form of spreadsheets export, instead of the unlimited access to Piwik and/or Google Analytics. Collection of such data dumps from various centres will shed a light on the understanding of the navigation of the CLARIN users.

A part of the problem is that the individual URIs of the centres need to be checked and the use of Persistent Identifiers (i.e. Handle¹⁶) makes it untraceable without manual clicking and checking of all the URIs recorded. Nevertheless, apart from Handle, the University of Leipzig (2.8% of all Outlinks) and the SIL International (2.1%) received more visitors than others.

¹⁶ <https://www.handle.net> (Accessed on 2018-03-19)

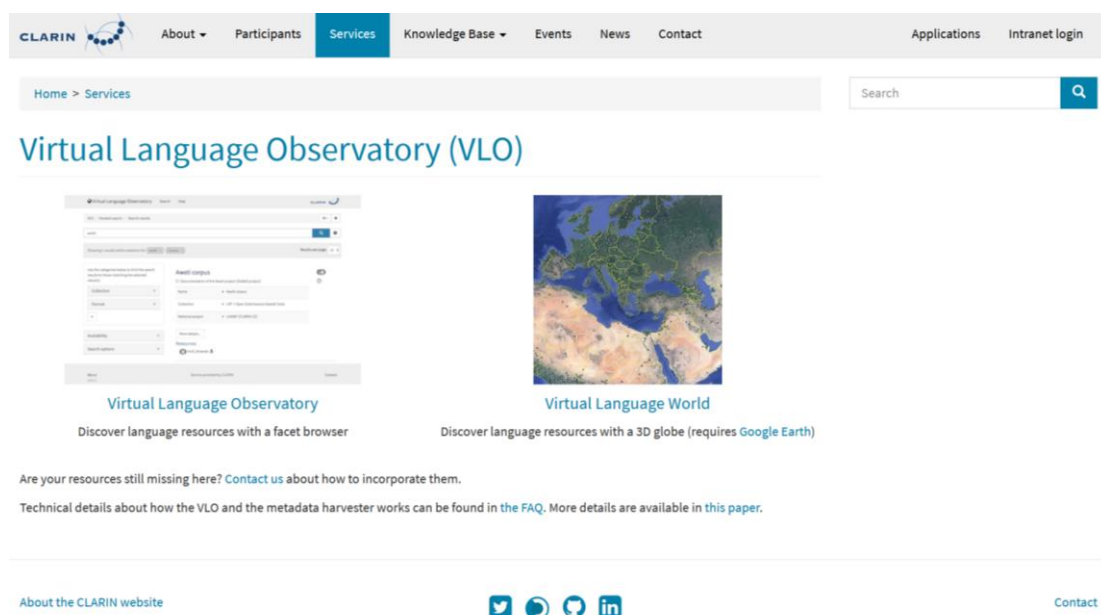


Figure 6. VLO introduction page at the main home page

Before moving on to the VLO itself, let us double-check in what position the VLO is. Figure 7 outlines the top 10 highest number of outgoing access from the entire clarin.eu domain (i.e. not only from the clarin.eu home page, but including it). The VLO tops the ranking as 11.2%. Other VLO related technical services such as centres.clarin.eu (Centre Registry 4.9%) and catalog.clarin.eu (Changing address including Component Registry etc. 4.5%) follows Handle persistent identifiers (6.9%). It is also notable that LINDAT and WebLicht (see section) are among the pages visited frequently by the users.

	URL	Unique Clicks	Percentage
1	vlo.clarin.eu	998	11.2%
2	hdl.handle.net	615	6.9%
3	centres.clarin.eu	439	4.9%
4	catalog.clarin.eu	404	4.5%
5	www.clarin.eu	386	4.3%
6	infra.clarin.eu	255	2.9%
7	lindat.mff.cuni.cz	227	2.6%
8	www.clarin-d.de	215	2.4%
9	docs.google.com	181	2.0%
10	weblicht.sfs.uni-tuebingen.de	118	1.3%

Figure 7. Top 10 outlinks from the whole clarin.eu domain

3.2 At the VLO

From the VLO's point of view, the trend of in- and out- channels is different (Figure 8). 22% of the visits originate from web referrers. 600 out of 1102 visits from websites (54%) are the VLO introduction page (with additional 5.8%). Interestingly, the Stackexchange website has a post about a Korean language corpus and the VLO is mentioned. As a result it gained a high rate of access (7.4%) during this period. Similarly, 5.0+ % are observed due to the University of Vienna offering a Moodle link to the VLO. Unlike the main website, a low number of users landed with the VLO via search engines (4.8%). 29% of the users find the website directly. Regarding the outward traffic, we can see a clear trend for *Korean* probably caused by the abovementioned stream ("korean" (1.7%) and "korean corpus" (2.6%)). At the first glance 31% of the users who went through to internal pages may have done so by browsing, because the VLO is a search engine which, in principle, should increase internal searches (24%). However, this assumption cannot answer why internal searches are less than page browsing. When it is discovered that the internal pages contains the URL syntax pattern such as "vlo.larin.eu/search?1", the

classification by Piwik becomes slightly dubious. It is nevertheless important to note that unlike on the CLARIN main website, much higher numbers were recorded for internal searches both in and out directions of the traffic. Yet another puzzle piece is the difference between 637 (inbound internal searches) and 1218 (outbound internal searches) as well as the existence of both identical search keywords and different ones. In general, more iteration of observations, analyses, and experiments would be needed to solve this kind of mystery, for example, by understanding the details of the mechanism of auto-generated URLs in the VLO, as well as what the Web Analytics records and classifies. In the meanwhile, 21% exited without doing anything.

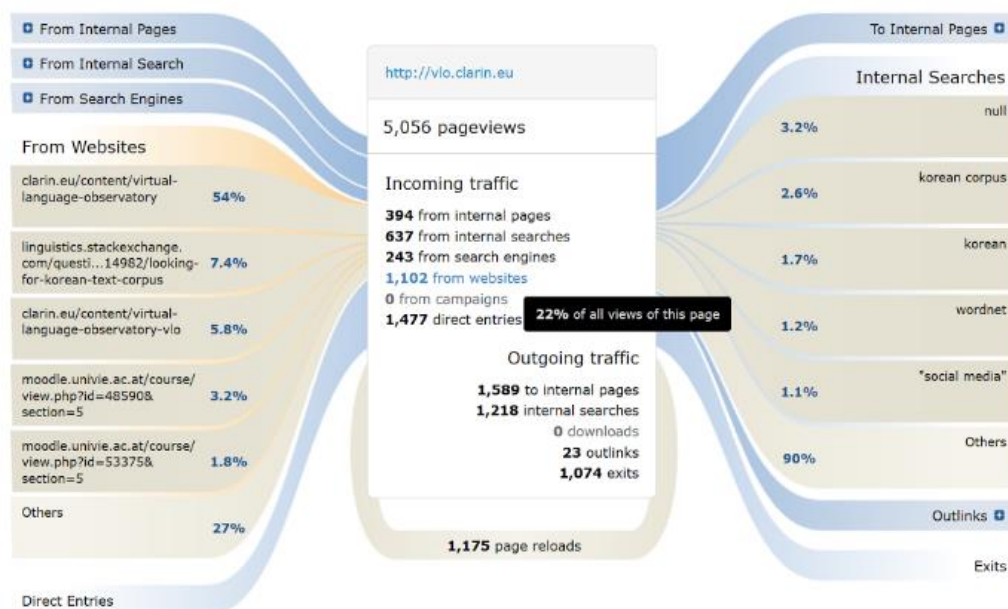


Figure 8. User flow at the home page of VLO

On the other hand, overall search keywords left some clues on the user needs (Figure 9). Interesting cases of Korean have already been introduced. In addition, some users search something very specific such as “hzsk” (0.9%, probably intended for the CLARIN B centre of Das Hamburger Zentrum für Sprachkorpora (HZSK)¹⁷), “GECO” or “geco” (0.9%, also intended for IMS GECO Datenbank provided by the CLARIN B centre of Universität Stuttgart¹⁸), and “germanet” (0.3%, also intended for the service by University of Tübingen¹⁹). It is obvious that they look for German data and tools. It seems that such access is made by the CLARIN’s internal users, rather than the experts outside CLARIN who know exactly what CLARIN offers. The tendency towards language names cannot be ignored and this trend was also found during the two years of Sugimoto’s analysis (2017).

¹⁷ <https://corpora.uni-hamburg.de/hzsk/> (Accessed on 2018-03-19)

¹⁸ <http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/IMS-GECO.html> (Accessed on 2018-03-19)

¹⁹ <http://www.sfs.uni-tuebingen.de/GermaNet/> (Accessed on 2018-03-19)

KEYWORD	SEARCHES	SEARCH RESULTS PAGES	% SEARCH EXITS
null	 2.1% 86	1.2	12%
hzsk	0.9% 39	3.2	67%
korean corpus	0.9% 36	3.6	61%
corpus	0.6% 23	3.3	48%
russian	0.6% 23	2.3	83%
treebank	0.6% 23	2.5	57%
korean	0.5% 20	2.6	25%
geco	0.5% 19	1.9	53%
german	0.5% 19	3.3	47%
GECO	0.4% 17	1.9	88%
wordnet	0.4% 16	3.1	31%
chn	0.4% 15	1.4	100%
dutch	0.3% 14	4.2	50%
germanet	0.3% 14	2.1	50%

Figure 9. Search keywords used in the VLO

It is also very easy to learn what the users downloaded (Figure 10). However, as the number is significantly lower than the visits in total, it was decided to display only the highest ranking URIs in this paper. It is perhaps fair to mention that the trouble of this type of analysis is that 120 links have to be manually clicked and checked to know exactly what the downloaded contents are about. Although some URIs could give some hints of content in the syntax, opaque URIs, especially persistent identifiers like Handle, make it impossible to guess the content of the target resource. Given that the number of outlinks is much bigger, there are limits for the manual analyses. This is one of the very interesting and unfortunate pitfalls of persistent identifiers in terms of Web Analytics. This paper does not mean to say that opaque URIs should be avoided. Rather, it only suggests that both the creators and implementers of persistent identifiers may need to consider this aspect for improvement or solution in the future, if Web Analytics deploying substantial amount of manual work is considered to be important.














DOWNLOAD URL	UNIQUE DOWNLOADS	DOWNLOADS
vlo.clarin.eu/ - Others	57	59
 corpora.uni-leipzig.de/downloads/ukr_newscrawl_2011_1M-text.tar.gz	2	3
 vlo.clarin.eu/record72-1.ILinkListener-cmdi-toggler-link&docId=CLARIN Centres/oai_clarin_pl_eu_11321_270.xml&q=KPWr&index=8&count=13	2	2
 vlo.clarin.eu/record75-1.ILinkListener-tabs-tabs-container-tabs-1-link&docId=CLARIN Centres/oai_clarin_pl_eu_11321_270.xml&q=KPWr&index=8&...	2	2
 cocoon.huma-num.fr/exist/crdo/schang/gcf/crdo-GCF_1016.xml	1	2
 clarin.phonetik.uni-muenchen.de/BASRepository/Corpora/CH-Jugendsprache/SNF_jspr_i4_S2_001_061019_Beziehungsnetz/SNF_jspr_i4_S2_001_0610...	1	1
 clarin.phonetik.uni-muenchen.de/BASRepository/Corpora/SC10/CLARINDocu.zip	1	1
 clarin.vdu.lt/xmlui/bitstream/handle/99999/10/ALKSNIS_v2.zip?sequence=1	1	1
 corpora.uni-hamburg.de/repository/file:kolas_kolas-1.0-documentation/PDF/andresen-knorr-kolas-dokumentation.pdf	1	1
 corpora.uni-hamburg.de/repository/file:kolas_kolas-1.0/ZIP/kolas-1.0.zip	1	1
 cts.informatik.uni-leipzig.de/teidumps/pbc/bible/parallel/deu/elberfelder1905.xml	1	1
 hdl.handle.net/11041/alipe-000853/ali-baptiste-101227-2.xml	1	1
 hdl.handle.net/11041/sidr000758/olac.xml	1	1
 slidr.org/logo/LogoOrtolang_small.png	1	1

Figure 10. Top download URIs within the VLO

Those additional (potential) analyses clarify that multi-dimensional analyses, combining user behaviour analysis, in this case, for keyword searching and downloading, with transition analysis, can make a significant contribution to the understanding of the users as the principal creatures of the ecosystem environment.

4 WebLicht and Content Search Aggregator

“WebLicht is an execution environment for automatic annotation of text corpora. Linguistic tools such as tokenizers, part of speech taggers, and parsers are encapsulated as web services, which can be combined by the user into custom processing chains.”²⁰ Consequently, the structure of the website/web application is very different from the main website and the VLO, resulting in no transition view produced by Piwik. In fact, the user flow exists in terms of the data processing chain, but not in terms of web pages. Therefore, we need to look at other statistics. 70% of visits to WebLicht are referrers, while 29% are direct entry. As the CLARIN-D is the developer of the WebLicht, the referrers are mostly from the German domains, except for the top score of “idp.clarin.eu” (29%). Similarly, Germany dominates the visits by country (82%), while there is also interest from Austria (3%), South Korea (3%) and the United States (1%) (Figure 11). WebLicht is perhaps something CLARIN has failed to promote. Although it can handle many languages (for example, there are more than 40 language choices for plain text processing), the service is almost exclusive to Germany, the major CLARIN consortium member. It seems that CLARIN would need to review the outreach strategies of WebLicht in order to go beyond the German niche market. The visit duration is substantially longer (11 minutes 57 seconds on average) than for the VLO (4 minutes 18 seconds) (Sugimoto 2016). 27% spend more than 10 minutes, proving the characteristics of the data processing service. This engagement promises that new users potentially become heavy users.

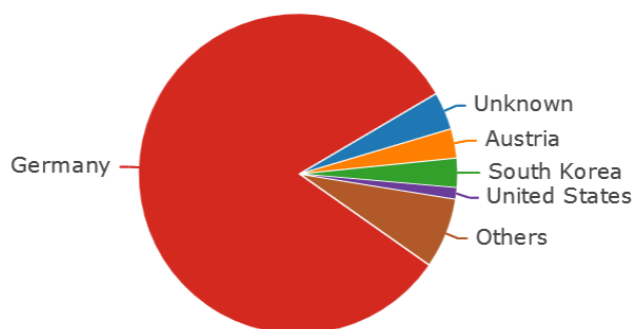


Figure 11. WebLicht visits by country

As for the Content Search Aggregator (Figure 12), the transition view is valid. A high ratio of page reload was detected (39%) in comparison with the main website (9.3%) and the VLO (23%), whereas web referrers come second at 34%. A very low amount or no users arrived internally (i.e. via web pages (0%) and search (2.3%)). In fact, less than 10% accessed from the CLARIN main website. On the other hand, CLARIN-D successfully converted their users to the Content Search users (over 75% of referrers). The implications of those results need to be further investigated. Incoming internal searches indicates the presence of German speaking users, as the most searched keywords are all German including “armut” (poverty, 17.4%), “Forsythie” (Forsythia, a type of shrub, 4.3%), “selbstmord” (suicide, 4.3%), and “diachrone deutsche korpora” (diachronic German corpora, 4.3%). Regarding outgoing internal searchers, there are more varieties, but German words are still the most visible. “Leipzig” (1.6%), “selbstmord” (1.6%), “vom text zur phonologischen aussprache” (from text to phonological pronunciation, 0.8%), “mal eben” (just in a moment, 0.8%) are shown in the highest. The same marketing argument we made for WebLicht applies to the Content Search Aggregator.

²⁰ https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page (Accessed on 2018-03-19)

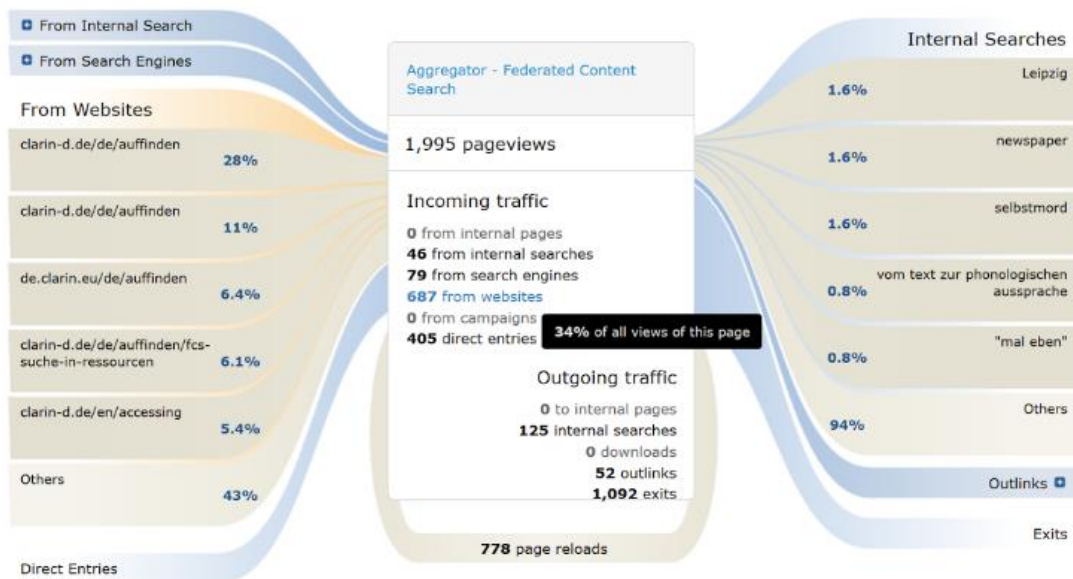


Figure 12. User flow at Content Search Aggregator

Compared to the VLO ecosystem, the situation of other applications is different. It is worth analysing the characteristics of the visits, but the user flow inferred from Piwik is rather limited. We can conclude that although it is necessary to monitor the flow from the main website, those services are rather the end points of the CLARIN ecosystem, thus, it is more productive to analyse the VLO ecosystem in this sense.

5 Identity Providers and Discovery Service

CLARIN provides a pragmatic solution for user authentication and authorisation. The recording of user sign-on and access to web services enables us to explore the statistics of different user types in CLARIN's web space to support our previous analyses. We analysed Identity Provider (i.e. only users with CLARIN credentials) and Discovery Service (i.e. all users trying to access log-in services)

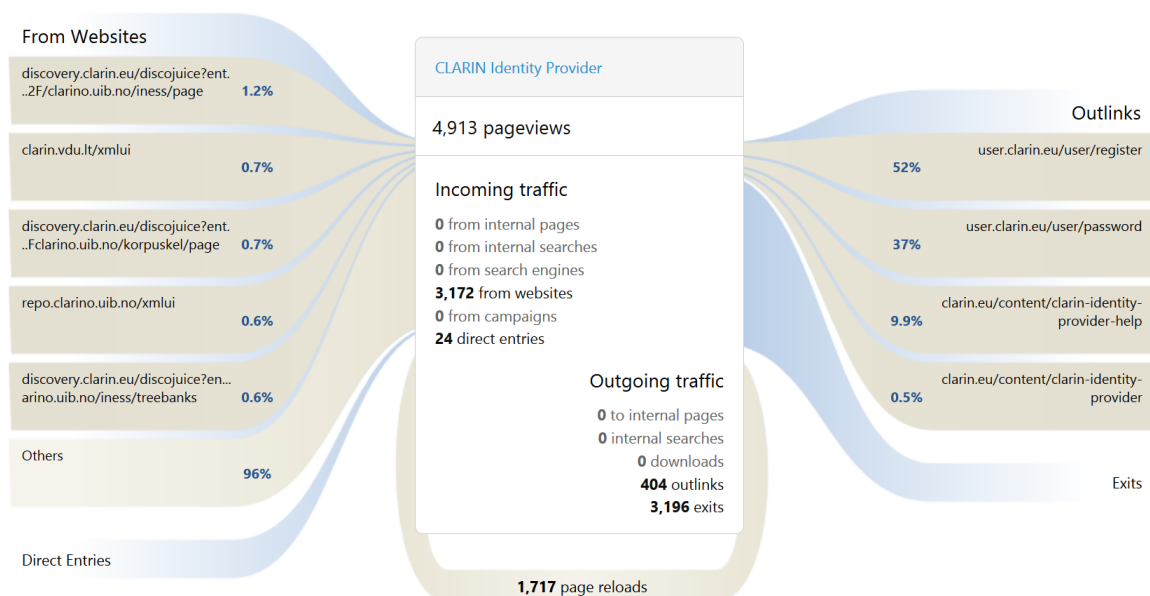


Figure 13. User flow for Identity Provider

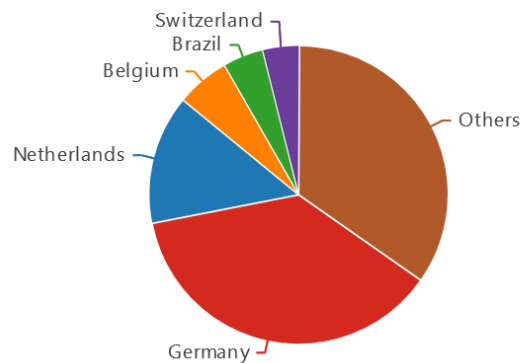


Figure 14. Access by country at Identity Provider

Although the URIs are relatively spread (Figure 13), as for the incoming web referrers of Identity Provider, there are only two countries which gain visibility: Lithuania and Norway. The Lithuanian repository of CLARIN-LT²¹ (0.7%) and the Norwegian repository by CLARINO²² (0.6%) delivered more users to the Identity Provider. The Norwegian boom is further boosted by other URIs of (probably) INESS.²³ In contrast, interestingly those Northern countries do not appear in the access by country (Figure 14). It may mean that German and Dutch users, as well as Belgian, Brazilian, and Swiss users to a lesser extent, are interested to find Lithuanian and Norwegian resources. The assumption was mostly right that among the total amount of 99 views of Norwegian domains within the top 5, the users from Germany accessed Norwegian domains 38 times and users from Norway did 43 times. The large majority comes from those two countries. It is a common phenomenon that users use resources from their own country, therefore, German is something unique in this context, in a way fulfilling the aim of CLARIN to encourage trans-European access. However, it is again true that the German population bias as well as the influence within CLARIN are big. Outbound traffic seems to be rather technical and there is not much from which we can draw conclusions. In case of Discovery Service (Figure 15), it is the Netherlands which dominates the scene for the incoming traffic. Among the URIs, Corpus Hedendaags Netherlands (9.4%+6%)²⁴, Open Sonar (4.5% and 4.2%)²⁵, although WebLicht shows strength (9.1%). This is clearly represented in the pie graph depicting access by country (Figure 16). The swap of German and Dutch users is quite dramatic and interesting, but we need more evidence to explain this situation. Again, outlinks contain URIs too technical to mention.

In the meanwhile, both the Discovery Service and Identity Provider have a large proportion of exit (65% and 82% of outbound traffic respectively). This may imply that many users give up access due to this access restriction. In that case the Service Providers may want to reconsider their access policies. While the former acquired 41% from referrers, the latter is at 94%, which is probably naturally high as a sign-on screen appears when a link on a webpage is clicked. It is, however, noted that the technical mechanisms of those services are complicated, making the recording (and interpretation) of the user access in Piwik very tricky. In order to clarify the situation, the next step of investigation would be to carry out an experiment to understand what Piwik actually records behind the user interactions with those CLARIN services, using the Visitor Log function.

²¹ <https://clarin.vdu.lt/xmlui/> (Accessed on 2018-03-19)

²² <https://repo.clarino.uib.no/xmlui/> (Accessed on 2018-03-19)

²³ <http://clarino.uib.no/iness/page> (Accessed on 2018-03-19)

²⁴ <http://corpushedendaagsnederlands.inl.nl/> and <http://chn.inl.nl/> (Accessed on 2018-03-19)

²⁵ <http://opensonar.inl.nl> (Accessed on 2018-03-19)

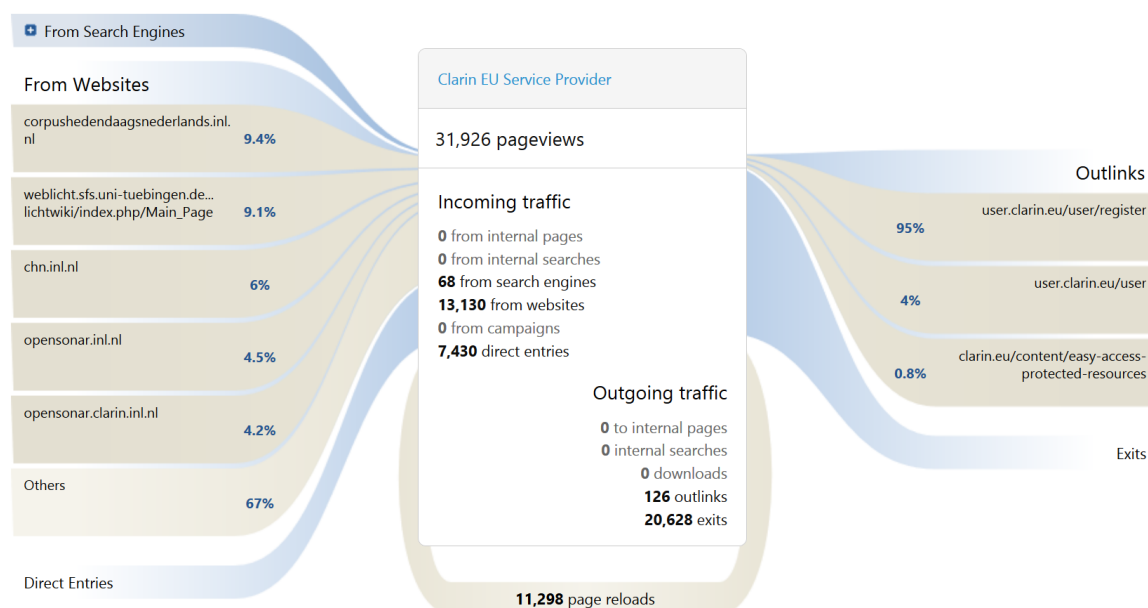


Figure 15. User flow at Discovery Service

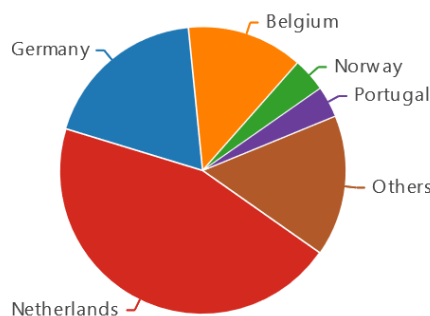


Figure 16. Access by country at Discovery Service

6 Conclusion

The transition view of Piwik in combination with other functions allows us to effectively evaluate the user traffic streams in the multi-layered web structure. It is easy to browse the types of inbound and outbound movements of the users. In particular, an unprecedented amount of statistics about the CLARIN and VLO ecosystem was analysed in detail. Whilst the role of the VLO at the centre of CLARIN's web infrastructure was confirmed, the complex user flow within the ecosystem uncovered some trends of this survey period. Some junction points like participating consortia and the VLO introduction page provide insights into the users moving out of the main website. In general, the existing users, characterised by direct entry and "intended access by search engine" (keywords are used to specify CLARIN and VLO), seem to influence the statistics to a large extent. In addition, the fact that many specific search terms are recorded in the VLO also adds to the evidence of CLARIN's internal community users. It may be still too early to draw a conclusion that CLARIN websites rely on internal community users. However, initial results collected so far are affirmative, even when they are compared to the outcomes of Sugimoto (2017) who suggested a heavy usage of the VLO by a CLARIN partner in Austria²⁶. In a way, CLARIN fails to catch attention from external community. CLARIN should definitely consider the transition of heavy user base from CLARIN members to the outsiders. For example, the CLARIN annual conference could be open to a larger community, thus, the infrastructure can be more widely recognised and used. In addition, CLARIN could reduce the internal networking and research mobility between CLARIN centres, and increase workshops and seminars in fringe domains such as

²⁶ Austria is often regarded as one of the core technical members of CLARIN.

philology and language-related subjects. In particular, the researchers who do not normally deploy computational linguistic methods would need crash courses to obtain practical skills and knowledge to use CLARIN resources and tools.

The high volume of flow from Germany can be seen in different traffic records, but the population bias is not yet taken into consideration. Nevertheless, as one of the core members of the CLARIN consortia, Germany hugely influences the web traffic. On one hand, CLARIN benefits from the driving force, on the other hand, the European infrastructure seems to need more effort to expand the user base outside Germany. The value proposition of CLARIN clearly states (CLARIN ERIC 2017) that “as generic infrastructure services can be used across borders, CLARIN members can benefit from the fact that the costs of construction and operation of such services can be shared between members” and “access to CLARIN resources (data, tools and methods) will also lead to more advanced research and open new research avenues across borders and disciplines”. For this reason, the reduction of CLARIN activities in Germany and the expansion of CLARIN programmes in less popular countries may be a good option for widening the user diversity. More knowledge transfer from active CLARIN countries to less active countries would also be a new strategy agenda for cross-border synergies. The impact of a sudden increase in particular access paths such as “Korea” became easily visible from the beginning to the end of the access paths, supporting the detailed analysis possibility of Piwik.

Although WebLicht and Content Search Aggregator provided less useful information about the user flow, and are thus not extremely suitable for the analysis of the user movements within the CLARIN ecosystem, they underpin the large amount of German users. Identity Providers and Discovery Services are also particular in the sense that they are the layers to go through to CLARIN services. The analyses revealed that Norway and Lithuania gain popularity, mainly due to the access from Germany. The dramatic swap of the Netherlands and Germany poses a question to be answered.

There are also some areas where further research is needed to clarify the situation and provide correct interpretation. For example, it is a challenge to scrutinise the websites after a major overhaul (for example, Goosen and Eckart (2014) and CLARIN ERIC (2016)). Web addresses may change over time due to the introduction of new underlying software and/or restructuring of the website. Such a change introduces a complicated list of page URLs for transition analysis. It would be wise for the web analytics team and development team to closely communicate about the web development plans, so that the troubles of web traffic evaluation could be minimised and CLARIN’s tasks can be more efficiently coordinated, for instance, by extending the members of the VLO Task Force (Haaf et al. 2014). Besides, the marketing strategies created by web analysis and the development of websites could go hand-in-hand for the efficient and continuous improvement of the infrastructure. The tight cooperation would potentially save money and address the needs of the right users and other stakeholders. Therefore, it is important both in terms of the technical, organisational, business, and financial stability of CLARIN. More cooperate governance²⁷ needs to be implemented. In addition, the technical mechanism behind authorisation and authentication in relation to the recording of Piwik is still unclear. Moreover, a pitfall of opaque persistent identifiers was recognised. It is not a big problem for the scale of analysis in this paper, but as the web access grows, it would make detailed and interesting analyses more difficult.

The preliminary results of this paper successfully displayed new in-sights into the end-users of CLARIN. In addition, this is probably the first time to synthesise the statistical analyses of both the dissemination website and the web applications of CLARIN in terms of user traffic. Moreover, it is also a reconfirmation that it is important to monitor the statistics over time. A comprehensive implementation of Business Intelligence would require more data from different areas such as financial reports and user engagement reports. Nevertheless, it is hoped that this small research project has brought some ideas about the visitors and environments of the CLARIN’s virtual ecosystem in the framework of web analytics and would be a valuable contribution to the development and sustainability of CLARIN.

References

[Chugh and Grandhi 2013] R. Chugh, and S. Grandhi. 2013. Why Business Intelligence? Significance of Business Intelligence Tools and Integrating BI Governance with Corporate Governance. In *International Journal of E-Entrepreneurship and Innovation*. 4 p1–14. <http://doi.org/10.4018/ijeei.2013040101> (Accessed on 2018-03-19)

²⁷ https://en.wikipedia.org/wiki/Corporate_governance (Accessed on 2018-03-19)

- [CLARIN ERIC 2016] CLARIN ERIC. 2016. CLARIN Newsflash September 2016 | CLARIN ERIC. <https://www.clarin.eu/CLARIN-Newsflash-September-2016> (Accessed on 2018-03-19)
- [CLARIN ERIC 2017] CLARIN ERIC. 2017. Value Proposition. <https://www.clarin.eu/value-proposition>. (Accessed on 2018-03-19)
- [Eckart, Hellwig, and Goosen 2015] T. Eckart, A. Hellwig, and T. Goosen. 2015. *Influence of Interface Design on User Behaviour in the VLO*. In *CLARIN Annual Conference 2015 Book of Abstracts*. <https://www.clarin.eu/sites/default/files/book%20of%20abstracts%202015.pdf> (Accessed on 2018-03-19)
- [Finnis, Chan, and Clements 2012] J. Finnis, S. Chan, and R. Clements. 2012. *Let's Get Real -How to Evaluate Online Success?*. <https://www.keepandshare.com/doc/3148918/culture24-howtoevaluateonlinesuccess-2-pdf-september-19-2011-11-15-am-2-5-meg?da=y> (Accessed on 2018-03-19)
- [Goosen and Eckart 2014] T. Goosen, and T. Eckart. 2014. Virtual Language Observatory 3.0: What's New? In *CLARIN Annual Conference 2014 in Soesterberg, The Netherlands*. http://www.clarin.eu/sites/default/files/cac2014_submission_2_0.pdf (Accessed on 2018-03-19)
- [Haaf, Fankhauser, Trippel, Eckart, Eckart, Hedeland, Herold, Knappen, Schiel, Stegmann, and van Uytvanck 2014] S. Haaf, P. Fankhauser, T. Trippel, K. Eckart, T. Eckart, H. Hedeland, A. Herold, J. Knappen, F. Schiel, and D. van Uytvanck. 2014. CLARIN's Virtual Language Observatory (VLO) under scrutiny-The VLO task-force of the CLARIN-D centres. In *Clarín 2014 Conference [CAC2014]*. http://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/3210/file/Haaf_Fankhauser_CLARINs_virtual_language_observatory_under_scrutiny_2014.pdf (Accessed on 2018-03-19)
- [Sugimoto 2017] G. Sugimoto. 2017. Number game. In *ArXiv:1706.05089 [Cs]*. <http://arxiv.org/abs/1706.05089> (Accessed on 2018-03-19)
- [Van Uytvanck, Zinn, Broeder, Wittenburg, and Gardelleni 2010] D. Van Uytvanck, C. Zinn, D. Broeder, P. Wittenburg, and M. Gardelleni. 2010. Virtual language observatory: The portal to the language resources and technology universe. In N. Calzolari, B. Maegaard, J. Mariani, J. Odjik, K. Choukri, S. Piperidis, M. Rosner, D. Tapias (Eds.), *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC\textquotesingle10)* (pp. 900–903) Valletta, Malta: European Language Resources Association (ELRA).

Digital Classics and CLARIN-IT: What Italian Scholars of Ancient Greek Expect from Digital Resources and Technology

Monica Monachini

ILC-CNR

Pisa, Italy

`monica.monachini@ilc.cnr.it`

Anika Nicolosi

Dip. DUSIC

Parma University, Italy

`anika.nicolosi@unipr.it`

Alberto Stefanini

Novareckon

Novara, Italy

`alberto.stefanini@virgilio.it`

Abstract

This paper presents and discusses the findings of a survey carried out to assess the use of digital resources and digital technologies with respect to work in ancient Greek scholarship, with the aim to identify the factors that are likely to constrain its use as well as to elicit needs and requirements of ancient Greek scholars in Italy. The survey is in line with the principles behind the user engagement strategy developed by CLARIN-ERIC and constitutes one of the national efforts undertaken by CLARIN-IT to contribute to the wider impact of CLARIN on Digital Classicists. The survey, as well as other surveys carried out in the sector in the last decade, points out that most of the available resources do not respond to users' requirements. This motivated us to develop a mock-up of a digital editor of Archilochus, which, mostly grounded on previous studies by Nicolosi, draws on the outcomes of the survey. The experiment includes a sample prototype to submit for evaluation by end-users. The final aim is to identify good practices and new models to enable new approaches to the study of classical texts and profile a new workbench for scholarly digital edition.

1 Introduction

Interest for the humanities and social sciences in language technologies has never been as strong as it is now. The main conferences in the Digital Humanities are seeing an increase in participation by computational linguists while at Computational Linguistics' conferences the humanities and social sciences represent an important line of research. The necessity of meeting the needs of an audience of different users opens up new challenges for language technologies: easily usable tools, adaptable to different types of content become crucial. The quality of resources, in particular the quality of digital editions of texts, is receiving increasing attention. For this reason, it is crucial to identify user requirements in relation to textual (and linguistic) analysis tools, in view of contributing to the advancement of this specific field of science. Attention to a new or different approach to a traditional discipline determines, not unsurprisingly, the development of new learning habits and, on the basis of the good practices inherited from the previous tradition, allows the development of a different and more modern research methodology and of new practices in didactics.

Scholars and scientists require modern, well-established research infrastructures to conduct internationally competitive research. Researchers in science and engineering have witnessed the potential offered by infrastructures and already use them for their work. The humanities already have a tradition of data and knowledge aggregators, with archives and libraries that contain texts and can be regarded as the research infrastructures of the past. When it comes to the adoption and expansion of digital research infrastructures, however, the humanities still lag behind other scientific fields. Research in that field, of course, also depends on the

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

opportunities offered by modern digital technologies, with the foundation for research being based on the amount of data digitally available around the world, ready to be processed by computers. Scholars of the sectors, however, were rather slow to adopt them.

CLARIN-ERIC (www.clarin.eu), the European research infrastructure, was created to make digital language resources available to scholars, researchers, students from all disciplines, especially in the humanities and social sciences. This involves providing digital repositories where data, corpora, lexicons, tools are catalogued, stored and retrieved in a simple way as well as developing technological solutions that can be intuitively used by users. CLARIN represents the perfect framework for bringing together producers of language technology with its users. CLARIN considers it essential to find out from users what resources and tools they (would like to) use, what solutions they prefer and what training support they need.

The idea that user needs should be a central part of the design and development process of any ICT infrastructure is not necessarily new, and pointing out that the CLARIN audience are humanists and social scientists is not enough. There is a wide range of people working within the academic sector: some may be more proficient in one area but less so in another, so there are different needs and different ways of engaging with users. Accordingly, CLARIN defines different methodologies and approaches to user engagement. Surveys are deemed to be suitable for making deeper analyses of landscapes, opportunities, barriers, etc.; they are useful for identifying previously unnoticed problems or areas of dissatisfaction and are especially helpful if we are not sure where to start or what to prioritize, or if we need to show evidence to support our decisions. Surveys can be crucial not only to elicit information about users' requirements and barriers to uptake, but also to reach and inform people about what we are doing. They might find that CLARIN already has solutions for them that they did not know about.

2 Motivations for a CLARIN-IT Survey of Digital Classics

In order to study the current interest in the use of digital resources and related tools in one specific context of the Humanities, Ancient Greek scholarship, we launched a survey based on a questionnaire to ascertain the current practice and the related needs within a group of practitioners in the field. On larger scale, the work represents one of the first attempts undertaken within the context of CLARIN-IT (Monachini and Frontini, 2016) to contribute to the wider impact of CLARIN on the specific Italian community of those interested in the application of Digital Humanities to the field of Classics and to ancient world studies. During the last decade, similar surveys were sporadically carried out. They pertain to the fields that do not strictly concern Digital Classics (DC) although the general outcome of their remarks is relevant, and it is evident that they may also be applied to DC resource design.

These studies concern a wide spectrum of scientific interests within the digital humanities realm and involve scholars from several countries, mostly English native speakers (USA, UK, CDN etc.). Our study, instead, focuses on the specific scientific community dealing with Digital Classics; it collects the views of a restricted sample of Italian digital humanists with focus on ancient Greek philology.

2.1 Previous Surveys

(Babeu, 2011) provides a summary of several surveys on the subject. For our purposes it is enough to quote the key outcomes of the studies by (Toms and O'Brien, 2008), (Audenaert and Furuta, 2010) and (Warwick et al., 2008).

Toms and O'Brien's study is of a behavioral nature and was, similarly to our own, conducted on a small sample of digital humanists who responded to a questionnaire published on the web. The main conclusion of the study is that "the digital humanist (...) gives value to primary and secondary materials (books) and uses more browsing than searching on the internet". His/her preferred research strategy is based on linking rather than on searching

and concatenating works, usually by referring to the material of interest through the bibliography quoted in the single article. Toms and O'Brien's study draws the figure of a lonely scholar, with a few joint publications, more interested to communicating than collaborating with colleagues. Digital humanists' priorities are to have access to primary sources and to integrate lessons with material from web searches; under this profile, they are interested in text presentation, i.e. having "multiple views" of the material or text being analyzed. Their interest particularly goes to tools for granular analysis of texts, at various levels, and in the most sophisticated text analysis and annotation tools with a variety of mark-up languages.

The study by Audenart and Furuta mainly concerns the relationship of digital humanists with the primary source, analogically or digitally reproduced, and results in a list of recommendations for the design of digital libraries¹. The authors argue that the existing resources essentially aim to disseminate material, while in general, there is a lack, in their opinion, of environments to support text analysis and understanding². The authors reiterate an argument already raised by Toms and O'Brien, namely that the available environments are aimed at finding information rather than using it. In a series of semi-structured interviews with a panel of eight researchers, they looked for answers to three key questions: *1. Why are scholars interested in examining the original textual material? 2. What kind of information do they search for? 3. How and when they use ICT and for what purpose?* The answers are that scholars aim to make use of the original material because either it is not readily available or no reliable transcripts are readily available. In many cases, even when transcripts available are deemed appropriate, the specialist considers essential to access the original, by direct visual inspection. Usually, scholars want to have access to any documents that deliver a certain interpretation and to information about everyone who contributed to them (author, public, publisher, illustrators, and scribes). The process of transmission of the text is usually the primary interest of the scholar. This led the authors to identify a model called SCAD consisting of four components: primary sources or Sources (including original drafts or copies); Context (cultural, socio-political and economic), Authors / actors, and Derived forms, or works that re-use the text in question as a source. The authors conclude with major final recommendations for the CSE: its ultimate goal is the usability of the resource extended over time; it, therefore must provide extensive documentation and a clear understanding of the users' needs. This involves a programmatic consultation with users, constant maintenance and update of the interface, content and functionality of the resource.

The study by Audenaert and Furuta may to some extent be compared with that conducted in the LAIRAH (Log Analysis of Internet Resources in Digital Humanities) project by Warwick, Galina, Terras, Huntington, and Pappa aimed, on the one hand, to identify twenty broadly used resources with Log Analysis techniques and, on the other hand, to interview their creators in order to understand the reason behind their popularity. In short, the authors, draw general still valid conclusions: who uses digital technology tends to prefer general to specialized resource; in particular, humanists have sophisticated mental models and high specialized skills in their field, but find it difficult to apply these skills in a digital environment. They need a wide spectrum of resources in order to discover new ways of thinking about what is already known, since discovering new data or facts is quite a complicated process; they use digital resources, only if they match their mental models and their research methods, thus refusing unfriendly interfaces or confused data and abhorring resources that require specialized training. They are worried

¹This study is part of a larger project to design a creativity support environment (CSE) for in-depth analysis and study of paper-based materials.

²The available environments, they claim, have developed resources that address the individual research needs of their developers or they are modeled on theoretical definitions of what the research practice should be in the digital environment. Many times this results in the recommendation of practices that have been intimidating to many scholars, such as the claim that they manually encode documents into XML

about the accuracy of the data, i.e. want "high quality content", that means having detailed information about the sources of digital resources. While many of the above considerations may seem obvious in that they correspond to the general usability criteria of a web resource, it is interesting to see how they are reiterated by a detailed analysis. Some points emphasize the importance of having high quality digitization and accurate information on data and processes adopted. It may also be noted that the results of this study are consistent with the findings of Audenaert and Furuta 2010. While none of the resources selected by the study concern Digital Classics, the overall relevance of these findings is clear for DCs as well. The authors emphasize the importance for users to play an active role in determining the design criteria of a digital project (designers as users) and insist that resource planners should never infer user requirements from their own behavior. Many of the projects have shown that their users were much more varied than they thought. Another criterion that determines the success of a digital resource is its sustainability. Data that are not deposited in institutional archives, which guarantee their conservation, easy access and documentation, soon fall into disuse.

Only the survey by (Toms and O'Brien, 2008) encompasses a broader spectrum of literary interests, with Latin prevailing – 17%: most of the interviewees are working on modern and contemporary literature, with only 13% interested in the classical and post-classical period.

3 The CLARIN-IT Consultation

For the reasons above, our study was carried out on a restricted sample of Italian digital humanists, interested in ancient Greek philology. It is a relatively small field (the number of scholars in Italy is about 130 people) but this area of study is traditionally of great interest in Italy and also includes university students and schoolteachers. Moreover, Italian Ancient Greek scholars are an active part of an international community (spreading especially in Europe, North and South America) and this field of studies has great potential: it is worth remembering that Ancient Greek studies are an essential part of our Western Cultural Heritage, and it is crucial to spread the knowledge of these studies. For these reasons, the Italian Ancient Greek community is a small but excellent sector where to test the new opportunities offered by Digital Humanities. The scope now is national and therefore narrow, but, thanks to CLARIN, it would be interesting to extend the analysis across borders and to other fields. This research sector, indeed, clearly characterized by international cooperation, requires an international and well-coordinated effort.

The survey (supplementary to a master degree thesis discussed at the University of Parma (Stefanini, Nicolosi and Monachini, 2017)) was carried out from May to September 2016 and is now available on-line through the CLARIN-IT channels³. In a questionnaire-based survey, the sample should be statistically representative of the target population. The questionnaire was sent to selected Italian researchers whose main focus of study is ancient Greek language, although their interests span over a broader area, encompassing Greek and Latin literature. The sample shows different professional roles: full professor, associate, researcher, and other (mainly Italian researchers working abroad and schoolteachers). The survey aims to evaluate the impact of digital techniques within the specific reference community of ancient Greek scholars in Italy. At this research stage, the sample is numerically consistent with the survey target because of its specific expertise (Ancient Greek): it is about 10 percent with respect to the initial potential target population (see Figure 1).

3.1 Questionnaire key points

The survey focuses on the digital resources and tools needed to support an excellent and usable digital edition of an ancient text. For this reason, first, we ask applicants to specify their field of expertise and evaluate the tools they use and know.

³at <http://www.clarin-it.it/it/content/sondaggio-current-practice-digital-classics-tools>.

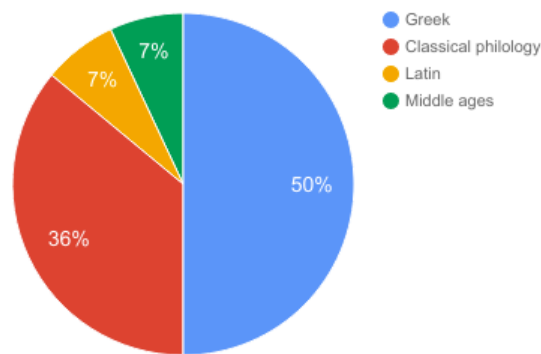


Figure 1: Interests of the respondents

The questionnaire has four sections (see Figure 2 below):

CURRENT PRACTICE with Digital Classics tools

Indicate which tool you use in your current practice: *

Your answer

evaluate the extension of the textual data base of the tool

Your answer

which functionalities of these tools do you consider more useful?

Your answer

which functionalities are lost or are not easy to use?

ADVANCED FUNCTIONALITIES for Digital Classics

In the current practice you may find tools for:

☐ finding on the web information about bibliographic references

☐ translating texts in a digital form, comparing different versions of the same text (OCR)

☒ linguistic analysis (semantic, thematic, morphosyntactic analysis etc.)

UTILITY of digital resources and tools for philological studies

Digital resources and tools may interact in order to:

☐ Match different versions of a text, with provision of witnesses and digital critical apparatus

☐ Make available different readings with computer-based linguistic/stylistic analyses

☐ Make available, whenever possible, a digital copy of the primary source (code, papyrus, epigraph etc.)

☐ Provide translation into one or more contemporary language

Say to which extent you deem useful an experimentation of these practices:

Your answer

Please indicate your field of expertise:

FUNCTIONALITY EVALUATION

From 1 (= not at all) to 5 (= very much), which functionalities do you deem more relevant your study practice? *

	1	2	3	4	5
syntactic analysis	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
metric text analysis	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
text variants	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
set of conjectures or readings	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
interactive support to reviewing text variants	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
several critical editions of the same text	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
make available digital copy of the primary source (code, papyrus, epigraph etc.)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
translation into contemporary languages	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
hypotheses and make them available to others	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 2: The questionnaire

1. **Current practice:** which Digital Classics tools researchers use in their studies;
2. **Advanced functions:** which available/unavailable function is/would be more useful;
3. **Usefulness** of digital resources and tools for philology, through ranking of four key functionalities:
 - match the different versions of a text – mentioning the sources – and provide critical apparatuses
 - make available diverse interpretations supported by computer-based linguistic/stylistic analysis

- make available a digital copy of the primary source (code, papyrus, epigraph etc.)
- provide one or more translations of a text in contemporary languages

An evaluation of the usefulness of one or more of these functions is required.

4. **Quantitative evaluation:** a 1-5 scale allows the respondents to assess the usefulness of a set of functions considered crucial for digital editions dealing with classical studies. Table 1 below summarizes the results.

Function	Average score
primary sources	4.00
variants	3.57
critical editions	3.57
collaborative hypotheses	3.36
logical and syntactic analysis	3.36
conjectures	3.29
metric analysis	3.14
translation	3.00
reviewing variants	3.00

Table 1: Quantitative Evaluation

3.2 Overview of replies

All the answers were timely and accurate. All the respondents showed interest in the topic and consensus emerged with respect to the need to develop and improve this research field.

The key outcomes of the survey may be summarized as follows.

1. Current Practice with Digital Libraries. All the respondents are familiar with the most important tools in the field (mainly Perseus DL and TLG), but also other available resources are mentioned (i.e. Trismegistos, Perseids and Alpheios, Musisque Deoque and PHI). They are generally considered good tools, but all these resources receive some criticism concerning their coverage and/or their usability and/or their availability.

2. Advanced Functions. They received particular attention by all respondents who insisted on functions such as, syntactic analysis, text search, digital copies of the primary source, bibliography, and translation in contemporary languages. Search by lemma, morphologic analysis paired with the availability of on-line dictionaries, are deemed the most useful functions, although they are not always available.

3. Usefulness of Digital Resources and Tools. There is a common requirement to improve the *syntactic analysis*, improve and refine *search mechanisms*, such as syntax-based searches, and make Application Programming Interfaces (API) available for several functions. Some users ask for hypertext links, with syntactic and grammatical analysis (e.g. tree) of the texts. In general, it is considered crucial to have annotated, searchable and interoperable, interlinked data.

There is agreement among respondents about the importance of text *variants* and of apparatuses as complete and comprehensible as possible. One of the replies says rather optimistically that those are current practice already, but admits that they are not available at the same time or linked to each other.

Some replies highlight the need to provide *primary sources*. As well as previous surveys have highlighted, it is important but it is not enough alone.

Translations in contemporary languages obtained a rather low score. This might be due to the bearing of the sample towards research rather than teaching. Collaborative hypotheses

functions did not receive a high score: this could be due to the need for a scientific board that guarantees the output or due to the habits of classics scholar who often works alone. It is worth noting that the result is to some extent contradicted by qualitative replies (see Table 1, above).

All the respondents highlighted the importance of carrying out experimentations in the field and insisted on the need to develop and/or make tools more reliable and usable. They found several inadequacies or unsatisfied desiderata: in particular, they complain about the absence of linguistic analysis and the lack of a complete and reliable critical apparatus. Some of them ask for lemmatization and annotation of texts and would expect far greater interoperability of data. They highlight the need to increase the material available in some fields of study, i.e. ancient Greek poetry, and ask for better usability of tools; they would welcome more attractive tools, equipped with user-friendly interfaces; they also point out that there are no tools able to integrate textual data and bibliography links, or hypertext links with other texts or resources available. (Table 2 shows the main deficiencies of available DH tools).

R1	<i>Relational syntax analysis functions are still missing. There are no tools that integrate textual data and bibliography links.</i>
R2	<i>Alignment with translation and syntactic and semantic analysis.</i>
R3	<i>Some tools need the addition of texts and the improvement of existing ones; others require the addition of several editions for the same Greek source.</i>
R4	<i>Some tools lack APIs (possibly Restful).</i>
R5	<i>Metric analysis may be useful, for lyrical sections, in particular.</i>
R6	<i>Even advanced tools offer poor - or nil - statistical disambiguation of morphological analysis and lemmatization. Other instruments of undoubted value are practically unusable due to the strong license restrictions. The tool's usefulness is very much tied to the quality of the reference editions.</i>
R7	<i>Failure to digitize the critical apparatus makes the tool unreliable. Research is conducted on the basis of the edition taken as a reference for each author, without the possibility to consult the variants.</i>
R8	<i>The possibility to combine word search and search of syntactic constructions.</i>
R9	<i>Lack of critical apparatus.</i>
R10	<i>Difficulty in understanding the reference source; when information about the source is found, it is often not so clearly identified.</i>
R11	<i>Advanced search, searching for co-occurrence of terms, creating concordances, selecting texts (single texts or groups of texts).</i>
R12	—
R13	<i>The visualization of the results is unsatisfactory.</i>
R14	<i>Hypertext links.</i>

Table 2: Main inadequacies

Finally, many respondents pointed out that models and software for authoring, editing, indexing and presenting a digital edition are important research directions. Digital editions may provide scholars with copious, very complete materials to ease their research and their studies, with a deeper insight into useful research methods.

As part of the overall strategy, the survey outcome is currently made public on CLARIN-IT (Nicolosi, Monachini and Stefanini, 2017) at <http://hdl.handle.net/20.500.11752/OPEN-86>, together with its questionnaire, so as to open our consultation to anyone willing to contribute (Figure 3).

ILC4CLARIN Repository Home / View Item

Search

Survey Data: Current practice of digital resources and tools for studies on Digital Classics

Please use the following text to cite this item or export to a predefined format: [BIBTEX](#) [CMDI](#)

Nicolosi, Anika; Monachini, Monica and Stefanini, Alberto, 2017, *Survey Data: Current practice of digital resources and tools for studies on Digital Classics*, Digital Repository for the CLARIN Research Infrastructure provided by ILC-CNR, <http://hdl.handle.net/20.500.11752/OPEN-86>.

Share: [f](#) [t](#) [g+](#)

[OPEN](#)

Authors	Nicolosi, Anika ; Monachini, Monica ; Stefanini, Alberto
Demo URL	https://docs.google.com/forms/d/e/1FAIpQLSepPWyzL0fvmD8mx6Qdy-S3OVjr_jpSCdKvO_fGtVmGvNOWw/viewform?c=0&w=1
Date issued	2017-10-31
Type	corpus
Language(s)	Ancient Greek (to 1453) , English Italian ,
Description	<p>This dataset contains the original responses to a questionnaire run from May to September 2016 on a sample of Italian digital humanists with focus of interest on ancient Greek philology about Current practice of digital resources and tools for studies on Digital Classics: namely Ancient Greek. The survey is now available on-line at http://www.clarin-it.it/content/sondaggio-current-practice-digital-classics-tools.</p> <p>The majority of questions were closed questions where respondents had to tick a box, occasionally multiple choice was allowed. A few questions required free text provision. The questionnaire was designed using 'Google Forms' and was run on the same platform in the autumn 2016.</p> <p>The results of the survey are briefly presented to the CLARIN Annual Conference 2017 (18-21 September 2017, Budapest). Abstracts available here: https://www.clarin.eu/sites/default/files/Monachini-Nicolosi-Stefanini-CLARIN2017_paper_3.pdf.</p>
Publisher	Università di Parma
Subject(s)	survey data digital classics users' needs
Collection(s)	ILC4CLARIN : OPEN Data & Tools

What can you do?

[DEPOSIT](#) [CITE](#)

Browse

> All of the Repository

My Account

Login

Statistics

Statistics **BETA**

General Information

Deposit

Cite

Submission Lifecycle

FAQ

About

Help Desk

Figure 3: Access to the Survey in the CLARIN-IT Repository

4 Action Plan: Implementing the Outcomes of the Survey

The outcomes of the survey motivated us to develop a mock-up of a scholarly digital editor for the ancient Greek poetry and to test its suitability and usefulness. These are the main steps of our action plan.

Existing on-the-shelf solutions, despite being often excellent products, provide an interface that reproduces the printed page of a commentary book⁴. We believe that a far more flexible and user-friendly solution may be envisaged⁵. For these reasons, we developed a mock-up of a scholarly digital editor of Archilochus, which draws on texts, translations and commentaries edited by (Nicolosi, 2013), while being mostly based on the survey. The mock-up provides, through an extensive use of windows and hyperlinks, a set of digital resources and other facilities, which allow the users to inspect the ancient text at many levels, thus easing its critical assessment. The experiment concerns a few fragments of the Greek poet to provide a prototype for evaluation by its intended end-users, in view of developing a full scholarly digital edition.

⁴For example 'The Classical Text Editor was designed to enable scholars to work on a critical edition or on a text with commentary or translation to prepare a camera-ready copy or an electronic publication without bothering much about making up and page proofs' (<http://cte.oeaw.ac.at/>). It also provides features oriented to a digital publication, i.e. the ability to export the publication in XML format according to the TEI standard, including formatting and styles adopted, graphic objects included in the publication, and references to external graphic representations

⁵For digital editions see (Pierazzo, 2014), (Sahle, 2008) and (Ruecker, 2008)

The final aim is to set a good practice, identify new models and new typologies of approach to the study of classical texts and profile a new workbench for scholarly digital edition. The intended audience for this tools is twofold, university students and scholars.

4.1 The Genette Model

The first recommendations made in the survey have been implemented on the basis of the model presented in (Genette, 1982) which allows a text to relate to other texts in several ways: *intertextual*, for quotation, plagiarism, allusion, *metatextual*, i.e. through a critic, reflexive relationship, *architextual*, as belonging to a literary genre, *paratextual*, i.e. with its textual periphery or *hypertextual*, for parody, spoof, sequel, and translation. These relationships may also include references to existing databases – where ancient texts’ witnesses are digitized – to imagery (objects, landscapes and tools referred by the poet) and to geographical databases, where places referred by the text are described. They may also refer to linguistic resources such as dictionaries, syntactic analysers and automatic translators.

4.2 The Mock-up

It is necessary to make a distinction between mock-ups and prototypes. A mock-up, still offering a detailed design of the final tool, is an intermediate step that only preludes to full software implementation and it represents obviously a rather low-cost operation. When adapting an existing text to a digital format, the key concern is the best user profit, i.e. how to exploit the ample repertoire of solutions and resources that digitalization may offer. The presented mock-up still doesn’t follow standards and design patterns; now it is envisaged only for a test and it is a standalone application. It will be revised, according to the evaluation, and it will become a network-based system. The mock-up (Nicolosi, Monachini and Stefanini, 2017) - which is mostly based on text, translations and commentary edited by (Nicolosi, 2013) - is now available at <http://hdl.handle.net/20.500.11752/OPEN-83> (Figure 4).

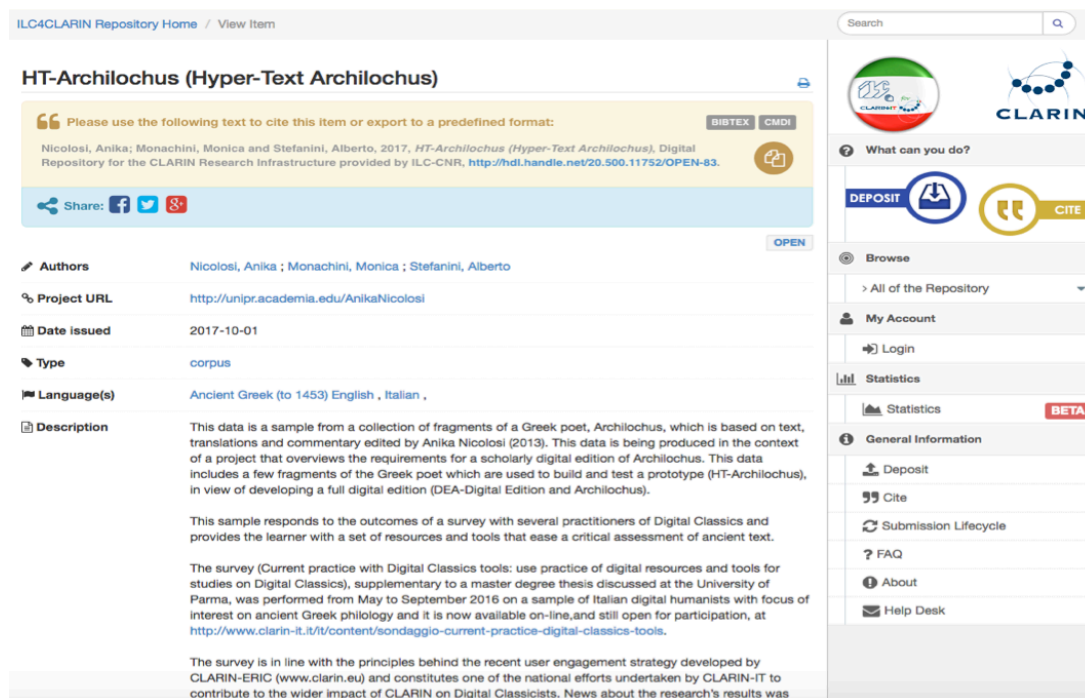


Figure 4: Access to the mock-up in the CLARIN-IT repository

It may be considered as an augmented scholarly and born digital editor compared to a simple digitization of a paper text, as it has a set of features that a simple digital edition does not present such as:

- a basic hypertext structure
- multimedia integration of text, images and graphics
- integration of material from multiple sources
- textual search tools
- reading support tools such as online translation, vocabulary, syntactic tree consultation

The mock-up view (Figure 5) shows a screen-shot of the set of the general functions, encompassing Italian translation, grammatical and metric analysis, and bibliography, the syntactic tree of the fragment⁶, the sources, loci similes, and some linguistic peculiarities and geographic information through invocation of the geographic database *Pleiades*. It is also possible to link the fragment to other relevant resources.

Figure 5: General Functionalities

Each button corresponds to a different function. The buttons ‘*introduzione generale*’ and ‘*bibliografia*’ switch to full pages where a general introduction to the fragment and its bibliography are given. It is also worth noticing that the Greek text is hyperlinked in order to access to Greek word study tools and on-line dictionaries.

The set of textual analysis functions (Figure 6) provides the textual analysis with the full commentary and critical apparatus. The former window shows the first sentence of the comment only by clicking on the ‘*more*’ button it switches to a separate slide where the complete text is given; the notes function is open by clicking on the relevant button on the text. Finally, a comparison between the numbering of fragments in different editions is provided.

4.3 TEI Compatibility

In view of preserving interoperability through TEI compatibility, fragments uploaded in the mock-up will be provided in the TEI format. This allows texts to be interoperable and linkable with other resources of interest. It enables links to vocabularies, ontologies and terminologies

⁶The use of tree-banks can improve didactic aspects and can improve the portfolio of competencies in relation to skill about language knowledge of the student.

[homepage](#)
[Previous slide](#)

30: Isolata la proposta di Giangrande 1972, 37-40, che, sulla base del confronto con Anacr. fr. 43 (PMG 338), 7 P. (= 82,7 Gent.), interpreta ἐν δορί come ἐν τῷ κύρῳ, relegando il poeta giambico nel solito cliché del trickster emarginato e sovversivo. Ad un inverosimile jeu d'esprit che adombri in generale oggetti riconducibili ad una manifattura lignea pensa invece Pocock 1961, 179s. [Less](#)

Fr. 2 W.²

ἐν δορί μὲν μοι μάζα μεμαγμένη, ἐν δορί
 δ' οἶνος Ἰσμαρικός· πίνω δ' ἐν δορί κεκλιμένος

Apparato, Varianti e Congetture:
 δ' ἐν δορί κεκλιμένος: δ' om. Syn. [Less](#)

Commento (Nicolosi, pp. 64-65):
 ἐν δορί κεκλιμένος: ammesso che la triplice
 occorrenza debba avere ogni volta lo stesso
 significato, diviene lecito domandarsi come
 sciogliere la contestata durezza del presunto
 nesso ἐν δορί κεκλιμένος [More](#) [Less](#)

62-64):
 le principali ipotesi
 oscillano tra il
 significato tradizionale di 'lancia'
 (cf., ad es., Hom. *Il.* II 382, V 40 *al.*,
 VIII 95 *al.*, X 31 *al.*, XI 96 *al.*, XIII
 130, XIV 494, XV 473 *al.*, XVI 114
al., XVII 7 *al.*, XX 423 *al.*, XXI 17 *al.*,
 XXII 112 *al.*, XXIII 893 *al.* e *Od.* X
 162 *al.*, XI 532, XIV 277, XIX 448)³¹
 – talora anche attribuendo al nesso
 il valore metaforico 'in armi', 'sotto
 le armi'³² – e quello ugualmente
 attestato, pur se meno frequente,
 di 'legno della nave' e quindi, per
 metonimia, 'nave' (cf., ad es., Hom.
Il. XV 410, XVII 744 e *Od.* IX 384)³³.
 Quest'esegesi, sostenuta –
praeunte Davison 1960,3 – da
 Bruno Gentili (1965, 129-134) e poi
 ribadita dallo studioso in contributi
 successivi³⁴, muove da una
 peculiare lettura di uno dei due
 principali testimoni, Syn. *Epist.* 130
 Hercher (= Garzya; Garzya-
 Roques)³⁵ ... [More](#) [Less](#)

fr.	W. ^{1,2}	A. ^{1,4}	Tard.	L.-B.	D. ^{1,3}	Crö.	B. ¹	B. ²	B. ³	B. ⁴	H.-C.	Schn.	Gaisf.	L.	Br.
2	2	2	2	7	2	3	2	2	3	2	2	2	45	56	-

Edizioni critiche:
 Br. = Brunck; L. = Liebel; Gaisf. = Gaisford; Schn. = Schneidewin; H.-C. = Hiller-Crusius; B.^{1,4} = Bergk^{1,4}; Crö. = Crönert; D.^{1,3} = Diehl^{1,3}; L.-B. = Lasserre-Bonnard; Tard. = Tarditi; A.^{1,4} = Adrados^{1,4}; W.^{1,2} = West^{1,2}

23/11/2017

Figure 6: Textual Analysis Functions

which are published as Linguistic Linked Open Data (LLOD), available in the semantic web world, and guarantees that data are searchable, augmentable, shareable, navigable, connectable⁷.

4.4 Evaluation

The mock-up allowed us to carry out an overall evaluation in view of designing a full digital edition. This was done with semi-quantitative criteria, by developing an appropriate metric and a questionnaire to evaluate users satisfaction with the mock-up. An extensive bibliography on methodologies for evaluating websites may be consulted: a recent work by (Fogli and Guida, 2015) provides a review of this bibliography together with a way to assess the quality of use of a website, encompassing a balanced set of features of the site, so as to mediate between the site owner's point of view (which obviously would like to save on implementation costs) and end-users requirements (they may require expensive developments in terms of ease of use).

The mock-up was submitted to an evaluation by a sample of prospective users in view of future developments, to better focus on the requirements from the product's perspective. Evaluation followed a proper protocol to collect feedback from a small sample of learners. This experimentation involved a group of university students, attending an MS-level Greek Philology course at the Parma University. Each student in the group was asked to write a small essay:

- using traditional bibliographic tools (task A)
- using the support of the mock-up (task B).

The student group was informed of the purpose of the experiment and was shown how to use the mock-up in advance. At the end of the test, each student filled in a questionnaire. The data collected through the questionnaire were analysed in order to highlight:

- the individual students' basic skills (through individual examination);

⁷(<https://www.w3.org/2005/Incubator/lld/wiki/Benefits>).

- the quality of the essay they produced (insufficient, satisfactory, good, excellent).

The final evaluation of the results is expected to be available at the end of 2018. A first evaluation was performed between September and December 2017. Results are positive and may be summarized as follows:

- the mock-up cannot fully substitute the book (neither it was intended to do it);
- however, it was judged to be a good complement of the book.

The majority of the interviewees considers the use of the mock up either useful or very useful, no one judge it to be not useful at all; a small percentage only (10%) considers it scarcely useful. Almost all the interviewees judge the mock-up easy and intuitive; two students only consider it too elaborate and un-natural or difficult to use. The mock-up at this stage is considered a good support to traditional study performed on the textbook, but not a substitute. Interviewees appreciate the availability of much more material with respect to the book (in form of text, apparatus, translations, witnesses, commentary, bibliography but also textual analysis, images and links to relevant sites). However, respondents also note some shortcomings: in particular, compared to the textbook, some overlapping layered windows prevent the simultaneous consultation of the relevant information and there is no way to insert commentary or study notes. Moreover, there are some difficulties similar to those in the printed book, such as, for example, the explanation of abbreviations in quotations. Finally, it is worth considering the replies about the difficulty of the tasks. Half of the interviewees judge the tasks A and B of the same difficulty. However, the majority of the remaining half of the respondents believe that performing task B (solved using mock-up support) was more complex than task A (solved using traditional bibliographic tools). In conclusion, at least half of the respondents consider the mock-up a useful study support.

The evaluation of the mock-up provides insight about usability of a full prototype, and then we will perform a cognitive walkthrough and a co-operative usability evaluation with 2-3 users. As the first evaluation step showed, we can already expect that it may be necessary to partly re-design the mock-up, at least from the point of view of the expected cognitive simplicity of the User Interface (i.e. about the screens that do not have clear visual structure, or that are unnecessarily complex), and the User Interface aesthetics, like Symmetry, Predictability, Economy, Proportion, and Simplicity, cfr. (Bhaskar et al., 2011), which are not fully implemented. User interface aesthetics, among other design considerations, are considered to be one of the determinants of user satisfaction. Once the design team has revised the mock-up accordingly, we plan to make it available through CLARIN-IT, together with the consultation questionnaire, in order to gather a broader outcome by the practitioners' community.

5 Conclusions

This paper presents and discusses the results of a survey carried out within the framework of CLARIN-IT in order to assess the actual use of digital resources and language technologies for Digital Humanities with respect to work in Ancient Greek scholarship. We concentrated, firstly, on the needs and requirements of Greek scholars, a large community manifesting complexity and an enormous level of heterogeneity. Our study, however, may also help in identifying gaps and drive the development of new technologies for ancient studies, thus addressing a set of R/D priorities that could be the base for establishing a consistent research and innovation agenda for Digital Classics, at large.

The focus is on the users; their feedback is paramount for identifying concrete factors that are likely to limit the uptake of new technology. Researchers must be primary actors in describing their expectations from digital data, tools and services in support of their studies, in view of enhancing existing resources and creating new resources and tools. The main message of the

Data Management Plan of PARTHENOS – one of the most important projects dealing with data science and aiming to build bridges and consolidating shared practices among the various domains of the humanities – states that: “the collection of user requirements and needs ... is based on the indications of a wide research community for the implementation of common policies and strategies” for managing data⁸.

The survey, as well as other surveys of digital methods in the sector carried out in the last decade, pointed out that most of the tools available do not pay enough attention to the criterion of usability. In some cases, there are researchers that either have not yet started using technology, or are in an early stage of doing so; some have difficulties in implementing them. To be successful, tools and services must be not only compatible with researchers’ workflows but, above all, must be reliable and easy to use (cf. also (Drude, 2016))⁹, combining the accuracy of traditional philology with a new, more intuitive and dynamic, still rigorous, approach.

A concrete action plan, emerging from the results of the survey, should lead to a workbench equipped with functionalities for inputting Greek lyrical text fragments in a simple and intuitive way and visualizing their encoding with specific TEI transcription; provide apparatus, literature and translation; link together primary sources and lexica; provide textual (and metrical) analysis and commentary, and offer search tools. To sum up, it is crucial to improve ancient studies by developing a common platform that responds to the desiderata of the scholars themselves.

CLARIN, the research infrastructure for the humanities and social sciences, can facilitate the take off of digital methods and solutions in various sectors and disciplines outside linguistics (insofar they have language as their object of study), such as philology. CLARIN, indeed, provides users with a variety of tools to analyse the data and delivers language services that, once integrated at an earlier stage of the process, may facilitate research tasks¹⁰. In addition, besides offering the opportunity to access and share data and tools, CLARIN represents the perfect framework from which to spread knowledge about the good practices related to a discipline (also with attention to possible educational aspects). Concluding, CLARIN is able to support and improve classical studies thanks to the application of concepts, methods and tools from the Digital Humanities to the field of Classics and of the study of the ancient languages.

⁸www.parthenos-project.eu

⁹This is in line with the PARTHENOS User Requirements report.

¹⁰We may remember, for instance, some VLO tools: “Lingua Interset” that is an universal set of morphosyntactic features, which all tagsets of all corpora/languages can be mapped to; The “TuTeAM corpus” that contains about 2800 entries from Ancient Greek and other modern languages; “Philostei” that is a system allows you to convert your book pages’ images into editable text (in TEI XML format); “Universal Dependencies” that is a project that seeks to develop cross-linguistically consistent tree-bank annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective; “The PROIEL Tree-bank” that is a dependency tree-bank with morphosyntactic and information-structure annotation.

References

- [Audenaert and Furuta2010] Neal Audenaert and Richard Furuta. 2010. What Humanists Want: How Scholars Use Source Materials. *JCDL '10: Proceedings of the 10th Annual Joint Conference on Digital Libraries*, 283–292. ACM, New York, NY. <http://dx.doi.org/10.1145/1816123.1816166>.
- [Babeu2011] Alison BAbau. 2011. Rome wasn't digitized in a day: Building a Cyberinfrastructure for Digital Classics. *Washington DC: CLIR - Council on Library and Information Resources*. <https://www.clir.org/pubs/reports/pub150/>.
- [Bhaskar et al.2011] N. Uday Bhaskar, P. Prathap Naidu, S.R. Ravi Chandra Babu and P.Govindarajulu. 2011. Principles of Good Screen Design in Websites *International Journal of Human Computer Interaction (IJHCI)*, vol. 2, 2, 48–57. <http://www.cscjournals.org/manuscript/Journals/IJHCI/Volume2/Issue2/IJHCI-25.pdf/>.
- [Drude2016] Sebastian Drude. 2016. PARTHENOS - Report on User Requirements. *PARTHENOS publishes the Report on User Requirements* March 1. <http://www.parthenos-project.eu/parthenos-publishes-the-report-on-user-requirements/>.
- [Fogli and Guida2015] Daniela Fogli and Giovanni Guida. 2015. WA practical approach to the assessment of quality in use of corporate web sites. *The Journal of Systems and Software*, vol. 99c, 52–65. <https://dl.acm.org/citation.cfm?id=2948289.2948354>.
- [Genette1982] Gérard Genette. 1982. Palimpsestes. La Littérature au second degré. *Parigi: Edition du Seuil*.
- [Monachini and Frontini2016] Monica Monachini and Francesca Frontini. 2016. CLARIN, l'infrastruttura europea delle risorse linguistiche per le scienze umane e sociali e il suo network italiano CLARIN-IT. *JCoL - Italian Journal of Computational Linguistics*, vol. 2, 2, 11–30. Special Issue on NLP and Digital Humanities.
- [Nicolosi2013] Anika Nicolosi. 2013. Archiloco. Elegie. *Bologna: Patron*.
- [Nicolosi, Monachini and Stefanini2017] Anika Nicolosi, Monica Monachini and Alberto Stefanini. 2017. HT-Archilochus (Hyper-Text Archilochus). *Digital Repository for the CLARIN Research Infrastructure provided by ILC-CNR*, October 1. <http://hdl.handle.net/20.500.11752/OPEN-83>.
- [Nicolosi, Monachini and Stefanini2017] Anika Nicolosi, Monica Monachini and Alberto Stefanini. 2017. Survey Data: Current practice of digital resources and tools for studies on Digital Classics. *Digital Repository for the CLARIN Research Infrastructure provided by ILC-CNR*, October 31. <http://hdl.handle.net/20.500.11752/OPEN-86>.
- [Pierazzo2014] Elena Pierazzo. 2014. Digital Scholarly Editing: Theories, Models and Methods. *HAL - Archives Ouvertes Fr.*, Grenoble.
- [Ruecker2008] S. Ruecker, M. Radzikowska, and S. Sinclair. 2011. Visual Interface Design for Digital Cultural Heritage: a Guide to Rich-Prospect Browsing. *Oxford: Routledge*.
- [Sahle2008] Patrick Sahle. 2008. A Catalogue of Digital Scholarly Editions. *v 3.0, snapshot 2008ff* <http://www.digitale-edition.de/>
- [Stefanini, Nicolosi and Monachini2017] Alberto Stefanini, Anika Nicolosi and Monica Monachini. 2017. Indagine e sperimentazione sulle pratiche d'uso di risorse e strumenti digitali nell'ambito della filologia greca. *Diss. Parma University* March 24.
- [Toms and O'Brien2008] Elaine Toms and Heather L. O'Brien. 2008. Understanding the information and communication technology needs of the e-humanist. *Journal of Documentation*, vol. 64, 102–130. <http://dx.doi.org/10.1108/00220410810844178>.
- [Warwick et al.2008] Claire Warwick, Isabel Galina, Melissa Terras, Paul Huntington and Nikoleta Pappa. 2008. The Master Builders: LAIRAH research on Good Practice in the Construction of Digital Humanities Projects. *Literary and Linguistic Computing*, vol. 23, 383–396. <http://discovery.ucl.ac.uk/13810/>.

Parliamentary Corpora in the CLARIN infrastructure

Darja Fišer

Department of Translation
Faculty of Arts, University of Ljubljana
Department of Knowledge Technologies,
Jožef Stefan Institute
darja.fiser@ff.uni-lj.si

Jakob Lenardič

Department of Translation
Faculty of Arts, University of Ljubljana
jakob.lenardic@ff.uni-lj.si

Abstract

This paper gives an overview of the parliamentary records and corpora from CLARIN countries with a focus on an analysis of their availability through the CLARIN infrastructure. Based on the results of the survey we provide a comprehensive overview of the corpora as well as draw a list of recommendations to optimize the depositing and cataloguing of the corpora in the CLARIN repositories in order to make them readily accessible for researchers from different disciplines. We also analyse the recall and precision of simple and faceted search of parliamentary corpora in the Virtual Language Observatory.

1 Introduction

Due to its unique content, structure and language, records of parliamentary sessions have always been a quintessential resource for a wide range of research questions from a number of disciplines in Digital Humanities and Social Sciences, such as Political Science (van Dijk 2010), Sociology (Cheng 2015), History (Pančur and Šorn 2016), Discourse Analysis (Hirst et al. 2014), Sociolinguistics (Rheault et al. 2015) as well as Multilinguality (Bayley et al. 2004). The good availability of parliamentary data in digitized form and granted access rights to public information in the EU countries have motivated a number of national as well as international initiatives to compile, process and analyse parliamentary corpora. The corpora were also the subject of a CLARIN-PLUS workshop¹ which aimed to bring together corpus developers and researchers using these resources. The aim of the workshop was to discuss technical issues related to proper structuring and archiving of such corpora and to address methodological questions about how to best use them in different disciplines. As examples of such use, the Finnish parliamentary corpus has already been successfully used in Discourse Analysis (Voutilainen, 2017), the Swedish corpus for the analysis of governmental policies related to Swedish film (Norén and Snickars, 2016), the Greek corpus for analyzing aggressive political discourse (Georgalidou 2017), the Norwegian corpus for developing dependency relations from LFG structures (Meurer, 2017) and the Lithuanian corpus for a stylometric analysis of parliamentary speech (Mandravickaitė and Krilavičius, 2015).

In order to gain an understanding how well the CLARIN infrastructure caters for this line of research, we conducted a survey for all member and observers CLARIN ERIC countries with which we aimed to identify the existing resources and check to which extent they are integrated in the CLARIN infrastructure. In this paper we provide a comprehensive presentation of the results and highlight aspects in which the accessibility of these corpora as well as the presentation of the relevant information can be optimised for researchers from different disciplines. Additionally, we evaluate the recall and precision of simple and faceted search of parliamentary corpora in the Virtual Language Observatory.

2 Corpora of parliamentary records within the CLARIN infrastructure

In total, we identified 15 corpora of national parliamentary data of CLARIN countries that are either available through a national repository or listed in the VLO. There exists one such corpus for each of the following 11 countries: The Czech Republic, Denmark, Estonia, Finland, France,

¹ <https://www.clarin.eu/event/2017/clarin-plus-workshop-working-parliamentary-records>.

Germany, Greece, Lithuania, Portugal, Slovenia and Sweden. There exist two corpora for Norway and Great Britain. We also took into account the Europarl corpus of the proceedings of the European Parliament, so there are 16 corpora in total.

In what follows, we describe each corpus in turn on the basis of the results of our survey² by providing the following information for each corpus:

- the name of the corpus, which also provides a hyperlink to the relevant handle where it can be accessed;
- the size of the corpus and the period it covers;
- the type of linguistic annotation included;³ and
- how the corpus is available (downloadable or through a concordancer or both).

2.1 Presentation of the CLARIN parliamentary corpora

The Hansard Corpus. This is the main corpus of the British Parliament. Covering the period between 1803 and 2005 and consisting of 1.6 billion tokens, it is the largest parliamentary corpus both in token size and temporal span. In addition to being tokenised, PoS-tagged, and lemmatised, the corpus is also characterized by deep semantic annotation pertaining to the classification of words based on historical concepts and thematic categories done with the *Historical Thesaurus Semantic Tagger* (Rayson et al., 2015). It is listed in the repository of the British observer CLARIN-UK and is presented as an online resource for querying through a dedicated concordancer. It is not listed in the VLO.

Parliamentary Debates on Europe at the House of Commons. This is the second, much smaller British parliamentary corpus. It is a thematically-focused corpus in that it contains only those parliamentary debates that correspond to the annual European Council meetings at the British parliament for the period between 1998 and 2005. It is roughly 190,000 tokens in size and its annotation consists of “mixed conversational analysis”. The corpus is available for download through the French ORTOLANG repository under CC-BY and is found on the VLO.

Czech Parliament Meetings. This corpus is the only parliamentary corpus that we have identified to consist of both transcripts and associated audio recordings – there are 88 hours of speech data from the Czech parliament for an unknown period corresponding to approximately 500,000 tokens. The annotation constitutes correction of errors, adding of proper punctuation and labelling of speech sections with information about the speaker. It is available for download on the website of the Czech repository LINDAT under the public CC-BY-NC-ND licence and is found on the VLO.

DK-CLARIN Almensprogligt korpus - offentlig del: tekster fra Folketinget. This corpus covers Danish parliamentary proceedings for the period between 2008 and 2010 and is 7.3 million tokens in size. It is available for download from the Danish repository CLARIN-DK under a non-specific public licence. As for annotation, the corpus is tokenised, PoS-tagged and lemmatised. The corpus is listed in the VLO.

Transcripts of Riigikogu (Estonian Parliament). This corpus consists of Estonian parliamentary proceedings from the period between 1995 and 2001 and is approximately 13 million tokens in size. It is unclear how the corpus is linguistically annotated. It is available for download on the corpus webpage and also accessible through the *Keeleveeb Query* concordancer provided by CLARIN-Estonia. It is listed in the VLO and is available under a non-specific academic licence.

Eduskunta Corpus. The *Eduskunta Corpus* is the corpus of Finnish parliamentary debates. There are 3 versions listed in the VLO:

- (1) Plenary Sessions of the Parliament of Finland, Downloadable Version 1
- (2) Plenary Sessions of the Parliament of Finland, Kielipankki Korp Version 1
- (3) Plenary Sessions of the Parliament of Finland, Kielipankki LAT Version 1

Each version covers Finnish parliamentary data for the period between 2008 and 2016. Versions (1) and (2) consist of the same data (2.2 million tokens), while the downloadable variant (1), which is available in the Finnish repository *Language Bank of Finland*, also provides the associated videos of the sessions. Corpus (1) is available for download under the CC-BY-NC-ND licence although the

² The complete results are available here: <https://office.clarin.eu/v/CE-2017-1019-Parliamentary-data-report.docx>

³ We list the annotation tools only in case the information is explicitly provided by the documentation.

access is restricted in that it requires a relevant institutional account – by contrast, corpus (2), while also available under CC-BY-NC-ND, is accessible through the *Korp* concordancer without restrictions. Corpus (3) is a variant of (1) that is currently in development and is set to be made freely available. It will provide reduced variants of the videos contained in the restricted version (1) processed through the *LAT* platform.⁴ The three versions of the *Eduskunta Corpus* are found on the VLO.

Parliamentary Debates on Europe at the Assemblée nationale. This is the French parliamentary corpus. Like the smaller British, it is a thematically-focused corpus in that it contains only those parliamentary debates that correspond to the annual European Council meetings at the French parliament for the period between 2002 and 2012. It is roughly 172,000 tokens in size and its annotation consists of “mixed conversational analysis”. The corpus is available for download through the French ORTOLANG repository under CC-BY and is listed in the VLO.

Parliamentary Debates on Europe at the Bundestag. Like their smaller British and French counterparts, the German parliamentary corpus is thematically-focused and contains only those parliamentary debates that correspond to the annual European Council meetings at the German parliament for the period between 1998 and 2005. It is roughly 417,000 tokens in size and its annotation consists of “mixed conversational analysis”. The corpus is available for download through the French ORTOLANG repository under CC-BY and is listed in the VLO.

Hellenic Parliament Sitzings. This corpus consists of Greek parliamentary proceedings from the period between 2011 and 2015 and is approximately 28.7 million tokens in size. The corpus is available for download under the academically-restricted CC-BY-NC licence from the Greek repository *clarin:el*. It is unclear how the corpus is annotated and it is not listed in the VLO.

Lithuanian Parliament Corpus for Authorship Attribution. This corpus consists of Lithuanian parliamentary proceedings from the period between 1990 and 2013 and is roughly 23.9 million tokens in size. In terms of annotation, the corpus is tokenised, PoS-tagged and lemmatised with *Lemuoklis*, which is a morphological analyzer for lemmatisation, and *MaltParser*, which was used for the generation of dependency tags (Kapočiūtė-Dzikiene, et al. 2015). The corpus is available for download through the CLARIN-LT repository under a non-specific public licence. It is found on the VLO.

Talk of Norway. This corpus is one of the two Norwegian parliamentary corpora. It covers the period between 1998 and 2016 and is 63.8 million tokens in size. It was annotated with the tools *angid.py* and *OBT* and is available for download through the *CLARINO* repository under the public NLOD licence. It is found on the VLO.

Proceedings of Norwegian Parliamentary Debates. This is the other Norwegian corpus. It covers the period between 2008 and 2015, consists of 29 million tokens and displays annotation in relation to the speaker, language variety, political party to which the speaker belongs and date and time. It is only available for online querying through the concordancer *Corpuscule*, also under the NLOD licence.

PTPARL Corpus. This corpus covers Portuguese parliamentary proceedings from the period between 1970 and 2008 and consists of 1 million tokens. It is tokenised, PoS-tagged and lemmatised with *LX-Tokenizer*, *LX-Tagger* (Branco and Silva, 2006), *MBT* and *MBLEM* (Généreux et al., 2012). The VLO entry links to a listing of the corpus in the ELRA catalogue,⁵ where the corpus is listed for download under a non-specific academic licence.

SlovParl. This is the Slovene corpus of parliamentary proceedings and there are two versions listed in the VLO – *Slovenian parliamentary corpus SlovParl 1.0* and *Slovenian parliamentary corpus SlovParl 2.0*. Both cover Slovene parliamentary proceedings for the period between 1990 and 1992 and differ from each other in the fact that the newer version is much larger in size – there are parliamentary proceedings from 54 sessions amounting to 2.7 million tokens in *SlovParl 1.0* in comparison with 232 sessions amounting to 10.8 million tokens in *SlovParl 2.0*. Both corpora are available for download under CC-BY in the *CLARIN.SI* repository and are extensively annotated (tokenisation, PoS-tagging and lemmatisation) with additional markup in relation to speaker and session typologies (Pančur and Šorn, 2016).

⁴ <https://lat.csc.fi/ds/asv/>

⁵ http://catalog.elra.info/product_info.php?products_id=1179

Riksdag's Open Data (Swedish: *Riksdagens öppna data*). This corpus, which in total consists of 1.25 billion tokens and is thus the second largest of the parliamentary corpora, is not listed in the VLO. Rather, it is only listed in the Swedish *Språkbanken* repository and is unique among the corpora in that there is no separate entry for the entire corpus but only for its subcorpora, of which there are 21 in total.⁶ Each subcorpus can be downloaded through the repository or – like the Finnish corpus – queried online through *Korp*. The corpus was tokenised, lemmatised, MSD-tagged (including additional markup in relation to semantic features, lemmagrams and compounding) with *Sparv*, which is the *Språkbanken*'s corpus annotation pipeline infrastructure (Borin et al., 2016). All the subcorpora are available under CC-BY.

Europarl. This is a multilingual parallel corpus of the sessions of the European Parliament. It contains documents from the period between 1996 and 2011 amounting to 588 million tokens. The corpus is sentence aligned and is freely available for download on a dedicated page under no specific licence. The corpus is listed in the VLO.

2.2 Summary and discussion

Table 1 summarizes the salient characteristics of the parliamentary corpora discussed in the previous subsection.

Table 1: Overview of the parliamentary corpora within the CLARIN infrastructure

NC	Size (mil tok)	Period	Anno	VLO	Ava il.	Location of avail.	Licence
uk ₁	1,600	1803-2005	T, PoS, L, additional semantic	/	C	External	/
uk ₂	0.19	1998-2015	Mixed conversational analysis	✓	D	ORTOLANG	CC-BY
cz	0.5	/	Semi-automatic alignment of transcriptions and audio records	✓	D	LINDAT	CC-BY
dk	7.3	2008-2010	T, PoS, L	✓	D	DK-CLARIN	Non-specific public
ee	13	1995-2001	TEI annotation	✓	D, C	External	Non-specific academic
fi	2.2	2008-2016	/	✓	D, C	FIN-CLARIN	CC-BY
fr	0.17	2002-2012	Mixed conversational analysis	✓	D	ORTOLANG	CC-BY
de	0.4	1998-2015	Mixed conversational analysis	✓	D	ORTOLANG	CC-BY
el	28.7	2011-2015	/	/	D	clarin:el	CC-BY
lt	23.9	1990-2013	T, PoS, L	✓	D	CLARIN-LT	CLARIN-LT public
no ₁	63.8	1998-2016	T, PoS, L	✓	D	CLARINO	NLOD
no ₂	29	2008-2015	Speaker, date, etc. markup	/	C	CLARINO	NLOD
pt	1	1970-2008	T, PoS, L	✓	D	External	Non-specific academic
si	10.8	1990-1992	T, PoS, L	✓	D, C	CLARIN.SI	CC-BY
se	1,250	1971-2016	T, L, PoS, semantic	/	D, C	SWE-CLARIN	CC-BY
eu	588	1996-2011	Sentence alignment, speaker markup	✓	D	External	/

⁶ cf. the resources listed under “Part of the Riksdag’s Open data” on <https://spraakbanken.gu.se/eng/resources>.

The parliamentary corpora are generally well integrated with the CLARIN infrastructure. Only 4 out of the total 16 corpora are not listed in the VLO; that is, the British *Hansard Corpus*, the Greek *Hellenic Parliament Sittings* corpus, the Norwegian *Proceedings of Norwegian Parliamentary Debates* and the Swedish *Riksdag's Open Data* corpus. *Riksdag's Open Data* is especially interesting in this respect since, as discussed in section 3.1, it consists of 21 subcorpora that are listed separately in the Swedish *Språkbanken* repository. Out of the 21 subcorpora, only one is listed in the VLO. This is the *Betänkande* subcorpus, which serves as a collection of summaries of voting results and decisions related to committee meetings.⁷ However, its metadata description in the VLO is fairly poor. Consequently, the subcorpus cannot be found with the most straightforward search queries like *parliament** or *parliament* corpus*. The remaining 20 subcorpora are not listed in the VLO and it is not immediately obvious why this is so.

In terms of availability, the majority – that is, 10 corpora (the Czech, Danish, Greek, Lithuanian, Portuguese, French, German corpora, the Norwegian *Talk of Norway* Corpus, the British *Parliamentary Debates on Europe at the House of Commons* corpus and *Europarl*) – are only for download. All of these corpora can be downloaded from a relevant CLARIN repository (e.g. *Czech Parliament Meetings* through LINDAT) except for *Europarl*, which is available on a dedicated webpage, and the Portuguese *PTPARL Corpus*, which is listed in the ELRA catalogue. 4 corpora can both be downloaded and queried through a concordancer. These are the Estonian *Transcripts of Riigikogu* corpus, which is available for download on a dedicated webpage and can be accessed through the *Keeleveeb Query* concordancer provided by CLARIN Estonia; the Finnish *Eduskunta Corpus*, which can be downloaded through the Finnish repository *Language Bank of Finland* and accessed through *Korp*; the Slovene *SlovParl* corpus, which can be downloaded through the *CLARIN.SI* repository and queried online through *noSketchEngine*;⁸ and *Riksdag's Open Data*, which can be downloaded from the *Språkbanken* repository and queried through *Korp*. 2 corpora can only be queried online – while the Norwegian parliamentary corpus *Proceedings of Norwegian Parliamentary Debates* is searchable through a concordancer provided by a CLARIN repository (that is, the *CLARINO Corpuscle* concordancer), the other corpus, *The Hansard Corpus*, is queried through a non-CLARIN dedicated concordancer.

The availability and thoroughness of metadata documentation is also fairly good. Information on size is available for all corpora, while information on the temporal period is missing only for *Czech Parliament Meetings*. Information on the level of linguistic annotation is likewise mostly readily available, missing only for the Greek *Hellenic Parliament Sittings* corpus and the Finnish *Eduskunta Corpus*. However, the location of the information on annotation is far from uniform – for instance, the description field of the *Lithuanian Parliament Corpus for Authorship Attribution* on the CLARIN-LT repository does not mention annotation (described only in one of the downloadable corpus files), while the *Riksdag's Open Data* subcorpora are systematically described in *Språkbanken*.

Licence information is also in most cases readily included, with most of the corpora being available under CC-BY. Here we would like to stress that there exists a slight discrepancy between the information as it is presented in the VLO on the one hand and on the relevant landing page on the other in the case of the *PTPARL Corpus*, *Lithuanian Parliament Corpus for Authorship Attribution* and *Talk of Norway*. In the relevant VLO entries for these three corpora, the licence is listed as unknown even though it is specified on the relevant landing pages – for instance, in the case of the *Talk of Norway* corpus, the licence in the CLARINO repository is specified as NLOD, which is public, so the VLO entry should follow suit and also list the corpus as publicly available.

⁷ <https://repo.spraakbanken.gu.se/xmlui/handle/10794/83>

⁸ https://www.clarin.si/noske/run.cgi/first_form

2.3 Additional national parliamentary corpora not in the VLO or CLARIN repositories

In our original survey, we identified 7 additional national parliamentary corpora, one corpus per each of the following countries: Austria, Bulgaria, the Czech Republic, the Netherlands, Germany, Latvia and Poland. However, these corpora are neither listed in the VLO nor available through a CLARIN repository, so were omitted from the current discussion. We provide a brief overview of these corpora, as they are generally well presented and would serve as welcome inclusions in relevant CLARIN repositories and the VLO. Note that the bolded names of the corpora below provide hyperlinks to relevant landing pages.

- **Korpusbasierte Analyse österreichischer Parlamentsreden.** Austrian parliamentary corpus for 2013-2015; 1.2 million tokens; tokenised and PoS-tagged (Sippl et al., 2016);
- **Corpus of Bulgarian Political and Journalistic Speech.** Bulgarian parliamentary corpus for 2006-2012; 10 million tokens; tokenised, PoS-tagged, and lemmatised;
- **CzechParl.** Czech parliamentary corpus for 1993-2010; 81.9 million tokens; tokenised, MSD-tagged, and lemmatised (Jakubíček and Kovár, 2010);
- **DutchParl.**⁹ Dutch parliamentary corpus for 1814-2014; 800 million tokens; tokenised, PoS-tagged and lemmatised (Marx and Schuth, 2010);
- **polmineR corpus.** German parliamentary corpus; size, period and annotation unknown;
- **SEIMA corpus.** Latvian parliamentary corpus for 1993-2016; unclear size; unclear annotation; and
- **Polish Parliamentary Corpus.**¹⁰ Polish parliamentary corpus for 1991-2017; 300 million tokens; tokenisation, MSD-tagging, lemmatisation, utterance-level segmentation, named entities (Ogrodniczuk, 2012).

3 Identifying parliamentary corpora through the VLO¹¹

In this section, we discuss the identification of the 12 VLO corpora both in terms of simple queries and the faceted search option and highlight problematic aspects.

3.1 Simple search

We first focus our discussion on simple search (in other words, using only the search field) on the basis of two salient search strings – *parliament** and *parliament* corpus*.

In the case of the simple search string *parliament**,¹² Table 2 lists the top 20 search results:

⁹ The Dutch CLARIN consortium has been involved in important projects on parliamentary data. One such project is *War in Parliament* (<http://www.clarin.nl/node/410>); another is the *DiLiPaD* project (<http://dilipad.history.ac.uk/>), which applies Linked Data to British, Dutch and Canadian parliamentary proceedings. Resulting datasets are available through the online PoliticalMashup environment (<http://politicalmashup.nl/>).

¹⁰ This corpus is set to be included in the CLARIN-PL D-Space repository in June 2018 (Ogrodniczuk, personal correspondence).

¹¹ The results presented in this section reflect VLO version 4.3.2. from January 2018.

¹² https://vlo.clarin.eu/?q=parliament*

Table 2: Results for the search string *parliament. The bolded results correspond to a subset of the national parliamentary corpora described in section 2.1.** ¹³

#	Name of VLO entry
1	Plenary Sessions of the Parliament of Finland, Kielipankki Korp Version 1
2	Plenary Sessions of the Parliament of Finland, Downloadable Version 1
3	Slovenian parliamentary corpus SlovParl 2.0
4	Slovenian parliamentary corpus SlovParl 1.0
5	Plenary Sessions of the Parliament of Finland
6	Czech Parliament Meetings
7	Corpus of the Proceedings of Estonian Parliament
8	TC-STAR Transcriptions of Spanish Parliamentary Speech
9	European Parliament Interpretation Corpus (EPIC)
10	Information in Sign Language on the Tasks of the Parliamentary Ombudsman of Finland
11	Lithuanian Parliament Corpus for Authorship Attribution
12	Europarl: European Parliament Proceedings Parallel Corpus 1996-2003
13	Corpus of the Proceedings of Estonian Parliament
14	Parliamentary Debates on Europe at the House of Commons (1998-2015)
15	Dataset of European Parliament roll-call votes and Twitter activities MEP 1.0
16	Parliamentary Debates on Europe at the Bundestag (1998-2015)
17	Parliamentary Debates on Europe at the Assemblée nationale (2002-2012)
18	Plenary Sessions of the Parliament of Finland, Kielipankki LAT Version 1
19	On the legislative authority of the British Parliament [Electronic resource] / James Wilson
20	The Present State Of Westminster Bridge: Containing A Description [...]

With this search string we are able to find the Finnish corpus (results 1, 2, 18), the Slovene corpus (results 3, 4), the Czech corpus (result 6), the Estonian corpus (results 7, 13), the European parliament corpus (result 12), the Lithuanian corpus (result 11) and the three thematic parliamentary corpora (results 14, 16, 17), so 13 hits for 9 out of the total 12 VLO corpora. The discrepancy between the higher number of VLO entries and smaller number of actual corpora is due to the fact that 3 corpora – that is, the Finnish, Slovenian, Estonian corpora – are associated with more than one VLO entry each. While this is to be expected in the case of the Finnish and Slovene corpus since each result corresponds to a different version, in the case of the Estonian corpus (results 7, 13), the two VLO entries point to exactly the same resource, so the lower-ranked version, which differs from the higher-ranked one only in that it contains a less detailed metadata description, is likely redundant. All in all, 75% of the corpora described in section 2.1 can be found by using the simple search string, but not all.¹⁴

We now turn to the phrasal search string *parliament* corpus*.¹⁵ As shown below, the relevant results are more scattered in comparison with the previous query and are in several cases ranked below the 20th hit.

¹³ After (18), the results of the simple *parliament** query begin corresponding to resources which turn out to be entries for singleton documents like specific letters, or transcriptions of a single speech, so not collections of a series of proceedings and are thus not relevant for our survey.

¹⁴ We believe that the the remaining three corpora (the Danish *DK-CLARIN Almensprogligt korpus - offentlig del: tekster fra Folketinget*, the Norwegian *Talk of Norway*, and the Portuguese *PTPARL Corpus*) are not yielded by this search string because the name or their metadata description does not contain the term *parliament(ary)*.

¹⁵ https://vlo.clarin.eu/?q=parliament*+corpus

Table 3: Results for the search string *parliament* corpus*¹⁶

#	Name of VLO entry
1	Lithuanian Parliament Corpus for Authorship Attribution
2	PTPARL Corpus
3	Slovenian parliamentary corpus SlovParl 2.0
4	Slovenian parliamentary corpus SlovParl 1.0
5	European Parliament Interpretation Corpus (EPIC)
6	Amaryllis Corpus - Evaluation Package
7	KOTUS Finnish-Swedish Parallel Corpus
8	Europarl: European Parliament Proceedings Parallel Corpus 1996-2003
9	GeFRePaC - German French Reciprocal Parallel Corpus
10	Corpus of the Proceedings of Estonian Parliament
11	Corpus of the Proceedings of Estonian Parliament
12	Grenelle II - Subpart 1: audio/video
13	Grenelle II - Subpart 2: audio/video
14	Grenelle II on environnement: multimodal annotation
15	Grenelle II - Subpart 2: audio/video
16	Grenelle II - Subpart 1: audio/video
17	Grenelle II on environnement: multimodal annotation
18	Corpus of Early Modern English Statutes 1491-1707
19	Helsinki Corpus of Swahili 2.0 (HCS 2.0) Annotated Version
20	Parliamentary Debates on Europe at the House of Commons (1998-2015)
21	Parliamentary Debates on Europe at the Bundestag (1998-2015)
22	Parliamentary Debates on Europe at the Assemblée nationale (2002-2012)
23	Plenary Sessions of the Parliament of Finland, Downloadable Version 1
24	[...]
25	Plenary Sessions of the Parliament of Finland, Kielipankki LAT Version 1
26	[...]
27	Plenary Sessions of the Parliament of Finland, Kielipankki Korp Version 1
28	Czech Parliament Meetings
29-45	[...]
46	Talk of Norway

On a positive note, this complex search string lists two more corpora than the simpler search *parliament**; that is, the Portuguese *PTPARL Corpus* (result 2) and the Norwegian *Talk of Norway* corpus (result 46). However, notice the relative low ranking of the three versions of the Finnish *Eduskunta Corpus* (results 23, 25, 27) and *Czech Parliament Meetings* (result 28) in comparison with the simpler search string in Table 2, in which case they were listed as top results. Why this is so is unclear, as all of these low-ranked entries have rich and granular metadata in which the term *corpus* is both used in the general resource description tab in the VLO and is selected as the value for *resourceType* under the extensive *All metadata* tab.¹⁷ By contrast, resources like the subparts of the *Grenelle* collection (results 12-17) display a poorer metadata description in which the term *corpus* is not used in the general description tab nor is it specified as the value for *resourceType*. Such results would be expected to be ranked lower than the three versions of the Finnish corpus and *Czech Parliament Meetings*. It is also unclear why the *LAT* version of the *Eduskunta Corpus* (result 25) is ranked higher than the *Korp* version (result 27), since it is, as described in section 2.1, still under development and therefore its VLO entry lacks a detailed metadata description in comparison with that of the *Korp* version. Additionally, *Talk of Norway* comes up as a very low-ranked result (46), so

¹⁶ The numbering in the table below again corresponds to the ranking in the VLO and, for reasons of space, we omit the irrelevant results beyond (20).

¹⁷ As of version 4.3.2. of the VLO, terms in phrasal search strings are conjoined by the AND operator, so a search string like *parliament* corpus* should provide an intersection of resources that pertain to the *parliament** search string and those resources that pertain to the *corpus* search string.

it is unlikely that a potential researcher would find easily. The fairly scattered presentation of the results in Table 3 is counterintuitive for anyone interested in finding parliamentary corpora with such a straightforward string as *parliament* corpus*.

3.2 Faceted search

In the VLO, resources can also be found by means of faceted search, which allows the user to narrow down a general query by applying filters under various facets such as Language, Resource Type, Genre, Modality, Subject and so forth. Following Odijk (2014), we focus on two facets that seem to be the most relevant for narrowing our search down to relevant parliamentary corpora. These are Resource Type, which should in the case of the simple search string *parliament** give us the option to select *corpus* as a value, and Subject, which should presumably narrow the search down to subjects related to parliamentary data.

In the case of the simple search string *parliament**, the Resource Type facet returns the following list of values, with the number of resources for each value listed in the parentheses: *Text* (51), *Sound* (37), *Info:eu-repo/semantics/dataset* (3), *Bioscoop* (2), *Corpus* (2), *Politics* (2), *Boek* (1), *Dataset* (1), and so forth. Surprisingly, the value *Corpus* lists only two resources: *Czech Parliament Meetings* and *Europarl: European Parliament Proceedings Parallel Corpus 1996-2003*. In other words, the vast majority of the corpora we can identify with the simple query *parliament** (cf. Table 2) are not captured in this facet and it is unclear why this is so, especially since a resource like *Plenary Sessions of the Parliament of Finland, Downloadable Version 1* has the value *corpus* under *resourceType* in the metadata description. Similarly, recall from section 2.1 that *Czech Parliament Meetings* is a corpus of audio records, yet selecting the value *Sound* fails to list it. In short, narrowing the search down through Resource Type does not yield the desired results.

Sticking to the same simple search string, the Subject facet presents the following list of values: *text_and_corpus linguistics* (37), *corpus* (6), *audio* (4), *débat politique* (4), *political debate* (4), *video* (4), *vidéo* (4), *discours politique* (3), *débats parlementaires* (3), *europe* (3) and so forth. On the one hand, values for the same type of subject are given twice (e.g. *discours politique* and *débat politique*) and selecting one value filters the results of the other. On the other hand, several values are clearly more suitable for the Resource Type facet, yet selecting for instance *corpus* (6) does not yield any of the parliamentary corpora in Tables 2 and 3, nor does selecting *audio* (4) yield *Czech Parliament Meetings*, contrary to expectations. Our findings correspond with Odijk's experience (2014).

4 Discussion and proposals

As parliamentary corpora are of great value for researchers from a wide range of disciplines and a lot of effort had already been invested in producing them, we propose that their developers and curators adopt the following suggestions to make them better accessible through the CLARIN infrastructure:

- create a virtual collection pointing to a landing page (ideally with a PID) for the corpora;
- add the missing corpora listed in section 2.3 to the repository of a certified CLARIN centre after which they will be automatically added to the VLO via metadata harvesting;
- improve the metadata of the existing corpora in order to make them more accessible for the end user.

For improving the metadata, follow the best practices below:

- use *parliament(ary)* in the title of the metadata file, so that it gets included in target queries (e.g. https://vlo.clarin.eu/?q=name:parliament*);
- use the word *parliament(ary)* in the title (and description) and provide descriptions in English that include one of these words or an equivalent term, which will lead to higher ranking;
- use a distinctive title (not e.g. 148 times Flemish parliamentary debate <https://vlo.clarin.eu/?q=Flemish+parliamentary+debate>);
- when providing highly granular metadata descriptions (many + detailed), make sure to use hierarchies (cf. <https://www.clarin.eu/faq/how-can-i-create-hierarchical-collection-cmdi>) so that the top node appears first in the VLO);

- include licencing information, which also helps with the ranking of hits in VLO, especially if the level is/maps to PUB or ACA.

However, a bigger limitation that needs immediate attention seems to be the VLO. We have shown that identifying parliamentary corpora in the current version of VLO is counterintuitive, since the best results (Table 2) are yielded by the simplest search string, whereas further specifications either by a narrower search string (Table 3) or the use of faceted search yields in substantially lower precision as well as recall. Additionally, while some corpora like *Talk of Norway* or the Danish *DK-CLARIN Almensprogligt korpus - offentlig del: tekster fra Folketinget* are listed in the VLO, they can only be identified after querying the full corpus name, which we believe is a limitation since many users will not know the official name of the resource they are looking for. On the other hand, a handful of highly relevant corpora, like the Swedish *Riksdag's Open Data* and the British *Hansard Corpus*, are not listed in the VLO at all, only in the national repositories, and we believe their inclusion would be beneficial for the comprehensive representation of CLARIN parliamentary corpora in the VLO.

5 Conclusion

In this paper we presented a survey of the parliamentary corpora in the CLARIN infrastructure. We have been able to find corpora for all the countries except Italy. While this is commendable, our survey highlights that not all the essential information about the corpora is easily available and, most importantly, that most of the existing corpora cannot easily be found through the Virtual Language Observatory. For this reason, we have drawn up a list of recommendations for corpus metadata in order to improve findability and ranking of the corpora by VLO as well as documented issues with the VLO that should be taken into account in future development of the service. This is of paramount importance as the VLO is the main gateway to the invaluable CLARIN resources.

In the future, we plan to create a Virtual Collection with all the identified parliamentary corpora and develop a model to ensure interoperability of the corpora and integrate them into a common concordancer in order to make them as readily accessible for researchers from different disciplines as well as for cross-border and cross-lingual projects which is where CLARIN is in the unique position to facilitate such endeavours. With this in mind, we will also collect showcases of successful applications of parliamentary corpora in Digital Humanities and Social Sciences, as such information valuably complements the corpora. We also plan to conduct a follow-up survey in order to evaluate the effect of the proposed recommendations as well as the uptake of the improved resources at regular intervals.

6 References

- [Bayley et al. 2004] Paul Bayley, Cinzia Bevitori, Elisabetta Zoni. 2004. Threat and fear in parliamentary debates in Britain, Germany and Italy, *Cross-Cultural Perspectives on Parliamentary Discourse*, 185-236.
- [Borin et al. 2016] Lars Borin, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer, Anne Schumacher. 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. http://www8.cs.umu.se/~johanna/sltc2016/abstracts/SLTC_2016_paper_31.pdf. Last accessed on 11 January 2018.
- [Branco and Silva 2006] António Branco and João Silva. 2006. A Suite of Shallow Processing Tools for Portuguese: LX-Suite, In *Proceedings of EACL2006 – 11th Conference of the European Chapter of the Association for Computational Linguistics*, 179–182.
- [Cheng 2015] Jennifer E Cheng. 2015. Islamophobia, Muslimophobia or racism? Parliamentary discourses on Islam and Muslims in debates on the minaret ban in Switzerland. <http://journals.sagepub.com/doi/pdf/10.1177/0957926515581157>.
- [van Dijk 2010] Teun A. van Dijk. 2010. Political Identities in Parliamentary Debates. <http://www.discourses.org/OldArticles/Political%20Identities%20in%20Parliamentary%20Debates.pdf>.

- [Généreux et al. 2012] Michel Généreux, Iris Hendrickx, Amália Mendes. 2012. “A Large Portuguese Corpus On-Line: Cleaning and Preprocessing.” *Conference: Computational Processing of the Portuguese Language (PROPOR)*.
- [Georgalidou 2017] Marianthi Georgalidou. 2017. Using the Greek parliamentary speech corpus for the study of aggressive political discourse. <https://www.clarin.eu/sites/default/files/4-georgalidou.pdf>.
- [Hirst et al. 2014] Graeme Hirst, Vanessa Wei Feng, Christopher Cochrane, Nona Naderi. 2014. Argumentation, Ideology, and Issue Framing in Parliamentary Discourse. <http://eur-ws.org/Vol-1341/paper6.pdf>.
- [Jakubiček and Kovář 2010] Miloš Jakubiček, Vojtěch Kovář. 2010. “CzechParl: Corpus of Stenographic Protocols from Czech Parliament”. In P. Sojka, A. Horák (eds.) *RASLAN 2010 Recent Advances in Slavonic Natural Language Processing*.
- [Kapočiūtė-Dzikienė et al. 2015] Jurgita Kapočiūtė-Dzikienė, Andrius Utkla, Ligita Šarkutė. 2015. “Authorship attribution of internet comments with thousand candidate authors.” *ICIST 2015 : 21st International Conference on Information and Software Technologies*, 433-448. Springer International Publishing.
- [Mandravickaitė and Krilavičius 2015] Justina Mandravickaitė, Tomas Krilavičius. 2015. Language usage of members of the Lithuanian Parliament considering their political orientation. *Deeds and Days* 64: 133-151.
- [Meurer 2017] Paul Meurer. 2017. From LFG structures to dependency relations. *Bergen Language and Linguistic Studies* 8: 183-201.
- [Marx and Schuth 2010] Maarten Marx and Anne Schuth. “DutchParl: The Parliamentary Documents in Dutch.” <http://politicalmashup.nl/new/uploads/2010/03/lrecfinalversionlong.pdf>. Last accessed on 7 January 2018.
- [Norén and Snickars 2016] Fredrik Norén, Pelle Snickars. 2016. Distant Reading the History of Swedish Film Politics—in 4,500 Governmental SOU Reports. <http://pellesnickars.se/2016/12/distant-reading-the-history-of-swedish-film-politics-in-4500-governmental-sou-reports/>
- [Odijk 2014] Jan Odijk. 2014. “Discovering Resources in CLARIN: Problems and Suggestions for Solutions.” <http://www.clarin.nl/sites/default/files/Searching%20with%20the%20VLO.pdf>. Last accessed on 11 January 2017.
- [Ogrodniczuk 2012] Maciej Ogrodniczuk. 2012. “The Polish Sejm Corpus.” http://www.lrec-conf.org/proceedings/lrec2012/pdf/653_Paper.pdf. Last accessed on 8 January 2018.
- [Oravecz et al. 2014] Csaba Oravecz, Tamás Váradi, Bálint Sass. 2014. “The Hungarian Gigaword Corpus.” http://www.lrec-conf.org/proceedings/lrec2014/pdf/681_Paper.pdf. Last accessed on 10 January 2018.
- [Pančur and Šorn 2016] Andrej Pančur, Mojca Šorn. 2016. Smart Big Data: use of Slovenian parliamentary papers in digital history, *Prispevki za novejšo zgodovino*, 56:3, 130-146.
- [Rheault et al. 2015] Ludovic Rheault, Kaspar Beelen, Christopher Cochrane, Graeme Hirst. 2015. Measuring Emotion in Parliamentary Debates Using Methods of Natural Language Processing. <http://www.cs.toronto.edu/pub/gh/Rheault-et-al-CPSA-2015.pdf>.
- [Voutilainen 2017] Eero Voutilainen. 2017. Parliamentary Records as Data for Linguistic Discourse Studies. http://videolectures.net/clarinplusworkshop2017_voutilainen_studies/.
- [Rayson et al. 2015] Paul Rayson, Alistair Baron, Scott Piao, Steve Wattam. 2015. “Large-scale Time-sensitive Semantic Analysis of Historical Corpora.” http://ucrel.lancs.ac.uk/samuels/papers/SAMUELS_ICAME36_Software_Demo_Handout.pdf. Last accessed on 7 January 2018.
- [Sippl et al. 2016] Colin Sippl, Manuel Burghardt, Christian Wolff, Bettina Mielke. 2016. “Korpusbasierte Analyse österreichischer Parlamentsreden.”

Many a Little Makes a Mickle – Infrastructure Component Reuse for a Massively Multilingual Linguistic Study

Lars Borin
University of Gothenburg
Sweden

Shafqat Mumtaz Virk
University of Gothenburg
Sweden

Anju Saxena
Uppsala University
Sweden

`lars.borin@svenska.gu.se, virk.shafqat@gmail.com, anju.saxena@lingfil.uu.se`

Abstract

We present ongoing work aiming at turning the linguistic material available in Grierson’s classical *Linguistic Survey of India* (LSI) into a digital language resource, a database suitable for a broad array of linguistic investigations of the languages of South Asia and studies relating to language typology and contact linguistics. The project has two concrete main aims: (1) to conduct a linguistic investigation of the claim that South Asia constitutes a linguistic area; (2) to develop state-of-the-art language technology for automatically extracting the relevant information from the text of the LSI. In this presentation we focus on how, in the first part of the project, a number of existing research infrastructure components provided by Swe-Clarín, the Swedish CLARIN consortium, have been ‘recycled’ in order to allow the linguists involved in the project to quickly orient themselves in the vast LSI material, and to be able to provide input to the language technologists designing the tools for information extraction from the descriptive grammars.

1 Introduction: South Asian Linguistics and the *Linguistic Survey of India*

1.1 South Asian Linguistics and the Areal Hypothesis

South Asia (also “India[n subcontinent]”) with its rich and diverse language ecology and a long history of intensive language contact provides abundant empirical data for studies of linguistic genealogy, linguistic typology, and language contact.

This region (normally understood in linguistic works as comprising the seven countries Bangladesh, Bhutan, India, the Maldives, Nepal, Pakistan, and Sri Lanka, as well as adjacent areas in neighboring countries, since language boundaries do not always coincide with national borders) is the home of hundreds of languages spoken by almost two billion people – more than a quarter of the world’s population. Most of the 661 living languages of South Asia (Simons and Fennig, 2018) are from four major language families (Indo-European>Indo-Aryan and Nuristani, Dravidian, Austroasiatic>Munda, Khasian and Nicobaric, and Tibeto-Burman (Sino-Tibetan); see Figure 1). In addition there are some language isolates and small families (Georg, 2017) and several creoles and pidgins.

South Asia is often referred to as a *linguistic area*, a region where, due to close contact and widespread multilingualism, languages have influenced one another to the extent that both related and unrelated languages are more similar on many linguistic levels than we would expect. However, with some rare exceptions (e.g., Masica, 1976) most studies are largely impressionistic, drawing examples from a few languages only (Ebert, 2006).

In this paper we present our ongoing work aiming at turning the linguistic material available in Grierson’s classical *Linguistic Survey of India* into a digital language resource, a database suitable for a broad array of linguistic investigations of the languages of South Asia, and especially for conducting a more thorough scrutiny of the South Asian areal hypothesis.

Given the CLARIN context, we will focus on some research infrastructural aspects of our work here, notably how the project was able to reap great benefits from repurposing existing infrastructure components provided by Swe-Clarín, the Swedish CLARIN consortium.

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

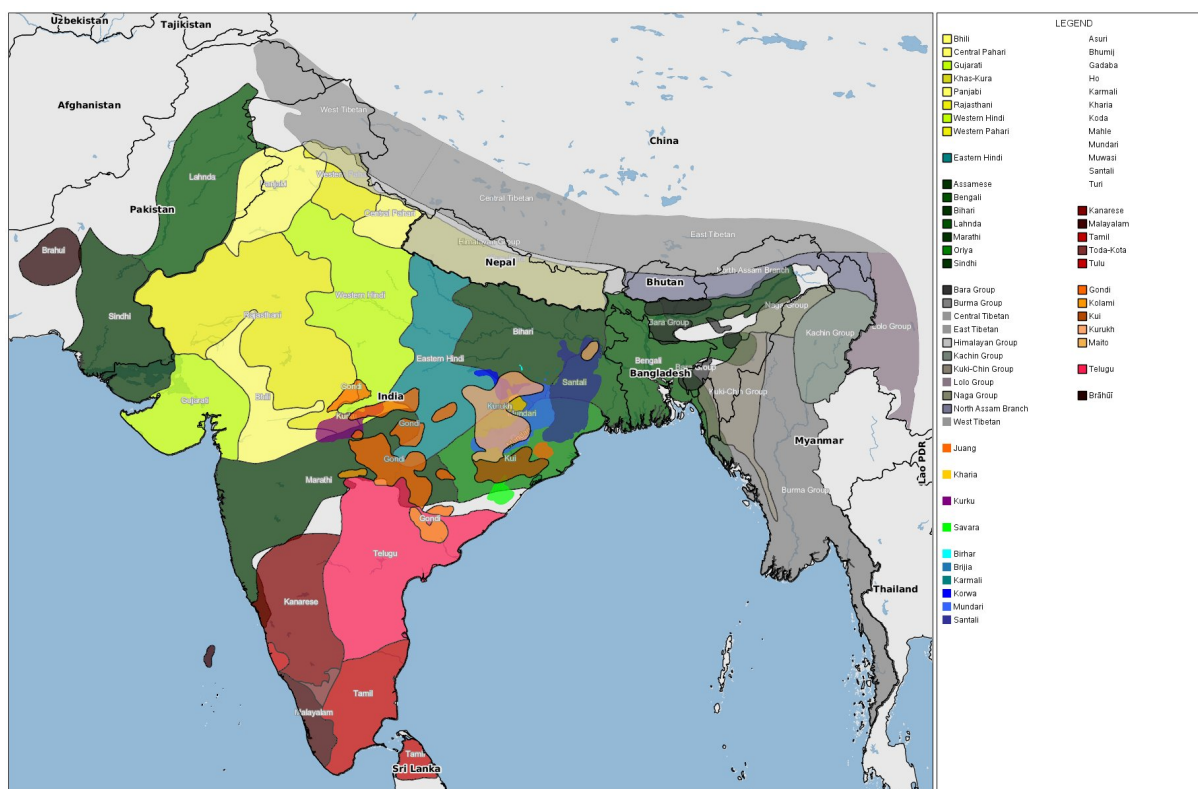


Figure 1: The four major language families of South Asia (from <http://llmap.org>)

1.2 Grierson's *Linguistic Survey of India*

The linguistic richness and diversity of South Asia was documented by the British government in a large-scale survey conducted in the late nineteenth and the early twentieth century under the supervision of Sir George Abraham Grierson and Sten Konow. The survey resulted into a detailed report comprising 19 volumes of around 9,500 pages in total, entitled *Linguistic Survey of India* (LSI; Grierson, 1903–1927). The survey covered 723 linguistic varieties representing the major language families of the region and some unclassified languages, of almost the whole of nineteenth-century British-controlled India (modern Pakistan, India, Bangladesh, and parts of Burma). For each major variety it provides (1) a grammar sketch (including a description of the sound system); (2) a core word list; and (3) text specimens (including a morpheme-glossed translation of the *Parable of the Prodigal Son*).

The LSI grammar sketches provide basic grammatical information about the languages in a fairly standardized format. The focus is on the sound system and the morphology (nominal number and case inflection, verbal tense, aspect, and argument indexing inflection, etc.), but as we will see below in Section 3, there is also some syntactic information to be found in them. Importantly, the sketches include information on some of the features that have been used in defining South Asia as a linguistic area, e.g. retroflexion, reduplication, compound verbs, word order, converbs/conjunctive participles, but go considerably beyond these, offering the possibility of a broad comparative study of South Asian languages.

The grammar sketches range in length from less than a page to over eighty pages, and the whole LSI comprises far too much text for it to be a realistic option to process it manually. Arguably, this language data qualifies as “big data”, not primarily by virtue of its volume, but by virtue of its inherent complexity and the tools needed to process it (Ward and Barker, 2013), especially for extracting and comparing grammatical features.

Thus, we are currently exploring information extraction methodologies which could help us turning the free-text descriptions of the LSI grammar sketches into formally structured tabular data suitable for large-scale automatic processing. At the present time, this is the main NLP focus of the project. Since

the grammatical descriptions are written in English, this of course means that the information extraction application that we are developing will be for English (see Section 3 below).

The core word lists which accompany the language descriptions are collected in a separate volume (Volume 1, Part 2: *Comparative vocabulary*). Each list holds a total of 168 entries. Most of the entries in the comparative vocabulary render concepts which cover a broad spectrum consisting of body parts, domestic animals, personal pronouns, numerals, and astronomical objects. There is some overlap with other concept lists used in language classification: For instance, 38 of the concepts are also found in the shorter (100-item) version of the so-called *Swadesh lists*, core vocabulary lists originally devised by the American linguist Morris Swadesh (1955) specifically for the purpose of inferring genealogical relationships among languages. Thus, the LSI comparative vocabulary clearly has one part that can be used in investigating genetic connections among the languages, but also another part – at least half of the entries – which we hypothesize could be used to find areal influences.

Notably, the LSI comparative vocabulary also provides some phrases and propositions (e.g., ‘good man’ ~ ‘good woman’ ~ ‘good men’ ~ ‘good women’, and ‘I, thou, etc. go’ ~ ‘I, thou, etc. went’), making it useful for comparative studies of some grammatical features, in addition to studies of lexical phenomena. In a preliminary study, some grammatical features have been semiautomatically extracted from the comparative vocabulary, and used as a kind of “silver standard” in some of our information extraction experiments.

The language data for the LSI grammar sketches were collected around the turn of the 20th century, hence obviously reflecting the state of these languages of about a century ago. However, we know that many grammatical characteristics of a language are quite resistant to change (Nichols, 2003), much more so than vocabulary. In order to get an understanding of the usefulness of the LSI for our purposes, we sampled information from a few of the sketches in order to see how well the LSI data reflect modern language usage. Our results show that while some of the lexical items are not used today in everyday speech, most other information reflects in many ways the modern language, and thus cannot be treated as representing an ‘archaic’ variety of, e.g., Hindi.

Despite its age, LSI still remains the most complete single source on South Asian languages. It has been used in a few studies with varying aims and objectives, but has not been exploited to the extent it could have been, important reasons arguably being its vast size and limited accessibility. This multi-volume work will generally be found only in select research libraries and it does not have any kind of index of its contents. A scanned version of LSI is now available on the University of Chicago’s *Digital South Asia Library* website,¹ although the page images displayed there are neither searchable nor digitally processable, effectively making this version equivalent to the printed LSI w.r.t. accessing its contents, although of course universally accessible to anybody with an internet connection. One of the major objectives of the study reported in this paper is to convert LSI into a digital resource stored in a way which makes it easy to access, explore, and process for deeper linguistic investigations of the languages described in LSI. This digital resource will have rich formally structured metadata as well as the full original text of the LSI.

1.3 Project Aims

On the linguistic side, the major objective of the project is to investigate the claim about South Asia as a linguistic area. The examination of genealogical, typological and areal relationships among South Asian languages requires a large-scale comparative study, encompassing more than one language family. Further, such a study cannot be conducted manually, but needs to draw on extensive digitized language resources and state-of-the-art computational tools. As mentioned already, there have been some earlier attempts to use LSI in areal studies (e.g., Hook, 1977), but because of the manual nature of these studies, the information in the LSI was used only to a very limited extent, and the results presented in a general, non-concrete manner. Further, no accompanying methodological discussion was offered (e.g., how the data was extracted and analyzed, and for which languages, etc.). We aim to investigate the *South Asia as a linguistic area* claim on the basis of a much broader array of linguistic data using state-of-the-art

¹<http://dsal.uchicago.edu/books/lsi/>

computational techniques and tools in this study. However, in this paper, we focus on the automatic extraction of linguistic features from the LSI data and the development of a typological database which can be used as a major source for the investigation of the above mentioned claim later (Section 3).

The development of general purpose methodologies and tools for large scale comparative linguistics and visualization of linguistic information is another primary aim of the project (Section 4). Our hope and aim is to build methodologies and tools which will be applicable not only to the LSI grammar sketches, but also to the multitude of descriptive grammars of the world's languages that are digitally accessible and available for linguistic investigations (see <http://glottolog.org>).

The full text of the LSI (only the Latin-script portions) has been digitized by a commercial digitization service using double keying, which has resulted in a digital version of very high quality. The amount of text that has been digitized so far is well in excess of one million words.

This will be the first large-scale digital resource on South Asian languages which will be completely automated, with a solid 'deep' structure with the possibility of doing searches for grammatical (morphological and syntactic) as well as lexical features, with links to the original LSI pages as well as rich visualizations. Building a database of this magnitude will also contribute at least indirectly to developing NLP tools for South Asian languages. Studies investigating a multitude of linguistic questions relating to lexicon, morphology, syntax, language contact between two specific languages as well as questions relating to areal linguistics and language change will benefit from this resource. We are already using the resource in our linguistic investigations, as we are building the database (cf. Borin et al., 2014; Saxena, 2016).

We also intend to initiate experiments for utilizing the text specimens for extracting additional linguistic data from the LSI, using the English version of the text as pivot, e.g., inferring basic subject-object marking through cross-language annotation projection (see, e.g., Xia and Lewis, 2007).

2 Recycling Research Infrastructure Components

In the first phase of the project, the linguists in the project team have needed to quickly orient themselves in the vast material of the LSI, both so that they would get an overview of the linguistic features present in the descriptive grammars, and so that they could provide input to the language technologists designing the IE application. In particular, we require gold-standard data on which we can evaluate our IE experiments. This dataset has been prepared using a standard methodological tool in large-scale comparative linguistics, viz. the linguistic questionnaire. In our case, the questionnaires contain mostly yes-no questions – e.g., “Does the language mark dual in at least one personal pronoun?” – and, inevitably, some dependencies among questions, e.g., if the answer to the pronominal dual question is “yes”, there are follow-up questions about first, second and third person pronouns.

The linguists in the project team will be greatly helped by having access to tools allowing them to browse and search the vast LSI material effectively. This is true for those designing the questionnaires, but in particular and to a much higher degree for those charged with filling out the questionnaires – typically linguistics master students – using the LSI grammar sketches.

For effective exploration of the digitized LSI already in the early stages of this project, and also in order not to spend too much project resources on useful but peripheral tool development, we have strived to reuse existing language tools and infrastructure to the greatest extent possible, even if these tools were not designed explicitly for the kind of large-scale comparative linguistic investigations which are being planned in this project, but rather for more traditional corpus-linguistic studies. Thus, the project team decided to recycle some existing e-infrastructure components – several of which were available through the Swe-Clarín infrastructure – rather than attempting to build a new system from scratch. In the following we describe how this was done.

2.1 LSI Grammar Sketches as Corpus

The text data, i.e., grammar sketches excluding tabular data (e.g., inflection tables) and text specimens, have been imported and made searchable using Korp, a versatile open-source corpus infrastructure (Borin

The screenshot shows the Korp web interface. At the top, there are navigation links: Modern | Parallel | Old Swedish | Litteraturbanken | Kubhist | Old texts | More. On the right, there are language options: Log in Svenska English and a settings icon. The main header features the Korp logo and a status bar: 125 of 232 corpora selected — 2.08G of 11.65G tokens. A dropdown menu shows 'tsunami..nn.1'. Below this is a search bar with 'tsunami (noun)' and a 'Search' button. There are checkboxes for 'also as' with options 'initial part', 'final part' and 'and', and 'case-insensitive'. A 'KWIC' section shows 'hits per page: 25', 'sort within corpora: not sorted', and 'Statistics: compile based on: word'. There are checkboxes for 'Show statistics' and 'Show word picture'. The main results area shows 'Results: 10,150' and a pagination bar from 1 to 406. A 'Show context' button is present. The context shows a snippet from 'ÅBO UNDERRÄTTELSE 2012' with the sentence: 'När man drog i snöret så kom det en sådan tsunami att hälften kom på golvet.' Below this, there is a list of corpora: 'ÅLANDSTIDNINGEN 2012', 'Teijo Ristola nämner katastrofer som tsunamin, skolskjutningarna i finska skolor, sjukhusbranden i Åbo, Tjerno', 'Förlagen publicerar nya böcker om Wagner, liksom en tsunami, sa Nike Wagner bland annat, i samband med öppnandet av en V', '8 SIDOR (does not support extended context)', and 'komma en jättevåg från havet efter jordbävningen, en tsunami'. On the right, there is a 'Corpus' section with 'Åbo Underrättelser 2012', 'Text attributes' with 'date: 2012-08-24', and 'Word attributes' with 'final part: [empty]', 'compound lemmas: [empty]', 'part-of-speech: noun', and 'compound word forms: [empty]'.

Figure 2: The user interface to Korp, Språkbanken’s and Swe-Clarin’s corpus infrastructure

et al., 2012b; Hammarstedt et al., 2017a; Hammarstedt et al., 2017b),² developed and maintained by Swe-Clarin leading partner Språkbanken (the Swedish Language Bank at the University of Gothenburg). Korp is used by several CLARIN centers in the Nordic countries, and also, e.g., in Estonia. Currently, the LSI corpus comprises about 1.3 million words, and contains data about around 550 linguistic varieties that we identified during the pre-processing step.

Korp is a modular system with three main components: a (server-side) back-end, a (web-interface) front-end, and a configurable corpus import and export pipeline (Hammarstedt et al., 2017b). The back-end offers a number of search functions and corpus statistics through a REST web service API. As the main corpus search engine, it uses Corpus Workbench (Evert and Hardie, 2011).

The front-end – an in-house development – provides various options to search at simple, extended, and advanced levels in addition to providing a comparison facility between different search results (Hammarstedt et al., 2017a). See Figure 2.

The corpus pipeline is a major component and can be used to import, annotate, and export the corpus to other formats. For annotations, it relies heavily on pre-existing external annotation tools such as segmenters, POS taggers, and parsers. Previously, it has mostly been used for Swedish text, and comes with very limited support for English in the vanilla distribution. For our purposes, we have incorporated the English Stanford Parser (Manning et al., 2014) for lexical and syntactical annotations. We have added word and text level annotations to the LSI data. The following is a list of all annotations that were added:

Word-level annotations: lemma, part of speech (POS), named-entity information, normalized word-form, dependency relation. These are all added automatically.

Text-level annotations: LSI volume/part number, language family, language name, ISO 639-3 language code, longitude, latitude, LSI classification, Ethnologue classification (Simons and Fennig, 2018), Glottolog classification,³ page number, page source URL, paragraph and sentence level segmentation. These have been added in a semi-automatic manner.

While most of the annotations are self-explanatory, there are a few which may need some explanation. The *normalized word form* is the form produced by removing the diacritics and other phonological char-

²<http://spraakbanken.gu.se/swe/forskning/infrastruktur/korp/distribution>
<https://github.com/spraakbanken/korp-frontend/>

³<http://glottolog.org>

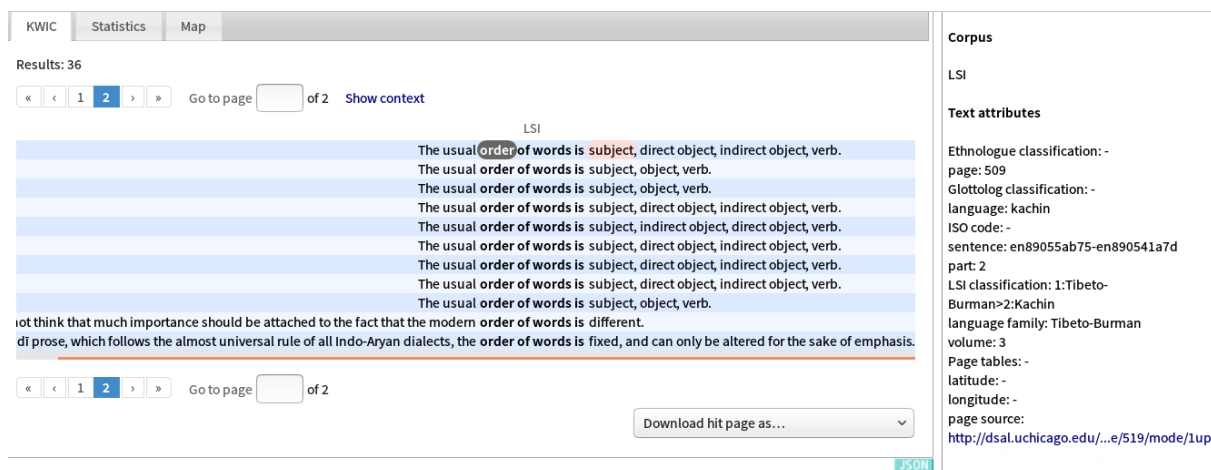


Figure 3: Korp KWIC view resulting from searching the LSI for the string “order of words is”

acters. The purpose is to make it easy to search the corpus by using the standard English keyboard without requiring the user to enter accented characters, since the LSI consistently renders language names and glosses in a kind of phonetic transcription which will most likely be unfamiliar to many users. Thus, the normalization allows the user to search for, e.g., *Bihārī* using the search string “Bihari”, or “bihari” (with case-sensitivity disabled).

The text-level annotations above mostly represent the metadata which were collected from different sources⁴ in addition to the LSI volumes themselves, and are maintained as part of the corpus. The page source URL, for example, is a link to the image version of the corresponding LSI page available from the University of Chicago’s *Digital South Asia Library*.

Figure 3 shows a screenshot of the Korp front-end displaying results of a simple corpus query in Korp’s KWIC (Key Word In Context) view. The query is one aiming at finding out about the basic word order of a language, which is one of the more prominent linguistic features utilized in typological language classification. This illustrates how this basic corpus tool can be repurposed for pursuing the kind of research questions that our linguist project members are interested in. It is true that Korp is not the ideal tool for this, but building a bespoke application would have been far beyond the means of the project.

The box to the right of the KWIC sentences shows annotations and metadata for the selected word (*Word* and *Text* level attributes), and also provides a link to the corresponding page image available at the *Digital South Asia Library* at the University of Chicago.⁵

The Korp software could be directly used in the project, without any other modification than setting up its configuration files for handling the LSI texts and using English language tools.

2.2 LSI Tables and Specimens as Lexicons

The LSI grammar sketches contain large amounts of tabular material, e.g., inflection tables, personal pronoun systems, etc., and also language specimens in the form of *interlinear glossed text*, both of which are not particularly suitable for displaying in a corpus KWIC view. Instead, these are imported and stored in another of Språkbanken’s infrastructure components, Karp (Borin et al., 2012a). Links are provided from the Korp KWIC metadata box to tables and specimens in Karp, but these can also be accessed directly through the Karp search interface.

Like Korp and other infrastructure components developed and maintained by Språkbanken, Karp, too, is structured with a server backend, a JSON-based web service API and a web application frontend. This means that functionality can be conveniently modularized. For instance, the Korp frontend calls

⁴For instance, location data come mainly from the Glottolog: <http://glottolog.org>.

⁵<http://dsal.uchicago.edu/books/lsi/> – page images only; no text search facility is available.

the Karp backend for the lexical information needed in order to execute lemma-based corpus searches, and conversely, the Karp frontend calls the Korp backend in order to offer example sentences for lexical entries.

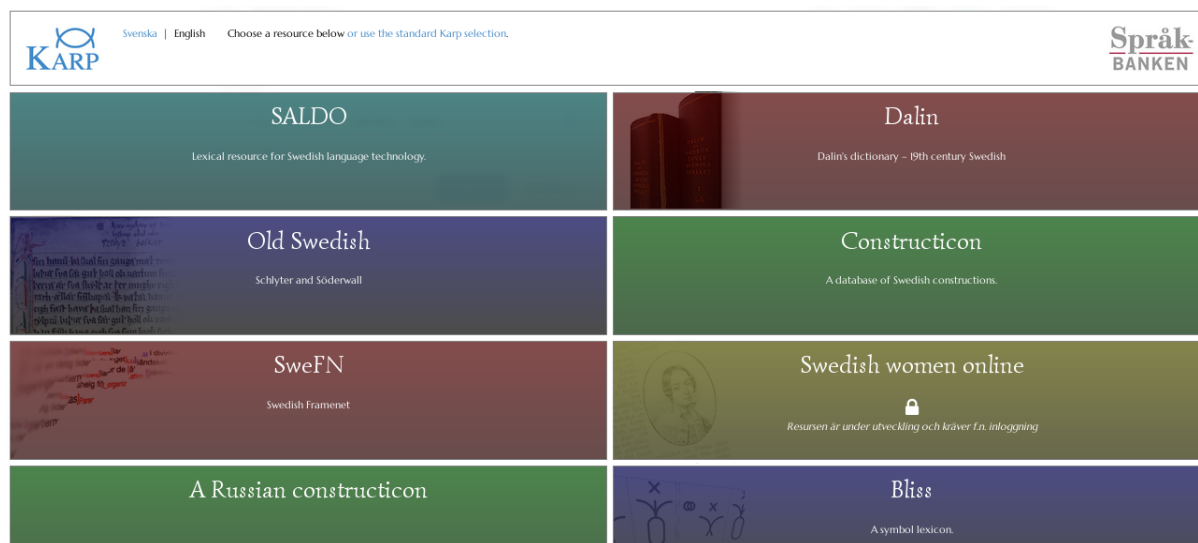


Figure 4: Karp, Språkbanken's and Swe-Clarín's infrastructure for accessing and editing lexical resources and other formally structured language data

Karp started out as a fairly run-of-the-mill web-based search and browsing environment for lexical data. All of Språkbanken's digital lexical resources are available through it, whether born digital – as SALDO, the main lexical resource used for Swedish morphological annotation in the import pipeline of Korp (Borin et al., 2013a) – or digitized versions of traditional dictionaries, including a number of Swedish historical dictionaries (Borin and Forsberg, 2011).

From its humble beginnings, Karp has developed into an infrastructure component for working with language data that has a formally defined (tabular) structure. In addition to lexical entries, this includes also data such as grammatical paradigm tables and encyclopedia articles.

Karp has an editing mode (Borin et al., 2013b), which was originally developed for building the Swedish FrameNet (Borin et al., 2010), but which has since been extended into a general editing environment for formally structured language data. Notably, the Swedish and Russian constructicons are built using the Karp editor (Lyngfelt et al., 2012; Janda et al., to appear 2018), as is the *Swedish Women Online* biographical database.⁶ See Figure 4.

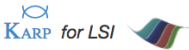
In the context of the present project, Språkbanken's Karp development team had to be involved in devising an "LSI mode", but in fact this fitted well with our ongoing effort aimed at turning Karp into a more general infrastructure for working with formally structured linguistic data, as described above. See Figure 5, illustrating a query aiming at finding out some linguistic features of the personal pronominal systems of the LSI languages. Again, as in the case of Korp mentioned above, this solution is not perfect, but it could have it up and running in a very short time compared to what it had meant to implement the perfect functionality from scratch, meaning that limited project resources can be put to better use, such as deeper linguistic analysis of the LSI data.

3 Automatic Extraction of Linguistic Information from Linguistic Descriptions

After having pre-processed the LSI data and stored it in a structured way, the next step is to extract information about particular grammatical features, and to build a typological database of the LSI languages. The developed feature database is to be used for investigation of the claim about South Asia as a linguistic area during the later stages of the project.

Automatic extraction of linguistic features from traditional linguistic descriptions is a novel task, and has high potential value in typological, genealogical, historical, and other related areas of linguistics

⁶<https://skbl.se/>


My lexicons
Log out

Sök i LSI
Freetext Search
Search History

Reset

Find entries where anything equals + or...

+ and...
+ except...

Search
Compile on...

Hits **12**

Page: 1 / 1

LSI ▾ 12 HITS (DISPLAYING 12)

Newari

nom

	FIRST	SECOND	THIRD
SG	ji, i. .	chha, chhi, thou. .	a-mi-sā, a-mi-se~, by them. .
PL	jhi-ji, jhi-pī, we. .	chhi-pī-gu, your. .	a-pī, they. .

obl

	FIRST	SECOND	THIRD
SG	ji-na, ji, by me. .	chha-nā, by thee. .	ō, by him. .
PL	jhi-ji-sena, ji-mi-se~, by us. .	chhi-mi-sā, chhim-se~, by you. .	a-mi-sā, a-mi-se~, by them. .

Figure 5: Karp view showing LSI tables of personal pronoun paradigms

that make use of databases of structural features of languages. There exist many typological databases of linguistic structures, including the *World Atlas of Language Structures* (WALS) (wals.info), the *Atlas of Pidgin and Creole Language Structures* (APiCS) (apics.org), the *South American Indigenous Language Structures* (SAILS) (sails.clld.org), AUTOTYP (github.com/autotyp/autotyp-data), and the *Phonetics Information Base and Lexicon* (PHOIBLE) (phoible.org). To the best of our knowledge, all the linguistic databases published so far have been manually constructed and curated, where human experts have turned information from field data or analyzed data into data-points in the database. The use of human expertise guarantees a certain level of quality and robustness, but is highly labor intensive and consequently costly. There are some 6,500 languages in the world, out of which descriptive grammars – ranging from brief grammar sketches to multi-volume reference grammars – are available for over 4,000 (see glottolog.org). Manually extracting information about 200–300 features from each of them is a very ambitious – and in practice unrealistic – undertaking.⁷ Significant amounts of analyzed language data (grammatical descriptions in discursive textual form) are increasingly

⁷Very relevant in this connection is the fact that one of the most ambitious and well-known linguistic-feature datasets, the WALS (Dryer and Haspelmath, 2013), even though it reports values for a total of 192 linguistic features in 2,679 languages, in reality most cells in the resulting matrix are empty. In version 2014 of the dataset available for download from <http://wals.info/download>, out of a total of 514,368 cells, no less than 437,903 are empty, meaning that less than 15% of the potential values have actually been filled.

being made available in digital form, and the field of natural language processing (NLP) offers tools that potentially can aid us in extracting information about linguistic features from such textual sources, at least for sources in English and some other languages. To take advantage of these advancements and to help the linguistic community in populating the linguistic feature databases, we have developed methods to automatically extract linguistic features from linguistic grammars.

For our initial study, we have identified a list of features that we think are interesting and will be useful to meet the objectives of the project. Some of these features are:

- (1) Apos: What is the order of adnominal property word and noun?⁸
- (2) NLpos: What is the order of numeral and noun in the NP?
- (3) NLBase: What is the base of the numeral system?
- (4) Aagr: Can an adnominal property word agree with the noun in number and/or gender?
- (5) AagrNum: Can an adnominal property word agree with the noun in number?
- (6) AagrGen: Can an adnominal property word agree with the noun in gender?
- (7) Reflexive: What kind of reflexive construction does the language have?
- (8) DefArticle: Are there definite or specific articles?
- (9) WOrder: What is the order of words?

For the purpose of extracting values and/or descriptions of these features from traditional reference grammars, including the grammar sketches in the LSI, we have experimented with three approaches to information extraction: (1) Pattern based; (2) dependency parsing based; and (3) semantic parsing based. In the following sections, we briefly describe each of the approaches and their results while leaving the details to be reported separately.

3.1 Pattern Based Feature Extraction

The pattern based feature extraction methodology is inspired by pattern based information extraction in general, and predicated on the observation that information about particular linguistic features in descriptive grammars is often given using particular descriptive patterns (at least we have observed this in the case of LSI). Taking advantage of this, one can look for the existence of particular keywords (or a combination of words) within the descriptive grammar to reach to the relevant text, and then process it further to extract the feature value of interest. We used a similar type of two-stage approach to retrieve the relevant sentences from our data using Korp’s standard search API first, and then to process them further using regular expression based patterns to extract the feature values. Suppose for example that we are interested in extracting information about the normal word order in a particular LSI language from the language description. As a first step, we can extract all sentences having the string “order of words is” from the description of a language (see Figure 3). Next, using the pattern `(.*) (order of words is) (.*)`, one can first split each sentence into three parts: the part appearing before the string “order of words is”, the string itself, and the part appearing after this string. The resulting parts can be processed further with more specific patterns (e.g. `(\w+)`, `(\w+)`, `(\w+)`) to extract the ‘order of words’ of that particular language.

This simple approach allowed us to get off the ground quickly, but it has serious limitations. This pattern based strategy will very strictly match particular sentence structures and/or contents. This probably will not cover all possible ways the same information could have been encoded unless one designs patterns rich enough to catch all possible instances. For such reasons, we have experimented with approaches inspired by syntactic and semantic analysis, and *Open Information Extraction* based techniques (e.g., Fader et al., 2011).

⁸An *adnominal property word* corresponds to an adjective or participle in English and many other languages.

3.2 Dependency Parsing Based Feature Extraction

Dependency parsing provides syntactic dependency information for the words of a text, which we exploit to extract feature values in this feature extraction strategy. After retrieving the relevant sentences using Korp’s search facility (as exemplified above), the sentences were parsed using the Stanford dependency parser (Manning et al., 2014), and the resulting dependencies were further processed using a set of rules to extract the required feature values. Again as an example, suppose that we are interested in extracting information about the order of adjective and noun in the Siyin⁹ language. Using Korp’s standard search interface, we can extract all sentences containing the lemma “noun” or “adjective” from the language description. One of the extracted sentences will be:

The adjectives follow the noun they qualify .

When we parse this sentence with the Stanford dependency parser, it will return the following dependencies:

```
det(adjectives-2, The-1)
nsubj(follow-3, adjectives-2)
root(ROOT-0, follow-3)
det(noun-5, the-4)
dobj(follow-3, noun-5)
nsubj(qualify-7, they-6)
acl:relcl(noun-5, qualify-7)
```

These dependencies can be processed further with a set of rules to extract the required information. We have worked out a specific set of rules (in the form of an algorithm) for each feature value that we are interested in. The details are beyond the scope of this paper, and will be reported elsewhere.

Table 1 shows how accurately the proposed feature extraction methodology was able to extract different feature values. For each feature, the accuracy value was computed using the following simple formula:

$$Accuracy = \frac{N_{correct}}{N_{extracted}}$$

Where $N_{correct}$ is the number of languages for which the feature value was correctly extracted, and $N_{extracted}$ is the total number of languages for which the feature value was extracted. To decide if an extracted value is correct or not, it was compared to the gold value which was retrieved manually by a human expert from the comparative vocabulary or from the language descriptions.

Feature	Accuracy (%)
Apos	0.818
NLPpos	1.0
NLBase	0.823
Reflexive	0.739
Aagr	0.857

Table 1: Evaluation results: Dependency parsing

Once again, this strategy will very strictly match particular sentence structures and contents of arguments. To address some of the limitations of this strategy, we report another strategy in the next subsection which is based on semantic parsing.

3.3 Semantic Parsing Based Feature Extraction

Shallow semantic analysis or semantic role labelling (SRL) is the process of identifying and labeling the semantic roles (also known as semantic arguments) associated with verbal or nominal predicates in a

⁹Siyin (csy) is a Tibeto-Burman language spoken in Burma.

Predicate	Semantic arguments
follow	ARG1:The_adjectives, ARG2:the_noun_they_qualify
qualify	ARG1: the_noun_they

Table 2: Semantic parse

given piece of text. Automatic semantic role labeling finds applications in many areas of NLP including information extraction (Surdeanu et al., 2003), and in this work we are using it for feature extraction – a sort of information extraction. In this strategy, after having parsed the sentences using a semantic parser (Björkelund et al., 2009), the parses are further processed to extract feature values. The further processing steps involve (1) checking for particular predicates for particular features; (2) inspecting the semantic arguments’ structure and contents; and (3) formulating the feature values. Using our previous example, i.e., the ordering of adjective and noun in the Siyin language, this time we can semantically parse the sentence *The adjectives follow the noun they qualify* to get the verbal predicates and their semantic arguments as given in Table 2.

The predicate ‘follow’ is one of those predicates that we had identified, independently, to be linked to the adjective–noun order feature: Using a development data set, we identified a set of predicates linked to each of the target features. This simply involved finding sentences in the descriptive grammars which were used to provide information about a particular feature, and then analyzing them to find the associated list of predicates.

The next step is to examine the semantic arguments of the predicate ‘follow’, and formulate the feature value. According to Propbank,¹⁰ for the predicate ‘follow’, ARG1 represents the thing following, while ARG2 represents the thing followed. In the analysis shown in Table 2, the string *The adjectives*, is ARG1 (i.e. the thing following), while the string *the nouns they qualify* is ARG2 (i.e. the thing followed). The substrings representing ARG1, and ARG2 can be further analyzed to formulate and return the feature value ‘2-N-ANM’ (the fact that adjectives follow the nouns). Had ARG1 contained *noun(s)*, ARG2 contained *adjective(s)* with predicate being ‘follow’, or ARG1 contained *adjective(s)*, ARG2 contained *noun(s)*, and the predicate being ‘precede’, ‘1-ANM-N’ (the fact that adjectives precede nouns) would have been returned as the feature value. We have used simple if-then-else conditions to examine predicates and their semantic argument strings for the purpose of extracting and formulating the feature values. In the future, we plan to experiment with more advanced techniques, such as active learning, for the feature extraction and formulation from the semantic parses.

In order to test the generality of our approach, for this experiment, rather than drawing on the LSI grammar sketches, we have worked with digitized reference grammars used in the Grambank project,¹¹ for a set of languages where the linguistic features of interest have already been extracted manually from exactly the grammars used in our experiment, thus providing us with a gold standard dataset.

The evaluation results of this strategy are given in Table 3. As can be seen, the system has varying precision and recall for different features, which highlights the difficulty/ease of automatically extracting the corresponding feature values using the described method.

As mentioned previously, there does not exist any work related to automatic linguistic feature extraction, which means we do not have any other system to compare the proposed system’s performance. Instead, we evaluate the system performance against a baseline calculated for each feature on the basis of the most frequent feature value. As can be noted, for four out of the five features, the proposed system was able to easily beat the baseline precision values, the exception being the feature ‘AagrNum’.

¹⁰The lexico-semantic resource on which the semantic parser is based

¹¹Grambank is an ongoing initiative at the Max Planck Institute for the Science of Human History at Jena, developing a database of structural (typological) features for a substantial part of the world’s languages. The grammars and gold-standard feature sets used in this experiment were kindly put at our disposal by Harald Hammarström.

Feature	Precision	Recall	F-Score	Baseline Precision
Apos	0.76	0.40	0.52	0.41
NLpos	0.85	0.30	0.44	0.75
AagrNum	0.69	0.21	0.32	0.77
AagrGen	0.64	0.14	0.23	0.27
DefArticle	0.84	0.27	0.41	0.27

Table 3: Evaluation results: Semantic parsing

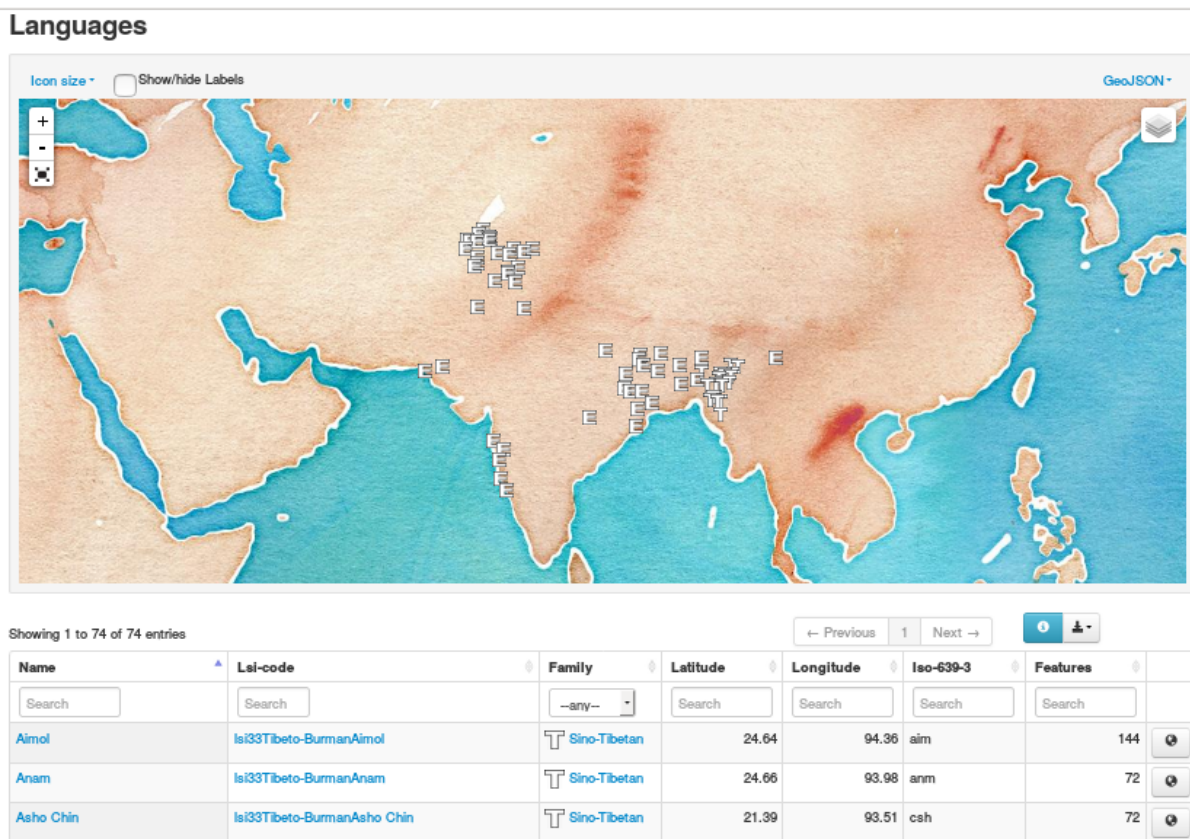


Figure 6: The LSI dataset in CLLD

4 Visualization for Linguistic Research: Visual Exploration of the LSI

An important aspect of the linguistic research driving our project is the relationship between *linguistic genealogy* (language family membership), *geography*, and *linguistic features*. Again, the digitized LSI offers such an abundance of data of various kinds, that we need very good tools for exploring this resource for the kind of large-scale comparative linguistic research necessitated by our project objectives. There are indications that data visualization and visual analytics have a crucial role to play in this connection (e.g., Havre et al., 2000; Chuang et al., 2012; Krstajić et al., 2012; Sun et al., 2013). So, we are developing a number of solutions for better visualization of languages and their features on maps to help the linguistic community working in the areas mentioned above, and to achieve the goals of the LSI project.

For the general case, we have adopted the *Cross-Linguistic Linked Data* (CLLD) framework developed by the Max Planck Society,¹² which is open-source and which we could simply install out of the box and configure to display all LSI varieties to which we could assign an ISO 639-3 language code. See Figure 6.

For the more specific purposes of working with the full LSI data, we have modified the mapping solution available in Korp into an interactive standalone application where the users can view the distribution

¹²<http://clld.org/>

of linguistic features in LSI varieties on a map. We provide switchable shape/color combinations for visualizing and differentiating family/feature characteristics. Figure 7 shows a snapshot visualizing the feature **s3sg** (“Is the form of the pronominal 3sg subject the same in intransitive and transitive clauses?”, i.e., an indicator of nominative–accusative vs. absolutive–ergative alignment) in languages belonging to the Indo-Aryan and Tibeto-Burman families. The user can select multiple families and multiple features at the same time by checking the appropriate check-boxes, and can also switch between color/symbol to visualize feature/family by selecting the appropriate radio button. In the map in Figure 7 we have selected feature values to be encoded by color, while the shape of the markers indicate language family (**I** for Indo-Aryan and **T** for Tibeto-Burman in Figure 7). In this map we can discern a clear areal distribution of this feature in South Asia, such that accusative alignment is mainly found in the east, regardless of language family. Such an interactive mapping facility provides a useful way to show the genetic relations and areal influences between languages spoken in different geographical areas and belonging to different language families.

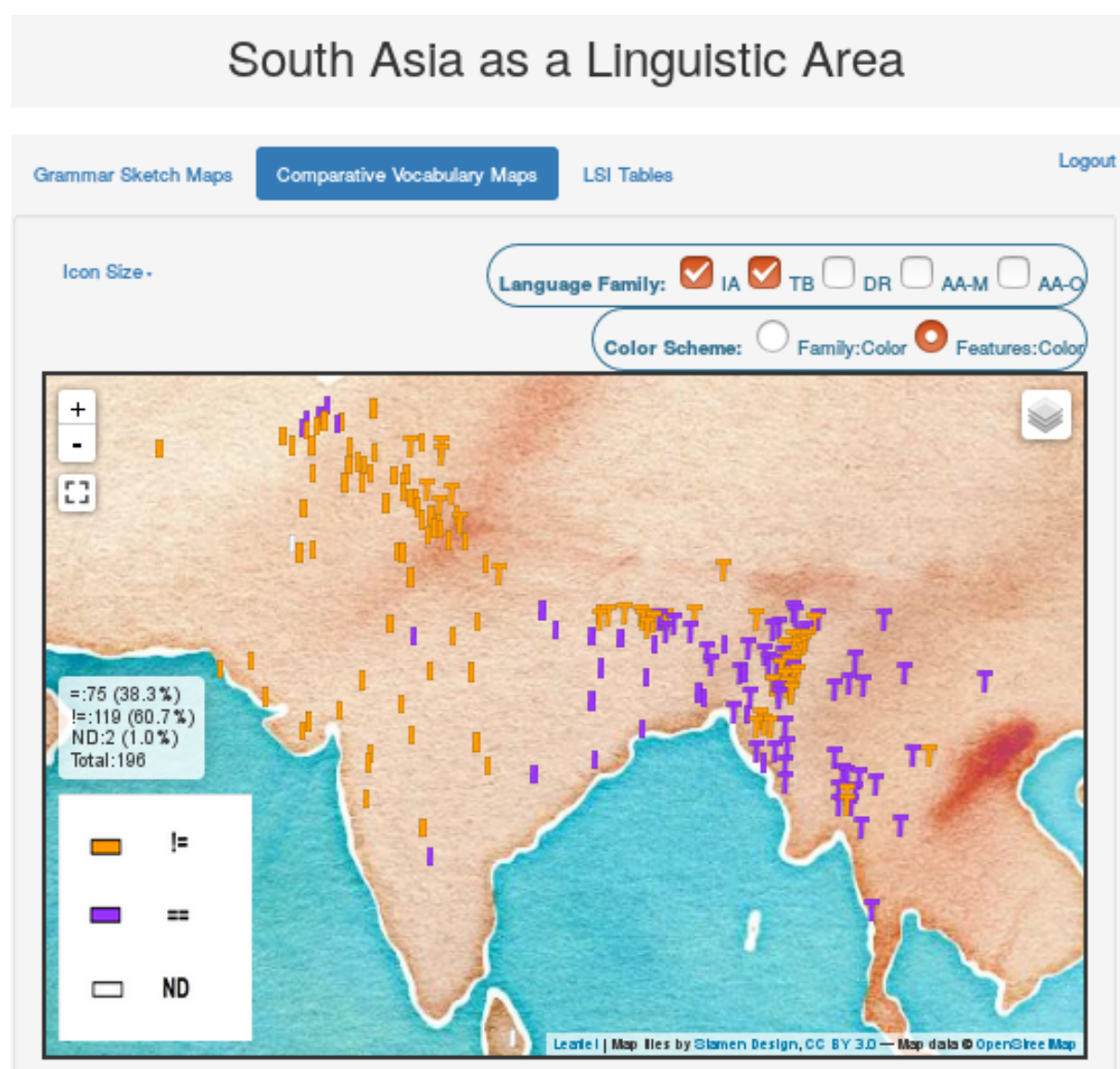


Figure 7: Map showing form of subject pronoun in relation to transitivity

This type of visualization is very helpful for comparison purposes, but not equally useful if we are interested to only explore/visualize feature values of individual languages. For that purpose, we have developed simple feature visualization solutions. Figure 8 shows a screenshot displaying feature values

extracted from the description of Lohorong.¹³ This type of expandable/compressible tree styled visualization makes it easy to visualize feature values of a language of interest.

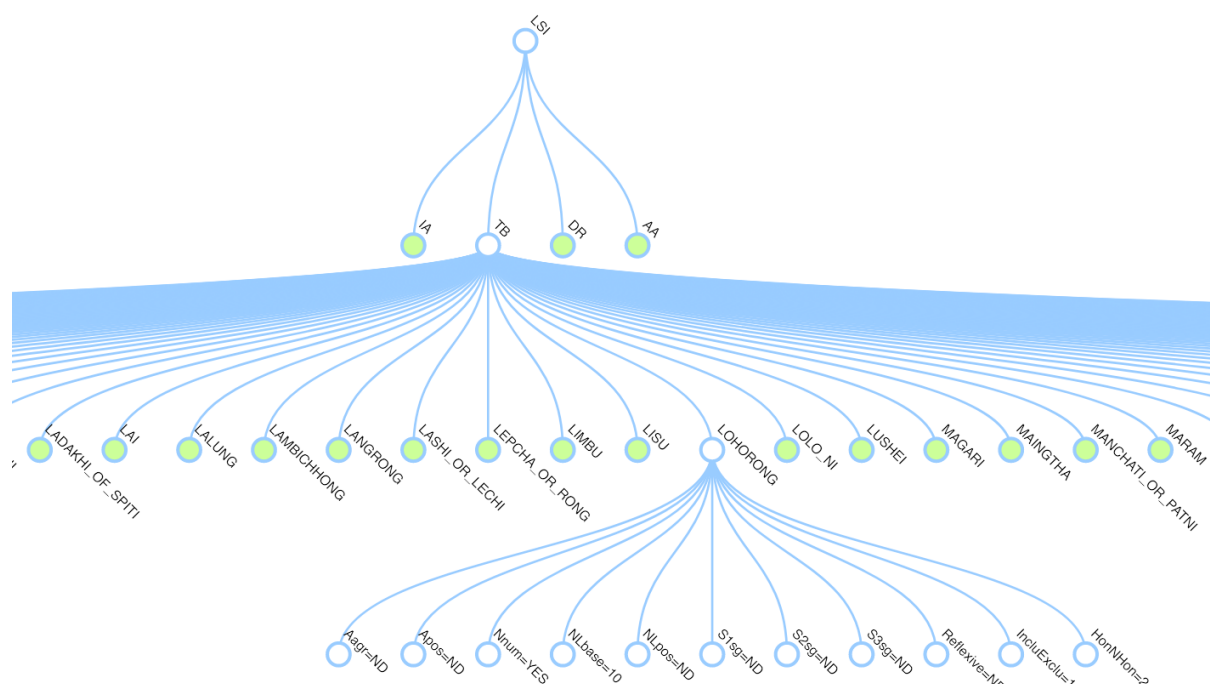


Figure 8: Linguistic features of Lohorong (lrr)

5 Conclusions and future work

Turning the LSI into a structured digital resource will provide a rich empirical foundation for large-scale comparative studies of the linguistic ecology of South Asia. In this day and age, it makes little sense to conduct such studies manually. Instead, they need to draw on extensive digitized language resources and state-of-the-art computational tools. This is the main goal of our ongoing work with the LSI.

In addition to this, we aim to contribute to the methodological development of large-scale comparative linguistics drawing on digital language resources, as well as to the methodological development of SRL based and open information extraction, adapting these paradigms to a different and hitherto unexplored domain. In the longer perspective, we hope that the solutions which we develop in our work will be more generally applicable to the text mining of descriptive grammars – which are increasingly available in digital form – so that the resulting formally structured linguistic information can be used to populate linguistic databases. Indeed, the outcome of our experiments on the Grambank data (described in Section 3.3) indicates that there are some grounds for optimism in this regard.

In order to get the project off the ground quickly, we needed tools for browsing, searching and visualizing the abundance of information present in the LSI. Recycling existing infrastructure components has turned out to be surprisingly effective. We have been able to use Korp and the CLLD framework more or less off the shelf. Rendering the LSI tabular data in Karp required modifications to the Karp infrastructure, and the geographical mapping solution shown in Figure 7 in practice is a new component developed in this project.

The linguists working with the questionnaires have expressed their satisfaction with Korp as an “information retrieval” interface to the LSI text. An added value in this context is that they have been asked

¹³Lohorong (lrr) is a Tibeto-Burman language spoken in Nepal.

to save the search results – sentences in the text – found by them to be the most relevant to determining a particular linguistic feature, thus providing invaluable input to our work on designing an IE system targeting linguistic information expressed in conventional descriptive grammars.

The status of the project is that most of the LSI has been digitized, the browsing, search and visualization applications described above have been implemented,¹⁴ and the manual questionnaire work and the development of the IE application is underway.

In the future, we would also like to take into account the phonological and other related information present in tabular data and the parallel annotated data present in the text specimens provided with LSI grammar sketches.

Acknowledgments

The work presented here was funded by the Swedish Research Council as part of the project *South Asia as a linguistic area? Exploring big-data methods in areal and genetic linguistics* (2015–2019, contract no. 421-2014-969), as well as by the University of Gothenburg and the Swedish Research Council through their funding of the Språkbanken and Swe-Clarin research infrastructures, respectively.

References

- [Björkelund et al.2009] Anders Björkelund, Love Hafdell, and Pierre Nugues. 2009. Multilingual semantic role labeling. In *Proceedings of CoNLL 2009: Shared Task*, pages 43–48, Boulder, Colorado. ACL.
- [Borin and Forsberg2011] Lars Borin and Markus Forsberg. 2011. A diachronic computational lexical resource for 800 years of Swedish. In Caroline Sporleder, Antal van den Bosch, and Kalliopi Zervanou, editors, *Language Technology for Cultural Heritage*, pages 41–61. Springer, Berlin.
- [Borin et al.2010] Lars Borin, Dana Dannélls, Markus Forsberg, Dimitrios Kokkinakis, and Maria Toporowska Gronostaj. 2010. The past meets the present in Swedish FrameNet++. In *14th EURALEX International Congress*, pages 269–281, Leeuwarden. EURALEX.
- [Borin et al.2012a] Lars Borin, Markus Forsberg, Leif-Jöran Olsson, and Jonatan Uppström. 2012a. The open lexical infrastructure of Språkbanken. In *Proceedings of LREC 2012*, pages 3598–3602, Istanbul. ELRA.
- [Borin et al.2012b] Lars Borin, Markus Forsberg, and Johan Roxendal. 2012b. Korp – the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012*, pages 474–478, Istanbul. ELRA.
- [Borin et al.2013a] Lars Borin, Markus Forsberg, and Lennart Lönnngren. 2013a. SALDO: A touch of yin to WordNet’s yang. *Language Resources and Evaluation*, 47(4):1191–1211.
- [Borin et al.2013b] Lars Borin, Markus Forsberg, Leif-Jöran Olsson, Olof Olsson, and Jonatan Uppström. 2013b. The lexical editing system of Karp. In *Proceedings of the eLex 2013 Conference*, pages 503–516, Tallin.
- [Borin et al.2014] Lars Borin, Anju Saxena, Taraka Rama, and Bernard Comrie. 2014. Linguistic landscaping of South Asia using digital language resources: Genetic vs. areal linguistics. In *Proceedings of LREC 2014*, pages 3137–3144, Reykjavik. ELRA.
- [Chuang et al.2012] Jason Chuang, Daniel Ramage, Christopher D. Manning, and Jeffrey Heer. 2012. Interpretation and trust: Designing model-driven visualizations for text analysis. In *ACM Human Factors in Computing Systems (CHI)*.
- [Dryer and Haspelmath2013] Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.
- [Ebert2006] Karen Ebert. 2006. South Asia as a linguistic area. In Keith Brown, editor, *Encyclopedia of Languages and Linguistics*. Elsevier, Oxford, 2nd edition.
- [Evert and Hardie2011] Stefan Evert and Andrew Hardie, 2011. *Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium*. University of Birmingham.
- [Fader et al.2011] Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of EMNLP 2011*, pages 1535–1545, Edinburgh. ACL.

¹⁴The LSI goes out of copyright towards the end of the project and our data will subsequently be made openly available.

- [Georg2017] Stefan Georg. 2017. Other isolated languages of Asia. In Lyle Campbell, editor, *Language Isolates*, pages 139–161. Routledge, London.
- [Grierson1903–1927] George A. Grierson. 1903–1927. *A Linguistic Survey of India*, volume I–XI. Government of India, Central Publication Branch, Calcutta.
- [Hammarstedt et al.2017a] Martin Hammarstedt, Lars Borin, Markus Forsberg, Johan Roxendal, Anne Schumacher, and Maria Öhrman. 2017a. Korp 6 – Användarmanual [Korp 6 – User manual]. Research reports from the Department of Swedish GU-ISS 2017-02, University of Gothenburg, Gothenburg. <http://hdl.handle.net/2077/53096>.
- [Hammarstedt et al.2017b] Martin Hammarstedt, Johan Roxendal, Maria Öhrman, Lars Borin, Markus Forsberg, and Anne Schumacher. 2017b. Korp 6 – Technical report. Research reports from the Department of Swedish GU-ISS 2017-01, University of Gothenburg, Gothenburg. <http://hdl.handle.net/2077/53095>.
- [Havre et al.2000] Susan Havre, Beth Hetzler, and Lucy Nowell. 2000. ThemeRiver: Visualizing theme changes over time. In *IEEE Symposium on Information Visualization, 2000. InfoVis 2000*, pages 115–123, Salt Lake City. IEEE.
- [Hook1977] Peter E. Hook. 1977. The distribution of the compound verb in the languages of North India and the question of its origin. *International Journal of Dravidian Linguistics*, 6:336–351.
- [Janda et al.to appear 2018] Laura A. Janda, Olga Lyashevskaya, Tore Nessel, Ekaterina Rakhilina, and Francis M. Tyers. to appear 2018. A construction for Russian: Filling in the gaps. In Benjamin Lyngfelt, Lars Borin, Tiago Timponi Torrent, and Kyoko Hirose Ohara, editors, *Constructicons in Contrast. Constructicography as a Fusion Between Construction Grammar and Lexicography*. John Benjamins, Amsterdam.
- [Krstajić et al.2012] Miloš Krstajić, Mohammad Najm-Araghi, Florian Mansmann, and Daniel A. Keim. 2012. Incremental visual text analytics of news story development. In *Proceedings of VDA 2012*, Burlingame, California. SPIE.
- [Lyngfelt et al.2012] Benjamin Lyngfelt, Lars Borin, Markus Forsberg, Julia Prentice, Rudolf Rydstedt, Emma Sköldberg, and Sofia Tingsell. 2012. Adding a construction to the Swedish resource network of Språkbanken. In *Proceedings of KONVENS 2012 (LexSem 2012 Workshop)*, pages 452–461, Vienna. ÖGAI.
- [Manning et al.2014] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of ACL 2014*, pages 55–60.
- [Masica1976] Colin P. Masica. 1976. *Defining a Linguistic Area: South Asia*. Chicago University Press, Chicago.
- [Nichols2003] Johanna Nichols. 2003. Diversity and stability in language. In Brian D. Joseph and Richard D. Janda, editors, *The Handbook of Historical Linguistics*, pages 283–310. Blackwell, Oxford.
- [Saxena2016] Anju Saxena. 2016. Indo-Aryan in typological and areal perspective. Keynote presentation at SALA-32, Lisbon, 27–29 April, 2016.
- [Simons and Fennig2018] Gary F. Simons and Charles D. Fennig, editors. 2018. *Ethnologue: Languages of the World*. SIL International, Dallas, 21st edition. Online version: <http://www.ethnologue.com>.
- [Sun et al.2013] Guo-Dao Sun, Ying-Cai Wu, Rong-Hua Liang, and Shi-Xia Liu. 2013. A survey of visual analytics techniques and applications: State-of-the-art research and future challenges. *Journal of Computer Science and Technology*, 28(5):852–867.
- [Surdeanu et al.2003] Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. 2003. Using predicate-argument structures for information extraction. In *Proceedings of ACL 2003*, pages 8–15, Sapporo. ACL.
- [Swadesh1955] Morris Swadesh. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, 21(2):121–137.
- [Ward and Barker2013] Jonathan Stuart Ward and Adam Barker. 2013. Undefined by data: A survey of big data definitions. *CoRR*, abs/1309.5821.
- [Xia and Lewis2007] Fei Xia and William Lewis. 2007. Multilingual structural projection across interlinear text. In *Proceedings of HLT 2007*, pages 452–459, Rochester, New York. ACL.

Implementation of an Open Science Policy in the context of management of CLARIN language resources: a need for changes?

Aleksei Kelli
Kaarli 3, 10119 Tallinn
Estonia
aleksei.kelli@ut.ee

Krister Lindén
Unioninkatu 40
00014 Helsingin yliopisto
Finland
krister.linden@helsinki.fi

Kadri Vider
J. Liivi 2
50409 Tartu
Estonia
kadri.vider@ut.ee

Penny Labropoulou
R.C. Athena/ILSP
Epidavrou & Artemidos
151 25 Maroussi
Greece
penny@ilsp.gr

Erik Ketzan
Birkbeck,
University of London
43 Gordon Square
WC1H 0PD London
United Kingdom
Eketza01@mail.bbk.ac.uk

Pawel Kamocki
ELDA, 9 rue des Cordelières
75013 Paris, France /
IDS Mannheim, R5, 6-13
68161 Mannheim
Germany
pawel.kamocki@gmail.com

Pavel Straňák
Charles University, Faculty of
Mathematics and Physics
Institute of Formal and
Applied Linguistics
Malostranské nám. 25
118 00 Praha 1, Czechia
stranak@ufal.mff.cuni.cz

Abstract

This article investigates the compatibility of the current CLARIN license categorization scheme with the open science paradigm. The first part presents the main concepts and theoretical framework required for the analysis, while the second part discusses the use of the CLARIN categorization system, divided into PUB (public), ACA (academic), and RES (restricted), and potential ways to change it. This paper serves to explore various suggestions for change and to begin discussion of a reformed CLARIN license category scheme.

1 Introduction*

The aim of this paper is to explore how, and if, the existing CLARIN license categories (PUB, public; ACA, academic; and RES, restricted) should be kept, modified, or replaced to further the goals of CLARIN's open science policy. This paper will be used as a starting point to evaluate and analyse the

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

compatibility of open science requirements with the way CLARIN manages language resources¹. In addition, this paper could support the development of the CLARIN open science policy itself.

In the first part of the article, the authors explore the theoretical framework and basic concepts necessary for the analysis. In the second part, the use of the CLARIN scheme and alternative categorization schemes are addressed.

This article has practical value because the authors are CLARIN Legal Issues Committee members with divergent views on how to move the implementation of open science policy forward within CLARIN. The integration of these different views gives more legitimacy to a possible outcome for CLARIN.

2 Open science definitions

CLARIN has expressed its commitment to **open science** (see CLARIN Value Proposition 2016). This commitment, however, requires additional clarification. To evaluate whether CLARIN follows open science requirements while managing language resources, it is necessary to define open science (OS).

According to the European Commission (2016) “Open Science represents a new approach to the scientific process based on cooperative work and new ways of diffusing knowledge by using digital technologies and new collaborative tools”. OECD (2015) defines open science as “efforts by researchers, governments, research funding agencies or the scientific community itself to make the primary outputs of publicly funded research results – publications and the research data – publicly accessible in digital format with no or minimal restriction as a means for accelerating research”.

There are several initiatives which provide criteria on the concept of “open”. The Berlin Declaration on **Open Access** (2003) requires that: 1) open access should cover research results, raw data and metadata, source materials, digital pictorial and graphical materials, etc.; 2) rights holders grant to all users a license to use, distribute, and to make and distribute derivative works; 3) a complete version of the work and all supplemental materials in an appropriate standard electronic format is deposited.

The policy document entitled “Ten years on from the Budapest Open Access Initiative: setting the default to open” (BOAI 2012) has some specific requirements for licensing and reuse (e.g. a recommendation to use CC-BY² or an equivalent license).

The Open Knowledge International³ sets the following key features of **openness**: 1) availability and access; 2) reuse and redistribution; 3) universal participation.

Open Knowledge International has also adopted the **Open Definition** (Open Definition 2.1), which has detailed conditions for the determination of open works and open licenses. It essentially allows conditions such as attribution, integrity, and share-alike. If there are additional restrictions on re-use of the data (e.g. non-commercial use, no derivatives), then the content is not open.

The director of the OpenScience project (which is dedicated to writing and releasing free and open source scientific software) defines open science through four fundamental goals: 1) transparency in experimental methodology, observation, and collection of data; 2) public availability and reusability of scientific data; 3) public accessibility and transparency of scientific communication; 4) using web-based tools to facilitate scientific collaboration (Gezelter 2009).

Since the main focus of this article is research data (language resources), then it is necessary to analyse the concept of open data. Open data as defined by OECD (2015) are “data that can be used by anyone without technical or legal restrictions. The use encompasses both access and reuse”. Legal rights covering research data can make the dissemination and reusability of data a complex issue. For instance, data can be categorized as personal and non-personal data. The General Data Protection Regulation (GDPR) defines personal data as “any information relating to an identified or identifiable natural person”

¹ 1 CLARIN deposition license agreements define language resources as “material owned by the Copyright holder as defined in this Agreement, including software, applications and/or databases”. Available at <https://www.clarin.eu/content/licenses-agreements-legal-terms> (7.3.2018). In this article the terms ‘CLARIN language resources’, ‘CLARIN resources’, ‘language resources’ and ‘resources’ are used as synonyms.

² CC BY (Creative Commons Attribution) is a good tool which strikes a fair balance between objectives of open science (enhance dissemination and reuse of research results) and interests of individual researchers to get credit for their work. According to OECD (2015) data set citations could serve as incentives supporting open science.

³ Open Knowledge International is a global non-profit organisation focused on realising open data’s value to society. Information available at <https://okfn.org/> (15.4.2017).

(Art. 4). Processing⁴ of personal data must be lawful (GDPR Art. 5 and 6). Many language resources contain personal data (e.g., in the form of a person's voice). Language resources also contain material protected by copyright and related rights. All of these must be considered in the dissemination of language resources.

Given that language resources encompass software, it is necessary to understand open source and free software distribution. According to the Open Source Definition provided by the Open Source Initiative (2007) open source software has to comply with the following criteria: 1) Free redistribution; 2) Source code; 3) Derived works; 4) Integrity of the author's source code; 5) No discrimination against persons or groups; 6) No discrimination against fields of endeavor; 7) Distribution of license; 8) License must not be specific to a product; 9) License must not restrict other software; 10) License must be technology-neutral.

According to the Free Software Foundation (FSF), the user of free software must have four freedoms: 1) the freedom to run the program; 2) the freedom to study how the program works, and change it; 3) the freedom to redistribute copies; 4) the freedom to distribute copies of the modified versions to others.

Generally speaking, the aim of open science is to make the material (publications, data, software) accessible and reusable.

3 Re-thinking the CLARIN framework for the management of language resources

3.1 The CLARIN classification system

The legal classification system of the CLARIN language resources is based on a tripartite division of resources: PUB (public), ACA (academic), RES (restricted)⁵, based on their license (CLARIN license classification system). In general, PUB resources are available to all, ACA resources require that users have a researcher status in order to be accessed, while RES resources can only be accessed by authorised users. Further conditions apply to all categories.

The researchers who created the license categories of CLARIN resources (Oksanen et al. 2010) provided arguments to explain their choices. First, the categorization was based on an extensive survey. Second, it was argued that the licensing categorization must take into account licensing terms, such as limiting the distribution to academia or to even more limited groups of users, that are not covered by standard licenses such as Creative Commons⁶ but which are commonly used for language resources.

The three categories are defined through specific requirements:

- PUB resources should have no use limitations (e.g. based on geographic location, purpose of use, etc.). Recommended licenses are the Creative Commons Zero (CC0)⁷ or the Open Database License⁸ (ODbL).
- ACA resources must be available for study, research and teaching purposes.
- The availability of RES resources is even more limited. Their use requires following specific ethical or personal data protection requirements (Oksanen et al. 2010).

The PUB, ACA, RES categories may also be subject to additional conditions such as non-commercial use (NC), non-derivative use (ND) and to redeposit modified resources with CLARIN (RED) (Oksanen et al. 2010).

Although the categorization scheme has been incrementally improved (see Kelli et al. 2015), the conceptual framework remains the same. The question is whether the division of resources into PUB, ACA and RES category scheme can or should be improved in the light of an open science policy.

⁴ Processing *inter alia* covers collection, storage, adaptation, retrieval, use, dissemination and erasure of personal data (GDPR Art. 4).

⁵ The tripartite division is not unique. For instance, ORCID also has three levels for access to data: 1) everyone; 2) trusted parties; 3) only me. Additional information available at <http://support.orcid.org/knowledgebase/articles/124518-orcid-privacy-settings> (17.4.2017).

⁶ For additional information, see Creative Commons. Available at <https://creativecommons.org/> (7.3.2018).

⁷ From the researcher perspective, CC BY would be a better option since it allows the researcher to get credit for his or her work.

⁸ For additional information, see Open Data Commons Open Database License (ODbL). Available at <https://opendatacommons.org/licenses/odbl/1.0/> (3.7.2017).

3.2 Open and Public

In many international and regional legal instruments, the concept of openness is not defined. However, it is used in different policy documents (Berlin declaration, BOAI 2012, etc.), by different institutions (OECD, EU), and organizations (Open Knowledge International, Open Source Initiative). The names of some standard license also include the term “open” (e.g. Open Database License). This, however, is not the only practice for naming licenses.

The term *public domain* originated from the French *domaine public*, which was coined in the first copyright legislation. Public domain referred to objects which were no longer protected by an exclusive privilege and belong to the king (i.e. the nation, hence the term public). The term *domaine public* is also used in French civil law to designate public property (roads, public schools etc.)). In France many scholars use the term *domaine commun* (common domain) instead of the archaic *domaine public*. In Germany, the term is *Gemeinfreiheit*, which translates as "common freedom". In international usage, the term 'public domain' refers not only to what is no longer protected by copyright, but also to what has never been protected by copyright (e.g. ideas). In French, there is a distinction between *domaine public* (no longer under copyright) and *fonds communs* (never under copyright), which is crucial given that moral rights are inalienable and non-transferrable.

Several well-known standard licenses such as the European Union Public License⁹ (EUPL), GNU General Public License¹⁰ (GPL), Eclipse Public License¹¹ (EPL) and Mozilla Public License¹² (MPL) use the term “public” in their title. There is also the term “free” used in the title of several standard licenses (e.g., Academic Free License¹³, Free Public License¹⁴). In “public license” the term “public” means “offered to the whole public”.

A public license can be accepted (i.e. concluded) by any member of the public, but it can be very restrictive (CC BY-NC-SA is a license that is public, but not open). Moreover, Creative Commons licenses are also (in their respective texts) expressly referred to as public licenses.

In sum, “public” means “directed or related to a public”, open in this context means at least accessible for everyone and for any purpose.

One option to move forward is to replace “public” with “open”. As a result, more resources would technically be classified as “restricted”, in contrast to “open”. This is a more general sense of restricted than what is intended by CLARIN RES, which is defined in contrast to public or academic.

License	CLARIN present class	Open definition class	New CLARIN class
CC-0	PUB	OPEN	OPEN
CC-BY	PUB	OPEN	OPEN
CC-BY-SA	PUB	OPEN	OPEN
CC-BY-NC	PUB	NOT OPEN	RES
CC-BY-NC-SA	PUB	NOT OPEN	RES
CC-BY-ND	PUB	NOT OPEN	RES
ODC-BY	PUB	OPEN	OPEN
GPL	PUB	OPEN	OPEN
META-SHARE Commercial No redistribution	PUB	NOT OPEN	RES

Table 1. CLARIN PUB licenses with their present and proposed classes.

⁹ Additional information on EUPL is available at https://joinup.ec.europa.eu/community/eupl/og_page/european-union-public-license-eupl-v11 (17.4.2017).

¹⁰ Additional information on GPL is available at <https://www.gnu.org/licenses/gpl.html> (17.4.2017).

¹¹ Additional information on EPL is available at <https://www.eclipse.org/legal/epl-v10.html> (17.4.2017).

¹² Additional information on MPL is available at <https://www.mozilla.org/en-US/MPL/> (17.4.2017).

¹³ Additional information on Academic Free License is available at <https://opensource.org/licenses/AFL-3.0> (17.4.2017).

¹⁴ Additional information on Free Public License is available at <https://opensource.org/licenses/FPL-1.0.0> (17.4.2017).

As an example for the changes this replacement would have, we listed in Table 1 some of the CLARIN PUB licenses and their present and proposed classes; for the "open" category, we follow the recommendations of the Open definition.¹⁵

Academic use

In CLARIN, academic (ACA) and restricted (RES) resources are both restricted for copyright or personal data protection reasons¹⁶. Note that licenses for "academic use" are not unique to CLARIN (see, e.g., Academic Free License¹⁷).

The concept "academic use" is admittedly vague and can cause confusion. The first question that arises is whether commercial research is covered or not. If not, then one option could be to replace the academic category with non-commercial (NC). This solution is problematic as well, however, as there is community-wide confusion regarding what types of use are "non-commercial" (Kamocki and Ketzan 2014). This argument is further supported by findings from the VLO, where the condition of "non-commercial use" is found across all three license categories (PUB, ACA and RES) - cf. Section 3.4.

Another feature of the ACA category is that it poses a requirement on affiliation of the user to a recognised higher educational or research institution (i.e. AFFIL=EDU). This is a crucial issue since it requires private firms and non-profit organizations to acquire a "home-for-the-homeless-researcher" status for their researchers so that they can access data in the ACA category. The affiliation condition can be upheld using the Eduroam network¹⁸, which is the secure, world-wide roaming access service developed for the international research and education community, and can thus cover both educational and research institutes. However, not all institutions from all European countries are yet connected to Eduroam. If a user is not part of the Eduroam network, they can apply for researcher status, in which case they do not need to apply for access to the ACA-labeled resources separately. In CLARIN there is already a technical solution called "home-for-the-homeless" by providing an ID for those that need to acquire individual access rights to RES-labeled resources, but are not yet securely identified. A similar technical solution can be provided as a "home-for-the-homeless-researcher", which in addition should require some documentation that a person is engaged in academic research. RES-labeled resources still require individual permission, e.g. due to personal data legislation.

This dichotomy of ACA meaning Academic Use vs. Academic User is reflected in the CLARIN license templates: the CLARIN ACA EULA mentions "educational, teaching or research", while the CLARIN ACA Deposition License Agreement (DELA) specifies two additional conditions: "ID: A user needs to be authenticated or identified.", and "EDU: A user needs to be affiliated with the community of academic researchers through a university".¹⁹ Either way, should CLARIN decide to keep the ACA category, this EULA and DELA should be either retired or made compatible. In the current state the depositor says they ask for the additional restrictions, but end-users are then not presented with those restrictions. This should of course be remedied in the EULA, although in practice end-users do not gain access unless they have been identified as having acknowledged researcher status.

It should also be pointed out that all ACA licenses are NORED, i.e. no redistribution. As researchers can anyway easily get access to the original point of distribution, there is no need to share ACA resources directly.

CLARIN is an exception, rather than the rule, in the use of an ACA license category. The license scheme by META, the Multilingual Europe Technology Alliance, contains no ACA category, but broadly distinguishes between commons (i.e. only for META-SHARE members) and restricted (with license categories for commercial/noncommercial, for-a-fee/not-for-a-fee, etc.).²⁰ The META-SHARE

¹⁵ For additional information, see Guide to Open Licensing. Available at <http://opendefinition.org/guide/> (7.3.2018).

¹⁶ ACA can be seen as a type of restriction which in CLARIN was considered important enough to be "upgraded" into a category of its own (just like educational use for the rights statements or embargoed access in the COAR vocabulary).

¹⁷ Additional information on the Academic Free License is available at <https://opensource.org/licenses/AFL-3.0> (17.4.2017).

¹⁸ Additional information available at <https://www.eduroam.org/> (7.3.2018).

¹⁹ CLARIN ACA EULA and CLARIN ACA DELA template are available at <https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN/ClarinetEULA> and <https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN/ClarinetSA> (7.3.2018).

²⁰ Additional information available at <http://www.meta-net.eu/meta-share/licenses> (9.3.2018).

catalogues use both the unrestricted/restricted dichotomy and the conditions of use for faceted browsing. The research licensing portal from our colleagues in DARIAH, contains public licenses.²¹

There are also differences in the regulations for research and/or educational purposes included in European and national legislations. It should be clear that the ACA category is meant as an interpretation of the license accompanying a resource; it does not say anything about the legislation covering research in a given country. A resource collected based on some national research exception and distributed with an ACA label, puts the burden on a user in another country to check their own legislation to see if their intended plans are conformant with it.

3.3 Overview of categorization schemes and use of the CLARIN scheme in the VLO

The use of classification systems (e.g. types of resource, domain, provenance information) in order to organize resources contained in digital catalogues is a common practice that contributes to efficient search and retrieval. Facets created on the basis of these systems allow users to browse through the catalogues and restrict their search space using multiple filters.

Facets related to access and usage are found in most digital catalogues and are among the ones most frequently applied by users, demonstrating the importance of users knowing if and how they can access a resource and under what conditions they can use it for their purposes. Although there are currently various efforts for standardization of metadata, there is not yet a single, widely accepted classification system for access and usage. Still, most of cases fall under the following options (not necessarily excluding one another):

- classification based solely on the license of the resource (e.g. CC-BY, AGPL, etc.);
- grouping of the licenses into categories (e.g. "open access", "free for educational purposes", etc.) and organization of the resources on the basis of their licences; these categories are mutually exclusive, i.e. a resource can only be assigned to one category based on its license;
- analysis of the licenses based on the conditions of use they regulate (e.g. "attribution", "non-commercial use", "fee required" etc.) and linking of the resources with the conditions of use of their license; in this case, a resource can be linked to one or more conditions of use.

The latter two options are not meant to replace licenses, but to support users in their search through the appropriate deployment of formal metadata elements and values.

The choice of the values used for these two options largely depends on the intended audience. They are usually selected and formulated in a way that users can have a general understanding of what they can do with the resources and understand in a user-friendly way some basic notions of the legal text.

For illustration purposes, we will briefly present here two classification schemes that are relevant to our purposes and that could help us in our discussion:

- the COAR controlled vocabulary of access rights: COAR is the Confederation of Open Access Repositories and one of its activities relates to the development of controlled vocabularies for bibliographic metadata to ensure interoperability between the various repositories²²; the access rights vocabulary²³ declares the degree of "openness" of a resource and has four values: *open access*, *restricted access*, *embargoed access* and *metadata access*. The last two values can be regarded as two types of specific restrictions (temporal restriction and content blockage), which are considered important enough to be promoted into values of their own.
- the rights statements that have emerged from a joint initiative of Europeana and the Digital Public Library of America²⁴: these include 12 rights statements ("high level summaries of the underlying rights status") mainly intended for use by cultural heritage institutions. Two main features are used in the creation of these statements: the copyright status and the declaration of permission or prohibition of selected uses, mainly use for educational purposes and use in commercial applications, which are the ones most frequently associated with cultural objects

²¹ Additional information available at <http://forschungslizenzen.de/> (9.3.2018).

²² For additional information, see COAR Vocabularies. Available at <https://www.coar-repositories.org/activities/repository-interoperability/coar-vocabularies/> (7.3.2018).

²³ For additional information, see Controlled Vocabulary for Access Rights (Draft V1). Available at http://vocabularies.coar-repositories.org/documentation/access_rights/ (7.3.2018).

²⁴ For additional information, see RightsStatements.org. Available at <http://rightsstatements.org/en/> (7.3.2018).

distributed via these institutions. What is also noteworthy is that there are specific statements for resources with unknown or doubtful copyright status, taking into account whether this has been investigated or not.

The CLARIN licensing categorization scheme is currently used for the facet "Availability" in the Virtual Language Observatory catalogue (VLO).²⁵ The VLO harvests metadata descriptions of language resources from CLARIN centres but also from any other source that uses the OAI-PMH harvesting protocol and has agreed to be harvested by CLARIN, such as the OLAC catalogue of language resources²⁶ () and EUROPEANA²⁷.

Given the fact that resources come from multiple sources that do not use the same metadata schema for describing them, there has been a mapping procedure to the CLARIN license categories from the original elements and values.²⁸ For resources whose metadata records included a metadata element for the license, the mapping was easy and straightforward. However, a large number of resource descriptions contained no element at all for licensing information or included a free text statement, such as "available for research", "please ask provider", "academic research only", etc.; where possible, mapping of these values to the license categories was decided. As a result, 509,971 resources have been tagged with one of the CLARIN labels, which amounts to around 31.5% of the total resources in the VLO.

Below are some statistics on how the categories have been used for different resources in the VLO as of January 7, 2018:

Categories of All Resources	Number of Resources	Category Distribution	Categorized
Public	269,673	52.3 %	31.5 %
Academic	138,740	27.2 %	
Restricted for individual	101,558	19.9 %	
Unspecified	1,109,949	–	68.5 %

Table 2. Overall distribution of license categories.

Comparing this with the distribution of resources containing Finnish:

Categories of Finnish Resources	Number of Resources	Category Distribution	Categorized
Public	24,437	96.4 %	98.9 %
Academic	51	0.2 %	
Restricted for individual	850	3.4 %	
Unspecified	275	–	1.1 %

Table 3. Distribution of license categories among Finnish resources.

Having examined the unspecified resources for Finnish, it seems that they have been mainly harvested from other sources than the Finnish CLARIN Centre and their metadata records have had no clearly specified license information.

For those who wish to analyse the license categories and subcategories, VLO already provides more advanced search options using keywords. Users can use "NC" as a search condition in VLO and get a list of "non-commercial" resources from all categories. Here are the figures as of January 7, 2018:

- Public (10,511)

²⁵ CLARIN Virtual Language Observatory. Available at <https://vlo.clarin.eu/> (7.1.2018).

²⁶ For additional information, see Open Language Archives Community. Available at <http://www.language-archives.org/> (7.3.2018).

²⁷ See <https://www.clarin.eu/faq-page/275#t275n3923> for a list of providers to the CLARIN VLO.

²⁸ The mapping has been the output of co-ordinated work between the VLO development team, the metadata curation experts from the Austrian Centre for Digital Humanities at the Austrian Academy of Sciences and the CLARIN Legal Issues Committee.

- Academic (70)
- Restricted for individual (959)
- Unspecified (580)

This serves as an example of why NC is not sufficient to describe Academic Use, as argued in Section 3.3, given that NC can be used in any category. Together, the figures also show that more than 100,000 Academic resources are available for academic activities also in a commercial setting, as it does not matter who or what entity produces the academic results. The key point is that when someone wishes to exploit the results commercially, they need to acquire the necessary rights. Note that currently ACA still comes with a need for the researcher accessing the data to be affiliated with an academic institution.

3.4 Alternative categorization

The question remains whether it would be reasonable to change the current categorization of resources. Considering the problems caused by license proliferation (e.g., the existence of conflicting clauses), it would, in theory, be preferable to rely on existing standard licenses (e.g. Creative Commons²⁹) rather than create new bespoke licenses to replace them. The problem is that the use of language resources cannot easily be based on well-known standard licences due to the many unique situations we have described as well as to the existence of legacy resources with licences that cannot be replaced. Additional permission and restrictions are fundamentally required. An alternative categorization scheme should therefore be considered.

One option is to divide resources into two main categories: **open** and **restricted** as proposed by P. Kamocki at the CLARIN annual meeting in Wrocław, 2015. This category scheme fits better, conceptually, with the open science doctrine, which is becoming increasingly supported and emphasized across the EU and globally. The transformation from PUB to Open would require moving some resources to a restricted category (when the license is not broad enough). The following diagrams exemplify the current and alternative categorization.

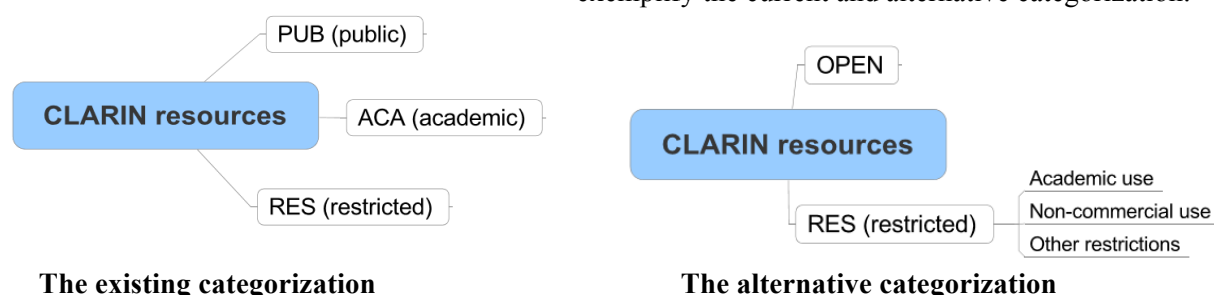


Figure 1. Existing and alternative categorization of licensing

Another alternative would be basically the same, but avoid using the Restricted label altogether. Reason for this would be that in a simple dichotomy of Open and the rest does not bring more information to the user, while labelling data with a not very positive, potentially even scary, label. In this proposed scenario, we would use one main category of Open, which is well defined: represented by established labels Open Access, Open Data and Open Source Software.

In addition to the Open category, we would also use a set of established labels for common license conditions and map them to existing licenses for quick user orientation. However, care must be taken to inform users to read the actual license³⁰, because “by” in CC-BY and for instance a copyright notice in MIT license are not exactly the same thing, just as the “SA” requirement as described in CC-BY-SA and in GPL-v2 differ. Still, for basic orientation it seems helpful to use these broad classification labels. We therefore propose to use established and easy to understand BY, NC, ND, SA labels from CC licenses, and possibly add some other that are used often in CLARIN³¹. If by reviewing current use of licenses in CLARIN we find that “research only” is a commonly used condition, we should add that label with some simple visualisation, as well.

²⁹ Additional information on Creative Commons is available at <https://creativecommons.org/about/> (17.4.2017).

³⁰ Users must be warned to read the licence text in all cases; as mentioned earlier, licensing categories and indication of conditions of use serve only as hints for the end-user and in no way replace the licences.

³¹ Work on gathering conditions of use linked to language resources, taking into account mainly CLARIN, META-SHARE and ELRA licenses has been initiated in the framework of the [W3C Linked Data for Language Resource Community Group](#)

4 Conclusion

As the open science doctrine becomes increasingly prevalent at national, regional and international levels, CLARIN's goals and policies should adapt to reflect this as it continues its mission of disseminating language resources as widely as possible.

Under its existing license category scheme, CLARIN resources are divided into three categories: public (PUB), academic (ACA), restricted (RES). This article analysed the existing categories and explored whether an alternative scheme, focusing on a division between “open” and “restricted”, would be more compatible with open science and be more useful for the CLARIN community.

Due to the cooperation of several authors with divergent suggestions, we have not yet reached final conclusions. The common understanding is that we need to continue our analysis among the CLARIN Legal Issues Committee. The article also serves as an indication of legal discussions relevant to CLARIN to the larger CLARIN community, so that additional voices may contribute.

Before making any final choices, we recommend a user survey to investigate the CLARIN community satisfaction with the current license category scheme, how accustomed they already are to using it and their interaction with other classification systems from other repositories. The current implementation of the classification system in CLARIN centres should also be taken into account.

Reference

- [Berlin Declaration on Open Access 2003] Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities of 22 October 2003. Available at <https://openaccess.mpg.de/Berlin-Declaration> (16.4.2017);
- [BOAI 10] Ten years on from the Budapest Open Access Initiative: setting the default to open (2012). Available at <http://www.budapestopenaccessinitiative.org/boai-10-recommendations> (15.4.2017);
- [CLARIN] CLARIN. Licenses, Agreements, Legal Terms. Available at <https://www.clarin.eu/content/licenses-agreements-legal-terms> (15.4.2017);
- [CLARIN license classification system] CLARIN. Licenses and the CLARIN license classification system. Available at <https://www.clarin.eu/content/licenses-and-clarin-license-classification-system> (7.3.2018);
- [CLARIN Value Proposition 2016] CLARIN Value Proposition (2016). Available at https://office.clarin.eu/v/CE-2016-0847-CLARINPLUS-D5_4.pdf (12.4.2017);
- [Estonian Copyright Act] Autoriõiguse seadus (valid since 12.12.1992). RT I 1992, 49, 615; RT I, 31.12.2016, 2 (in Estonian). Translation available at <https://www.riigiteataja.ee/en/eli/524012017001/consolide> (17.4.2017);
- [European Commission] European Commission (2016). Open innovation, open science, open to the world – a vision for Europe. Available at <https://ec.europa.eu/digital-single-market/en/news/open-innovation-open-science-open-world-vision-europe> (12.1.2018);
- [FSF] Free Software Foundation. What is free software? Available at <https://www.gnu.org/philosophy/free-sw.html#mission-statement> (13.1.2018);
- [GDPR] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). OJ L 119, 4.5.2016, p. 1-88. Available at <http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1515793631105&uri=CELEX:32016R0679> (12.1.2018);
- [Gezelter 2009] Dan Gezelter (2009) "What, exactly, is Open Science?", The Open Science Project. Available at: <http://openscience.org/what-exactly-is-open-science/> (7.3.2018).
- [Kamocki and Ketzan 2014] Paweł Kamocki and Erik Ketzan (2014) Creative Commons and Language Resources: General Issues and What's New in CC 4.0 (May 2014). Available at <https://www.clarin.eu/content/legal-information-platform> (12.1.2018);
- [Kelli et al. 2015] Aleksei Kelli, Kadri Vider, Krister Lindén (2015). The Regulatory and Contractual Framework as an Integral Part of the CLARIN Infrastructure. 123: Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wrocław, Poland. Ed. Koenraad De Smedt. Linköping University Electronic Press,

(cf. <http://www.cosasbuenas.es/static/ms-rights/> for the MS-rights ontology, which can be seen as an extension of the ODRL model). (Rodríguez and Labropoulou 2015).

- Linköpings universitet, 13–24. Available at <http://www.ep.liu.se/ecp/article.asp?issue=123&article=002> (14.4.2017);
- [OECD 2015] OECD (2015), “Making Open Science a Reality”, OECD Science, Technology and Industry Policy Papers, No. 25, OECD Publishing, Paris. Available at <http://dx.doi.org/10.1787/5jrs2f963zs1-en> (12.1.2018);
- [Oksanen et al. 2010] Ville Oksanen, Krister Lindén, Hanna Westerlund (2010). Laundry Symbols and License Management: Practical Considerations for the Distribution of LRs based on experiences from CLARIN' in Proceedings of LREC 2010: Workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management. Available at <https://helda.helsinki.fi/handle/10138/29359> (13.4.2017);
- [Open Knowledge International] Open Knowledge International. What is open? Available at <https://okfn.org/opendata/> (15.4.2017);
- [Open Definition 2.1] Open Knowledge International. Open Definition 2.1 Available at <http://opendefinition.org/od/2.1/en/> (15.4.2017);
- [Open Source Initiative 2007] Open Source Initiative. The Open Source Definition. Available at <https://opensource.org/osd> (13.1.2018);
- [Rodriguez and Labropoulou 2015] Victor Rodriguez-Doncel & Penny Labropoulou (2015). Digital Representation of Rights for Language Resources. In Proceedings of the 4th Workshop on Linked Data in Linguistics (LDL-2015), ACL-IJCNLP 2015, pages 49 - 58

Authorship and copyright ownership in the digital oral archives domain: The *Gra.fo* digital archive in the CLARIN-IT repository

Silvia Calamai

University of Siena, Italy
silvia.calamai@unisi.it

Aleksei Kelli

University of Tartu, Estonia
aleksei.kelli@ut.ee

Chiara Kolletzek

Lawyer and Records Manager, Bologna
chiara.kolletzek@live.it

Francesca Biliotti

University of Siena, Italy
francesca.biliotti@unisi.it

Abstract

The paper addresses the problem of authorship and copyright ownership in relation to a digital oral archive created through the digitisation of several analogue archives. The case study is provided by the *Gra.fo* digital archive (*Grammo-foni. Le soffitte della voce*, Scuola Normale Superiore di Pisa and the University of Siena, Tuscan Region PAR FAS 2007-13), a collection of around 30 Tuscan oral archives which is in the process of being documented in the CLARIN-IT repository¹. The problem is addressed from both an archival and a legal perspective, since speech and oral digital archives are at the crossroads of different fields of knowledge.

1. Introduction

Today, thanks to new and accessible technologies, oral recordings are enjoying a resurgence: on the one hand, technological progress has brought recording tools within everybody's reach; on the other, many existing analogue archives are being digitised in order to ensure their preservation. Both in the recording of new audio data and in the digitisation of already existing oral documents, three aspects must be taken into careful consideration. Firstly, long-term preservation of data and metadata is essential for the persistence of the data derived from a research project beyond its limited time span. Secondly, the choice of data and metadata formats is crucial in order to make data findable, available, interoperable and reusable. Thirdly, from a legal perspective, archives are covered with several rights. Oral recordings containing original contributions constitute copyright protected works. Persons involved in these recordings have inter alia related rights (the rights of performers) and are entitled to personal data protection. Archives are also protected as databases (see Kelli et al. 2015).

The case study for the present paper is provided by the project *Grammo-foni. Le soffitte della voce* (*Gra.fo*; Scuola Normale Superiore di Pisa and University of Siena, Tuscan Region PAR FAS 2007-13). *Gra.fo* discovered, digitised, catalogued and disseminated via a web portal (<http://grafo.sns.it/>) nearly 3000 hours of speech recordings stemming from around 30 oral archives collected by scholars and amateurs in the Tuscan territory, mostly in the 1960s and 1970s. The *Gra.fo* digital archive is a

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

¹ The authors wish to thank Pawel Kamocki and Irene Mecatti for their valuable comments and advice. Any errors or shortcomings are entirely the authors' responsibility.

heterogeneous collection of born analogue archives stemming from different disciplines and preserving a multitude of types of documents. Some of the archives originated from research projects conducted by linguists in order to document or investigate specific features of Tuscan dialects (e.g., the ‘Alto Mugello’, ‘Atlante Lessicale Toscano’ and ‘Carta dei Dialecti Italiani’ archives). Other archives were collected from anthropological, folkloric or ethnomusicological perspectives and thus concern folk music, folk literature and folk culture in general (e.g., archives ‘Edda Ardimanni’, ‘Roberta Beccari’, ‘Vanna Brunetti’, ‘Anna Buonomini’, ‘Paolo De Simonis’ - collection ‘Canti popolari del Mugello’, ‘FLOG’ - collections ‘Gilberto Giuntini’ and ‘Nunzi Gioseffi’, ‘Sergio Gargini’, ‘Benozzo Gianetti’, ‘Duse Lemetti - Gruppo Vegliatori’, ‘Museo del Bosco’). Yet others stem from history and sociology and deal with topics like working conditions in the twentieth century, the labour movement, women in the workplace, the Italian diaspora, the impact of industrialisation on rural society and memories of the First or Second World War (e.g., archives ‘FLOG - collection ‘Andrea Grifoni’, ‘ASMOS’, ‘Neri Binazzi’, ‘Cappelli di Paglia’, ‘Dina Dini’, ‘Elba’, ‘Roberto Segnini’, ‘Angela Spinelli’). It follows that the *Gra.fo* digital archive is also highly hierarchised: it is a digital archive of originally born analogue archives that in turn are made of subdivisions and that can be partitioned into further subdivisions (see Fig. 1a and Fig. 1b).

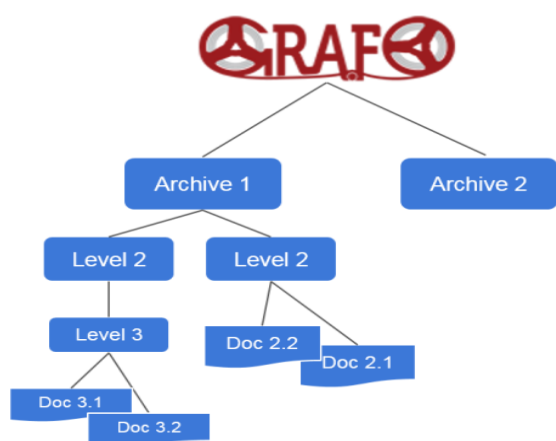


Fig. 1a The architecture of *Gra.fo* archives (as described in the paper).

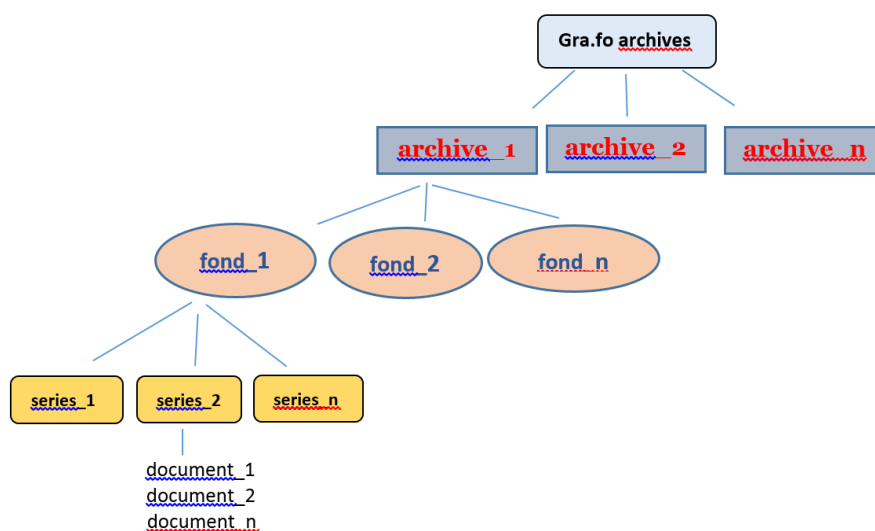


Fig. 1b The architecture of *Gra.fo* archives (with the original taxonomy used in the project).

As a collection of archives preserving valuable linguistic resources, the *Gra fo* digital archive is in the process of being documented in the CLARIN-IT repository. At the beginning of this documentation process, issues concerning searchability, granularity and consistency of the metadata descriptors were expected and described in Calamai, Frontini (2016). Several crucial legal issues concerning the relationship between the original archives (constituted by the recordings originally collected by the researchers) and the digital archives (created from the digital files derived from the original recordings), and the issues of authorship and copyright ownership applicable to these two entities also arose and are now being discussed. Since archives contain oral speech (human voice), personal data protection is a pertinent issue. The European Union's adoption of the General Data Protection Regulation (applicable as of 25 May 2018) makes the issue even more relevant. However, personal data protection concerning oral archives requires extensive analysis, which cannot be accomplished in this article because of its limited space and focus.

The paper introduces the description of the *Gra fo* archive in the CLARIN-IT repository (§2), posing the question of whether the original analogue or the digital archive should be the primary reference (§2.1) and trying to answer it (§2.2): what might be seen at first as a mere technical action (i.e., the metadata ingestion from an original dataset to another dataset) appeared in fact to be a rather challenging issue and required further development, at the crossroads of jurisprudence, archival science, and speech technology. Therefore, the legal aspects involved in the appropriate choice of metadata descriptors represent the focus of the paper. In particular, § 3 is devoted mostly to the discussion of the issues of authorship and copyright ownership (§ 3.1); a short section (§ 3.2) deals with the ownership of tangible objects (e.g., the analogue archives) – all of which lead to the conclusions closing the paper (§ 4).

2. The starting point

The CLARIN-IT repository² is managed by the Institute of Computational Linguistics “A. Zampolli” of the National Research Council (ILC-CNR). For the Italian SSH community, *Gra fo* represents a case study allowing the testing of the system in preparation for the documentation of other oral archives. Therefore, every decision made for the *Gra fo* archive – arduous though it may be – is an important benchmark for future developments.

The *Gra fo* digital archive enters the CLARIN-IT repository as an independent collection. The different archives comprised in the *Gra fo* archive are described as single items within the collection, while no description is provided for the single oral documents constituting the archives. This solution corresponds to the suggestion made by Calamai and Frontini (2016) as an alternative to the maximum-granularity and the minimum-granularity options.

2.1 Which archives need to be described for the metadata?

The description of the archives unveils the crucial problem of what should be the object of our description into the metadata set of information: the digital archives produced within the framework of *Gra fo* or the original ones produced by the researcher(s) who collected the recordings in the first place. Upon close inspection, most of the metadata descriptors could be interpreted in one way or the other leading to opposite options of description. For example, when indicating the contact person of a given archive, one could refer to the contact person(s) of the original archive (usually its author or owner, depending on the circumstances) or to the person(s) working for the digital archive (in this case, the *Gra fo* scientific coordinators); when stating the author of an archive, one could indicate the “creator” of the original archive (the researcher) or that of the digital one (the *Gra fo* consortium).

² <https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/>.

Such dichotomies could presumably apply to any archive that is transferred from the analogue to the digital domain. Yet in *Gra.fo* the situation is even more complex because the digital archives and the single digital oral documents accessible via the *Gra.fo* portal do not mirror the original ones, as they are the result of a meaningful interpretative activity and can therefore be considered derivative works (see the details in § 3.1.1.2). As described in Calamai, Biliotti, Bertinetto (2014), in speech recording fieldwork, a document (e.g., an interview, a narrative, etc.) can be distributed over various wireless carriers or portions of carriers, so that one and the same carrier may contain various unrelated documents while more than one carrier can refer to one and the same document (see also Kolletzek 2012). This led *Gra.fo* to consider the documental unit as independent from the carrier, which is viewed as a mere container, and to create new digital oral documents corresponding to the single communicative events (e.g. an interview, a narrative, etc.) contained in the original recordings (see Fig. 2).

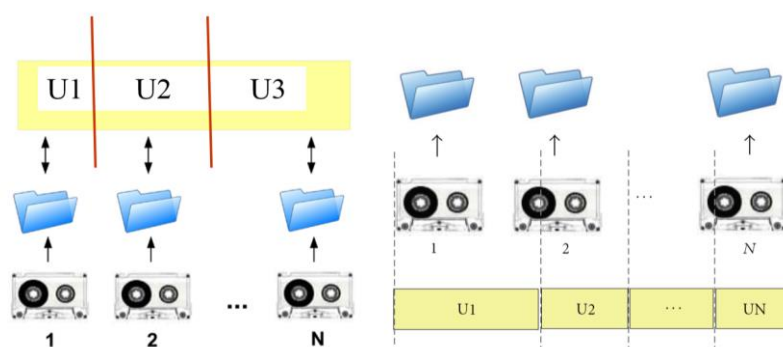


Fig. 2 The one-to-one relationship between the carrier and the documental unit, represented by ‘U’ (on the left) and the tricky relationship between the two in speech recording fieldwork (on the right).

After being edited from digital copies of the original recordings, these new digital documents are extensively described, transcribed (in some cases), and made available to the end user.

2.2 What we think we should describe

Our viewpoint is in line with the recently emerging idea that digitisation does not produce a mere copy of the physical reality; it rather produces a new reality that – as such – deserves recognition and a proper treatment (Sheridan 2017). In the *Gra.fo* project, the preservation process often produces something different from the original analogue document and the “final object” can be seen as the outcome of an interpretative process: digitisation is carried out by a technician who knows nothing about the content of the tapes, while the digital files that can be accessed via web (i.e., the audio document, the archival record, the potential accompanying documents) are created by an expert cataloguer.

Because the digital archives accessible via the *Gra.fo* portal do not mirror the original ones, we believe that, in describing the *Gra.fo* archives in the CLARIN-IT repository, the digital rather than the analogue archive should be the reference, provided that the source of the digital archive is clearly mentioned³. Accordingly, when stating the size of an archive, one should certainly indicate the number of digital oral documents it contains (rather than the number of open reel tapes or compact cassettes), since these are the documents that the user will find in the *Gra.fo* portal. Similarly, when stating the date of release of a given archive, one should refer to the date when the archive was made public in the *Gra.fo* portal. When indicating a contact person, one should always refer to the *Gra.fo* scientific coordinators, mentioning the contact person of the original archive only when that is deemed appropriate or useful for

³ This should be the case also when citing *Gra.fo* resources accessed via the CLARIN-IT repository.

some reason (in some cases the latter is merely the holder of the archive and is not willing to be contacted for issues related to it). Any relevant information concerning the original archive (who collected it, when and why, who owns it, etc.) will be provided in the Description box in the metadata record.

3. Authorship and copyright ownership in the *Gra.fo* oral archive

With respect to the metadata issues addressed in §2.2, one of the most challenging questions regards the metadata descriptor related to authorship and copyright ownership. The policies of the *Gra.fo* portal clearly state that all its contents are the fruits of the *Gra.fo* staff's labour. However, without the original archives, *Gra.fo*'s work would simply be non-existent. The researchers who collected the original recordings are the authors of the corresponding original archives and their authorship should be recognised. At the same time, the work of interpretation, editing, description and transcription carried out by the *Gra.fo* consortium certainly deserves recognition. The following paragraphs are devoted precisely to this thorny issue. The section on authorship and copyright ownership (§3.1 and §3.2) is organized as follows: the archival and legal frameworks are presented first, followed by their application to the *Gra.fo* archives, our case study. A shorter section (§ 3.3) is devoted to the ownership of tangible objects (e.g., the analogue archives). We will see that the archival definition and the legal definition of both concepts only partially overlap: both perspectives are nevertheless necessary to cope with authorship and ownership in the domain of a digital archive.

3.1 Authorship

3.1.1 The legal framework

Firstly, it is necessary to define what the meaning of “authorship” is, with respect to both the single document and the entire digital archive represented by *Gra.fo* (see Fig. 1). The distinction is relevant, especially if we consider that Italian law provides adequate forms of protection for

- A. the document/record itself, which can be defined (according to the ISAD(G) standards on archival description) as ‘recorded information in any form or medium, created or received and maintained by an organization or person in the transaction of business or the conduct of affairs’;
- B. the archive (or its subdivisions), defined by ISAD(G) as ‘the whole of the records, regardless of form or medium, organically created and/or accumulated and used by a particular person, family, or corporate body in the course of that creator's activities and functions’.

3.1.1.1 The document

Let us consider the document/records first. According to archival science, the “author” is “the individual or corporate body responsible for the intellectual content of a document. Not to be confused with creators of records” (ISAD (G)). This definition is inherent in any type of document as the expression of the intellectual activity of human beings and having informative content⁴.

According to the lawyers' point of view, two additional requirements are highlighted in the concept of “authorship”: the concept of “moral rights” on the one hand and the concept of “economic rights” on the other. In other words, the author has two sets of rights: 1) moral rights like the right of attribution (the right to be named as the author), the right of integrity (so that your work is not distorted, etc.), and 2) economic rights like the rights of reproduction (your right to make copies) and of distribution. Moral

⁴ As confirmed by the Latin origin of the word: *documentum* is derived from *docere* “to inform”, “to educate”, “to demonstrate”.

rights are not transferable but economic rights are. Nevertheless, such requirements are applicable only in the case that the document/record can be qualified as “work”, as will be discussed *infra*. The concept of authorship is, however, controversial and the intellectual property (IP) laws in continental Europe and common law countries diverge. Continental law countries – like Italy – usually define the author as a natural person and consider moral rights inalienable, not subject to transfer and imprescriptible. Common law countries accept legal entities as authors as well (e.g., in the case of work-for-hire) and authorship (or rather, copyright ownership) allows for transactions in both moral and economic rights (Ricketson, Ginsburg 2006, 358-363).

Let us tackle this issue one step at a time. Not all documents (in the archival sense) receive legal protection for the intangible content they represent. According to the Berne Convention for the Protection of Literary and Artistic Works (the Berne Convention),

The expression “literary and artistic works” shall include every production in the literary, scientific and artistic domain, whatever may be the mode or form of its expression, such as books, pamphlets and other writings; lectures, addresses, sermons and other works of the same nature; dramatic or dramatico-musical works; choreographic works and entertainments in dumb show; musical compositions with or without words; cinematographic works to which are assimilated works expressed by a process analogous to cinematography; works of drawing, painting, architecture, sculpture, engraving and lithography; photographic works to which are assimilated works expressed by a process analogous to photography; works of applied art; illustrations, maps, plans, sketches and three-dimensional works relative to geography, topography, architecture or science (Article 2.1).

The Convention then reserves to each Country the possibility “to prescribe that works in general or any specified categories of works shall not be protected unless they have been fixed in some material form” (Article 2.2).

According to the aforementioned principles, the Italian law on *Intellectual Property Rights* (hereinafter, the LDA) states that authorship is applicable to “original, creative works produced in literature, music, the visual arts, architecture, drama and cinema. Nevertheless, the list is merely illustrative, and not complete” (article 2). Therefore, a document/record might be considered as ‘work’ - and not only a mere vehicle of informative content – only in the case it meets the requirement of creativity, in the sense of ‘originality’ and ‘novelty’. Crucially, as stated by the jurisprudence of Italian courts⁵, the legal concept of ‘creativity’ does not coincide with that of ‘unconditional originality’ and ‘newness/novelty’; rather, it refers to a creative act – even if minimum – representing the ability to express a feeling, an idea, a fact, an action or whatever aspect of life in a personal way. The identification of the presence of such a requirement is anything but straightforward, since it refers to an interpretative activity, for which there are practically no objective criteria. The question whether oral records – mainly represented by interviews and answers to questionnaires – fall within the realm of the LDA is increasingly a matter of discussion among lawyers and scholars dealing with oral sources. Recently, two Italian lawyers coped with the issue of authorship in oral history interviews and stated that the author of an oral interview is the historian (Cortese, Giadrossi 2017).

With the clarification of this preliminary point, it is now possible to analyse the legal content of authorship with reference to the records being considered ‘works’, that is, having moral and economic rights. The definition of moral rights stems directly from the Berne Convention:

Independently of the author's economic rights, and even after the transfer of the said rights, the author shall have the right to claim authorship of the work and to object to any distortion, mutilation or other modification of, or other derogatory action in relation to, the said work, which would be prejudicial to his

⁵ i.e., Cass. Civ. sect. I, 12 March 2004 Decision No 5089.

honour or reputation (Article 6bis).

With a reference to the doctrine of moral rights, it is suggested that the Berne Convention might conceptualise the author as the natural person who created the work rather than any legal person or entity (Adeney 2006, 115). As for the economic rights, the Berne Convention presents an explanatory enumeration of all the business exploitation rights of the work: translation, reproduction, fair use, right of performance, right of broadcast, right of public replication, right of adaption. Exactly the same dichotomy can be found in the LDA⁶. Moreover, according to Italian law, the moral rights are inalienable, not subject to transfer and imprescriptible. Conversely, economic rights fall into the category of disposable assets, which originally lie with the author, and which the author may dispose of freely, according to the rules established by the LDA (Articles 12-19).

3.1.1.2 The archive

Let us now consider the archive from a strictly archival perspective. The concept of the creator of an archive is defined in ISAD(G) as “the corporate body, family or person that created, accumulated and/or maintained records in the conduct of personal or corporate activity. Not be confused with collector”. As in case of the documents/records reviewed above, the archival definition only deals with the ‘creator/producer’ – that is, it associates a set of objects with the person who created or collected them.

The point of view of the law is rather different, as stated in §3.1.1.1, and the component of ‘creativity’ is compulsory in the recognition of authorship. Moreover, the law highlights additional aspects: even in the case of archives, the need to protect moral and economic rights is recognised, irrespective of whether or not the recognition of authorship for the single records is present.

The Berne Convention, in Article 2.3, states that “Collections of literary or artistic works such as encyclopaedias and anthologies which, by reason of the selection and arrangement of their contents, constitute intellectual creations shall be protected as such, without prejudice to the copyright in each of the works forming part of such collections.” Although the Convention expressly cites the term “collections” and not “archive”, we think that the principle can also be applied to the latter. Such interpretation is substantiated also by the LDA, where the protection of databases is envisaged⁷: in this perspective, databases are considered ‘as a whole’, and they are characterised by a particular structure and by an internal architecture, which are the outcome of human creativity, in the sense described above (i.e., in the Italian LDA, the database *as such* is protected as a human intellectual creation, regardless of the individual records described in it. In other words, the LDA grants an independent specific protection to databases). In this case as well, Italian jurisprudence does not specifically use the label “archive”, but the rationale at the root of intangible heritage safeguarding projects such as *Gra.fo* allows us to consider digital archives as genuine databases. Therefore, the *Gra.fo* digital archive can be considered a derivative work of the analogue archives, on the basis of Art. 4 of the LDA:

Senza pregiudizio dei diritti esistenti sull'opera originaria, sono altresì protette le elaborazioni di carattere creativo dell'opera stessa, quali le traduzioni in altra lingua, le trasformazioni da una in altra forma letteraria od artistica, le modificazioni ed aggiunte che costituiscono un rifacimento sostanziale dell'opera originaria, gli adattamenti, le riduzioni, i compendi, le variazioni non costituenti opera originale (English Translation: Creative re-elaborations of an original work – like translations into other languages, transformations into other literary or artistic forms, modifications and additions embodying a substantial remaking of the

⁶ The LDA predates the Berne Convention by more than thirty years.

⁷ Arts 64 quinquies and 64 sexies.

original, re-arrangements and variations which cannot be considered original – shall also, without any loss or waiver of existing rights over the original work, be protected).

3.1.2 What about authorship in *Gra.fo*?

In this section we analyse in detail how the above-described legal concepts may be applied to the *Gra.fo* project. Firstly, it is relevant to verify whether records and archives (both the original ones and the digital archives created by *Gra.fo*) may be defined as ‘works’ and be protected by copyright protection (in this specific case, the Italian LDA). The answer is anything but simple and has relevant repercussions, in particular with respect to the moral and economic rights protection that the Italian LDA grants to a ‘work’ (a subject which may require more in-depth attention elsewhere). Secondly, it is worth bearing in mind the importance of making an exact attribution of authorship, since this fundamental topic is independent and separate from whether or not the record (or the archive) is considered ‘work’. If the record (or the archive) is labelled as ‘work’, it follows that the authorship attribution will also be associated with the legal safeguards that the LDA ensures to ‘works’. In the case that the record (or the archive) is not considered ‘work’, from a legal perspective the ‘author’ simply does not exist⁸.

While there is no doubt that the authors of the digital *Gra.fo* archive are the researchers who worked for the *Gra.fo* project, identifying the authors of the original archives is a more complex task. In the domain of oral archives, the author is not identifiable in a straightforward way, nor is he/she the only subject holding rights over an archive. Considering records, in particular, at least three entities are entitled to moral rights over an oral archive: the informant(s), the researcher(s) who collected the document and the individual/organisation commissioning the research. If we also consider economic rights, we should add the individual/organisation with whom/which the archive is deposited, who/which may have acquired some economic rights (Le Draoullec 2006, Stéphan 2013).

Discerning which of these entities should be given the status of ‘author’ in speech and oral archives might not be clear-cut. Looking at other experiences of archive preservation and dissemination in the world is certainly useful, but the fact that every country has its own applicable legislation⁹ means that the guidelines derived from single experiences are not easily comparable. Speech and oral archives contain very different types of recording material: from answers to questionnaires where the respondent’s task is limited to the translation of linguistic elements into his/her own dialect to very long monologues in which the interviewer remains silent and acts as a mere witness. The range of cases between these two extremes is ample and varied and introduces the thorny problem of genres and elicitation styles in oral material collection. Take, for example, the particular case represented by improvised oral poetry where the poet creates something totally original and can clearly be considered the ‘author’ (he/she can even profit from his/her works financially). Such an issue is obviously extremely complex also in the case of ‘collective’ archives created in the context of a geo-linguistic enterprise. The *Carta dei Dialetti Italiani* Archive (one of the greatest endeavours in Italian dialectology research: see Calamai, Bertinetto 2012) is a convenient example: one scholar conceived of and directed the enterprise; each region had its own research team that was directed by a coordinator; many different researchers carried out the fieldwork; and many speakers were interviewed. Thus, who should be recognised as the author? According to the LDA (Art. 7), we may consider Oronzo Parlangeli, who was at the time the scientific coordinator of the enterprise, the author of the *Carta dei Dialetti Italiani*. Nevertheless, each

⁸ In any event, from a strictly archival perspective, the metadatum related to ‘author’ remains valid, since such attribution will in any case be relevant for putting records and archives into their proper context.

⁹ Laws about authorship differ greatly from one country to another. In the United States tradition, according to MacKay (2016, pp. 75-76), ‘the speakers in the recorded interview automatically own their own words from the moment they are spoken, until or unless transferred to another entity through a legal release agreement’. According to French norms, however, researchers are the authors of their recordings, though they might share the role of co-authors with the interviewees in cases in which the latter participate in the exchange creatively (Stérin, 2016).

research project has its own history and requires specific interpretation – assuming that this is possible in the case of born analogue archives – in order to successfully identify the ‘author’.

3.1.3 What about copyright ownership in *Gra.fo*?

The creation of the *Gra.fo* digital archive, entailing complex decision making, certainly required creativity and thus can be considered a derivative work of the analogue archives, as already stated.

In the *Gra.fo* project, the archives’ copyright holders were asked to sign a legal agreement to grant their rights over the recordings to the *Gra.fo* project so that these could be digitised, catalogued, transcribed, and disseminated through the web portal (see Appendix 1). By means of the legal agreement between the copyright holders (in most cases, also the owners of the analogue archives) and both universities working on the *Gra.fo* project (SNS and UNISI), it was possible to digitise, describe, catalogue and analyse the oral documents contained in every single archive and then, finally, create a digital archive accessible via the web with specific search boxes.

Moreover, the main investor in the Project, the Tuscan Region, signed a legal agreement with SNS and UNISI in order to regulate copyright rights and results dissemination (Art. 13 of the Agreement): ‘Diritti di proprietà intellettuale e diffusione dei risultati. I diritti di sfruttamento economico e di utilizzo dell’innovazione eventualmente prodotta dal progetto di ricerca appartengono in misura uguale alla Regione Toscana ed ai soggetti attuatori i progetti medesimi’ (English Translation: Copyright rights and results dissemination. Economic rights and, when appropriate, rights of use of the innovation produced by the research project belong equally to the Tuscan Region and to the implementing bodies of the project itself”).

3.2 Ownership

3.2.1 The legal framework

The previous section addressed the complicated issue of authorship with respect to oral records and archives. We now tackle the issue of rights *in rem*¹⁰, commonly labelled “ownership” rights and only in appearance less complicated than intellectual property rights.

According to Italian law, the concept of ‘ownership’ in the commonly accepted meaning of the term corresponds to three different ideas: ‘property’, ‘possession’, and ‘custody’. Property refers to the whole set of rights *in rem* over an object: according to the Italian Civil Code, it allows the owner/title holder to make any lawful use of the good that is mainly expressed in the power of enjoyment and disposition (“the right to enjoy and dispose of things in a full and exclusive way” – article 832). Property grants its holder full control over the good. It is also protected by the Italian Constitution in Art. 42, where, in the 2nd paragraph, it is said: “Private property is recognized and guaranteed by law, which determines the ways it is acquired and enjoyed, as well as its limits in order to ensure its social function”.

3.2.2 What about ownership in *Gra.fo*?

Figuratively speaking, oral archives have a very complex life. Researchers who spent their lives doing fieldwork might have guarded their archives jealously. This is often the case for those who financed their own research and are, therefore, also the owners of their archives (e.g., among the *Gra.fo* archives, the ‘Vanna Brunetti’ archive). Others, who received funding for their research, might have consigned their recordings to the funding organisation (like the ‘Duse Lemetti – Gruppo Vegliatori’ archive, entrusted to the Municipality of Galliciano, which financed the research), or to an organisation

¹⁰ The Latin word *res, rei* means a material good/object.

that guaranteed the physical conservation of the materials (like the ‘Angela Spinelli’ archive, which was handed over to the Istituto Culturale e di Documentazione Lazzerini in Prato). Some may have come into possession of an archive through their friends (e.g. the ‘Edda Ardimanni’ archive is made up of recordings that were collected by different, unidentified researchers and then donated to Edda Ardimanni). Yet others might have inherited an archive from a deceased relative (like the ‘Sergio Gargini’ archive, which belongs to his wife, Anna Buonomini)¹¹. Therefore, the owner may correspond to the researcher who collected the recordings, to the organisation funding the research, to the organisation guarding the archive, or even to other persons that were not involved in the research at all (researchers’ heirs or friends).

According to the legal agreement between the copyright holders and the *Gra.fo* project, the original recordings were retained for the time necessary to work on them and then returned to their legitimate owners (who have the property, possession and custody of the original archives), while the portal and everything it contains (including the edited audio files and the relative descriptions) belong (property, possession and custody) to the Scuola Normale Superiore, the University of Siena and the Tuscan Region.

4. Conclusion

The analysis of this case study brings us to the following conclusion: for the digital archives, the authors are the researchers who worked at *Gra.fo*, while the owners are SNS, UNISI and Tuscan Region, which supported and financed the research project. As for the original archives and records, it certainly appears more difficult to provide a single, clear answer, given the complexity of born analogue archives. Undoubtedly, the identification of both the author and the owner in such cases requires a complex, in-depth analysis with respect to the origin, the aims, the rationale and the members of the research project of every single archive. It is not infrequent that information retrieval is only partially possible, due to the passage of time, the scarcity of accompanying materials and a general lack of sensitivity to the problem of long-term preservation of the various primary research data (open cassettes, open reels, field-notes) once the research project has expired.

Among Italian scholars, issues related to rights over oral recordings are now taken into great consideration compared to the past, when little or no attention was given to legal and ethical issues. Undoubtedly, Italy lacks a reference point comparable to France’s *Questions éthique et droit en SHS* (<https://ethiquedroit.hypotheses.org/>). However, the increasingly urgent need to agree upon some guidelines is being recognised around the world; 2015 saw the launch of the Italian Oral History Association’s (AISO) *Good Practices in Oral History* (<http://aisoitalia.org/?p=4795>), the culmination of a lengthy process of reflection and discussion among oral historians, anthropologists, and legal experts (Bonomo, Casellato, Garuccio 2016; Casellato 2017). What the conclusions of the meetings of the AISO’s Good Practices working group seem to suggest is that there is no general rule for establishing who the author of an oral document is and, moreover, the identity of the author cannot be decided *a posteriori*: only the agreements made between interviewer and interviewee in the context of the interview can tell us who the author of that document is (Sinello 2015). Nevertheless, under the section “Use of the Interviews”, in *Good Practices* it is clearly stated that the title holder (*titolare*, in Italian) of the interview is the person who conducted it. The problem remains crucial for those undocumented archives created in the 60s and 70s (when authorship, privacy and personal data protection were not common issues among linguists and historians) which now demand hard work on the part of curators to reconstruct their history. Therefore, the inclusion of the *Gra.fo* archives in the CLARIN-IT repository appears to be not only a metadata ‘translation’, but also a refined reflection on authorship, ownership,

¹¹ Details on the single archive are provided in Calamai, Bertinetto (2014).

and on the relationship between original source and digital objects. For this reason, speech technology, jurisprudence, archival science and digital philology should cooperate closely in order to offer sustainable solutions to scholars dealing with speech and oral archives.

References

- [Adeney 2006] Elizabeth Adeney. *The Moral Rights of Authors and Performers. An International and Comparative Analysis*. Oxford University Press, 2006.
- [Berne Convention] Berne Convention for the Protection of Literary and Artistic Works of September 9, 1886, completed in PARIS on May 4, 1896, revised in BERLIN on November 13, 1908, completed in BERNE on March 20, 1914, revised in ROME on June 2, 1928, in BRUSSELS on June 26, 1948, in STOCKHOLM on July 14, 1967, and in PARIS on July 24, 1971, and amended on September 28, 1979 http://www.wipo.int/treaties/en/text.jsp?file_id=283698#P85_10661
- [Bonomo, Casellato, Garruccio 2016] Bruno Bonomo, Alessandro Casellato, Roberta Garruccio. 2016. “Maneggiare con cura”: un rapporto sulla redazione delle *Buone pratiche per la storia orale. Il mestiere di storico*, VIII, 2: 4-21.
- [Casellato 2015] Alessandro Casellato 2017. *Il mestiere della storia orale. Stato dell’arte e buone pratiche*. Archivio Trentino, vol. 1/2016: 75-102.
- [Calamai, Bertinetto 2012] Silvia Calamai, Pier Marco Bertinetto. 2012. Per il recupero della *Carta dei Dialecti Italiani*. T. Telmon, G. Raimondi, L. Revelli (eds.) *Coesistenze linguistiche nell’Italia pre- e postunitaria. Atti del XLV Congresso internazionale di studi della Società di Linguistica Italiana (Aosta/Bard/Torino 26-28 September 2011)*, 2 vols., Roma, Bulzoni: 335-356. [ISBN: 978-88-7870-722-1].
- [Calamai, Bertinetto 2014] Silvia Calamai, Pier Marco Bertinetto. 2014. *Le soffitte della voce. Il progetto Grammo-foni*. Vecchiarelli.
- [Calamai, Biliotti, Bertinetto 2014] Silvia Calamai, Francesca Biliotti, Pier Marco Bertinetto. 2014. Fuzzy Archives. What Kind of an Object Is the Documental Unit of Oral Archives?. M. Ioannides et al. (eds.) *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection EuroMed 2014*, LNCS 8740, pp. 777–785.
- [Calamai, Frontini 2016] Silvia Calamai, Francesca Frontini. 2016. “Not Quite Your Usual Kind of Resource”. Gra.fo and the Documentation of Oral Archives in CLARIN. *CLARIN Annual Conference 2016*, Oct 2016, Aix-en-Provence, France. <https://www.clarin.eu/sites/default/files/calamai-frontini-CLARIN2016_paper_14.pdf>.
- [Cortese, Giadrossi 2017] Cortese Fulvio, Giadrossi Alessandro 2017. Quid iuris? «Buone pratiche per la storia orale». Archivio Trentino, vol. 1/2016: 105-123.
- [General Data Protection Regulation] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance) OJ L 119, 4.5.2016, p. 1-88 (<http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1515442739220&uri=CELEX:32016R0679>).
- [ISAD (G)] CONSEIL INTERNATIONAL DES ARCHIVES, ISAD (G): General International Standard Archival Description. Second edition. Adopted by the Committee on Descriptive Standards Stockholm, Sweden, 19-22 September 1999 (<https://www.ica.org/en/isadg-general-international-standard-archival-description-second-edition>) - Date of access: 2018.01.01
- [Kelli et al. 2015] Aleksei Kelli, Kadri Vider, Krister Lindén 2015. “The Regulatory and Contractual Framework as an Integral Part of the CLARIN Infrastructure.” 123: Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wrocław, Poland. Ed. Koenraad De Smedt. Linköping University Electronic Press, Linköpings universitet, 13–24 (<http://www.ep.liu.se/ecp/article.asp?issue=123&article=002>).

- [Kolletzek 2012]. Chiara Kolletzek 2012. “Tracciato sonoro: l’approccio archivistico alla descrizione dei documenti sonori.” *Archivi e computer*, 22, 2.
- [LDA] L. 22 aprile 1941 n. 633 “Legge italiana sul diritto d’autore” - <http://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:legge:1941-04-22;633!vig> - Date of access: 2018.01.01
- [Le Draoullec 2006] Ludovic Le Draoullec. 2006. “L’utilisation des corpus oraux à des fins culturelles: quels contrats mettre en œuvre?”. *Bulletin de l’AFAS* (URL: <http://afas.revues.org/622>; DOI : 10.4000/afas.622).
- [MacKay 2016] Nancy MacKay. 2016. *Curating Oral Histories, Second Edition: From Interview to Archive*. Routledge.
- [Ricketson, Ginsburg, 2006] Sam Ricketson, Jane C. Ginsburg. 2006. *International Copyright and Neighbouring Rights: The Berne Convention and Beyond*. Second Edition. Volume I. Oxford University Press.
- [Sheridan 2017] John Sheridan. 2017. *Creating the Disruptive Digital Archive*. Digital Preservation Coalition (<http://www.dpconline.org/blog/disruptive-digital-archive>).
- [Sinello 2015] Rachele Sinello. 2015. *Le vite degli altri. Verso la definizione delle linee guida italiane per la storia orale*, Thesis, Corso di Laurea triennale in Lettere, Università Ca’ Foscari di Venezia, A.A. 2014/2015, supervisor: Prof. Alessandro Casellato.
- [Stéphan 2013] Lena Stéphan. 2013. *Les archives sonores : conservation et valorisation du patrimoine oral. Mémoires Master "Archives numériques"*, École Nationale Supérieure des Sciences de l’Information et des Bibliothèques.
- [Stérin 2016] Anne-Laure Stérin. 2016. *Le chercheur mène un entretien enregistré* (<http://ethiquedroit.hypotheses.org/1133>).

Appendix 1. Legal agreement between the archives' owners and SNS and UNISI University (original version)

Spett.le

Scuola Normale Superiore

Piazza dei Cavalieri, 7

Pisa

Oggetto: cessione diritti su fonoregistrazioni

L'Associazione/Biblioteca/Archivio "....."

(C.F.), con sede legale a

..... in via, rappresentata dal

Presidente e legale rappresentante *pro tempore*, dott./prof.(ssa)

....., il/la quale agisce in virtù dei poteri conferitigli da

..... (ad es., lo statuto dell'Associazione)

Premesso che

- Grazie a un finanziamento regionale (Avviso pubblico FAS 2007 2013 Delibera CIPE 166/2007 PAR FAS Regione Toscana Linea di Azione 1.1.a.3. per il per il sostegno a progetti di ricerca in materia di scienze socio economiche e umane), la SNS e l'UNISI hanno dato vita alla realizzazione di un progetto denominato *Grammo – foni. Le soffitte della voce*, volto a raccogliere, salvaguardare e analizzare documenti vocali di interesse linguistico mediante la loro digitalizzazione e la loro catalogazione, attraverso la costituzione di un archivio telematico centralizzato presso il Laboratorio di Linguistica della SNS, che – una volta realizzato – sarà regolato da apposite norme miranti ad evitare i rischi di usi impropri dei beni vocali raccolti;

- il materiale audio-fonico in possesso di XYZ abbia un valore storico, documentario e scientifico tale da giustificare la digitalizzazione, la descrizione, la catalogazione e la successiva messa in rete, onde permettere agli studiosi di poterlo utilizzare in futuro, evitando che questo possa andare disperso per il deterioramento o l'obsolescenza dei supporti materiali sui quali è attualmente conservato;

cede

in via non esclusiva e gratuita alla Scuola Normale Superiore, nella qualità di mandataria dell'Associazione Temporanea di Scopo (ATS) costituitasi tra la stessa e l'Università degli Studi di Siena (di seguito UNISI) per la realizzazione del progetto *Grammo-foni. Le soffitte della voce*, i documenti audio-fonici di cui ha il legittimo possesso / proprietà indicati nell'allegato A affinché siano digitalizzati, catalogati e trascritti (parzialmente o integralmente), e successivamente costituiscano parte costitutiva dell'archivio telematico centralizzato che sarà realizzato presso il Laboratorio di Linguistica della SNS per finalità di studio, ricerca e pubblicazione, previo trattamento volto a rendere la fonte non identificabile, su apposito sito web dedicato con accesso regolamentato.

Dichiara

- che i documenti audio-fonici ceduti all'ATS sono in suo legittimo possesso e di non aver già ceduto ad altri in esclusiva il materiale di cui all'elenco allegato;

- che qualora terzi dovessero rivendicare la titolarità di diritti su tali beni non compatibili con la presente cessione i relativi files dovranno essere rimossi dalla banca dati della ATS e/o trattati compatibilmente con i diritti di tali terzi.

- di essere consapevole che la pubblicazione avverrà con la citazione espressa della XYZ da cui la ATS ha acquisito, mediante il presente atto di cessione, i diritti sui dati allegati;

- di esonerare l'ATS da qualsiasi responsabilità qualora i documenti audio-fonici subiscano danneggiamento durante le fasi di digitalizzazione nonché da eventuali danni (purché non causati da grave negligenza) che possano verificarsi durante le fasi di trasporto;

- di non voler sostenere alcuna spesa di trasporto dei supporti materiali contenenti i dati fino alla sede in cui saranno oggetto di digitalizzazione;
- di volere una copia digitalizzata su opportuno supporto (fornito da XYZ) del materiale ceduto
- di volere accesso gratuito all'archivio telematico, fatte salve le eventuali restrizioni temporanee imposte da altri enti cedenti, ragioni tecniche o altra causa ostativa.

Pisa, 2011

Il Direttore/Presidente di

.....

Allegato "A".

(Descrizione dei materiali temporaneamente ceduti da agli esclusivi fini del presente accordo)

Appendix 1. Legal agreement between the archives' owners and SNS and UNISI University (English translation)

Scuola Normale Superiore
Piazza dei Cavalieri, 7
Pisa

Transfer of rights over sound recordings

The Association, Library or Archive known as

.....,
(NIN), whose registered office address is
..... and which is represented by its President and legal
representative *pro tempore*,, who is acting by virtue of the
powers bestowed on him/her by (e.g., The Statute of the Association)

Given

- that, thanks to a regional grant (Public Notice FAS 2007 2013 Decision of the Interministerial Committee on Economic Programming [CIPE] 166/2007 PAR FAS Tuscan Region Course of Action 1.1.a.3 in support of research projects in socio-economic sciences and the humanities), the Scuola Normale Superiore (SNS) and the University of Siena (UNISI) have created a project named *Grammo-foni. Le soffitte della voce (Grammo-foni. The Attics of Voice)* whose purpose is to collect, preserve and analyse vocal documents of linguistic interest by digitising and cataloguing them and by establishing a centralised electronic database at the Linguistic Laboratory of the SNS which, once this web archive has been set up, will be regulated by specific norms for the purpose of avoiding the risk of improper usage of the collection of vocal assets;

- that the audio material owned by XYZ has such historic, documentary or scientific value as to justify its digitisation, description, classification and publication online so that scholars may access it in the future and so that it will not be lost forever through the deterioration or obsolescence of the tangible medium currently preserving it;

Agrees

to grant a non-exclusive, royalty-free licence to the Scuola Normale Superiore, as the agent of the Associazione Temporanea di Scopo (ATS – a temporary task force constituted between the SNS and UNISI for the fruition of the project denominated *Grammo-foni. Le soffitte della voce*), to use the audio documents listed in Attachment A of which ATS is the legal owner/title holder so that they may be digitised, catalogued and transcribed (either partially or fully) and so that they may in future constitute an integral part of the centralised online archive to be created at the SNS Linguistic Laboratory for studies, research and publication, subject to prior processing for the purpose of making the source unidentifiable, through the specific website dedicated to this project and through regulated access;

Declares

- that the audio documents transferred to the ATS are in his/her legal possession and that the exclusive rights to the material listed in the attachment below has not been ceded to others;
- that should third parties claim ownership of the rights over these assets incompatible with the present transfer the relevant files will be removed from the ATS database and/or handled in a manner compatible with the rights of said third parties;

- that he/she is aware that the audio documents listed in Attachment A, which ATS has obtained from XYZ with this Act of Transfer, will be published in the form requested by XYZ;
- that he/she exonerates ATS from any responsibility should the audio documents be damaged during digitisation or during transportation (unless, of course, the damage was caused by grave negligence);
- that he/she does not wish to incur any expense for the transportation of the tangible mediums containing the audio material to the location where they will be digitised;
- that he/she wishes to have a digitised copy of said material in an appropriate medium (supplied by XYZ);
- that he/she wishes to have free access to the online database, except when other transferring parties have imposed temporary restrictions or when there are other obstacles, like technical problems.

Pisa, 2011

Director/President of

.....

Attachment "A".

(Description of the materials temporarily transferred by for the sole purpose described in this agreement.)

Literary Exploration Machine A Web-Based Application for Textual Scholars

Maciej Maryl
Institute of Literary
Research of the Polish
Academy of Sciences
Warsaw, Poland
maciej.maryl
@ibl.waw.pl

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

Tomasz Walkowiak
Faculty of Electronics
Wrocław University of
Science and Technology
tomasz.walkowiak
@pwr.edu.pl

Abstract

This paper presents a design of a web-based application for textual scholars. The goal of this project is to create a complex and stable research environment allowing scholars to upload the texts they analyse and either explore them with a suite of dedicated tools or transform them into a different format (e.g. text, table, list, spreadsheet). The latter functionality is especially important for research focusing on Polish texts (due to the rich morphology and weakly constrained word order of Polish) because it allows for their further processing with tools built for English. This project utilises the existing CLARIN-PL applications and supplements them with new functionalities.

1 Challenge

Digital literary studies seem to be among the most rapidly developing strands of Digital Humanities. The main obstacle to the development of the field lies in the users' lack of programming skills and insufficient knowledge of how to use digital methods and operate the existing tools. The authors of this paper were involved in various educational activities as organisers and instructors of CLARIN-PL workshops dedicated to endowing Polish scholars with digital methods of textual scholarship.¹ The main lesson learned from these endeavours is that although there is a genuine interest in computational literary criticism, the learning curve remains steep and workshop participants do not eventually incorporate those tools into their research workflows because they find them too complicated.

For instance, all basic language tools developed by CLARIN-PL are offered through web applications, which are seemingly easy to use from the perspective of their designers.² For instance, potential users of a morpho-syntactic tagger (Figure 1) need to simply upload a file to the web application to have their text tagged and morpho-syntactically disambiguated. Although the result contains complete information about word forms (lemmas, Part of Speech (henceforth PoS) and values of grammatical attributes), it is available only in an XML file, i.e. the CCL format (Marcinićzuk, & Radziszewski, 2013) native for the KPW Corpus (Broda et al., 2012). Deciphering the output may be challenging for non-professional users. For instance, let us take a look at the representation of a simple phrase *ciemny włos jej* 'her dark hair' presented in Listing 1, below, where *base* tag specifies a lemma and a positional tag *ctag* encodes PoS and values of grammatical attributes.

¹ Materials from the workshops are accessible at <http://clarin-pl.eu/en/mediateka-2/>

² <http://ws.clarin-pl.eu/>

Morpho-syntactic tagger

Used tools ▾

Instructions ▾

Options

☐ GUESSER
 ☒ MORFEUSZ 1
 ☒ MORFEUSZ 2

Textarea

Load corpus

Litwo! Ojczyzna moja! ty jesteś jak zdrowie. Ile cię trzeba cenić, ten tylko się dowie, kto cię stracił. Dziś piękność twą w całej ozdobie widzę i opisuję, bo tęsknię po tobie. Panno Święta, co jasnej bronisz Częstochowy i w Ostrej świecisz Bramie! Ty, co gród zamkowy Nowogródzki ochraniasz z jego wiernym ludem! Jak mnie dziecko do zdrowia powróciłaś cudem (gdy od płaczącej matki pod Twoją opiekę ofiarowany, martwą podniosłem powiekę i zaraz mogłem pieszo do Twych świątyń progu iść za wrócone życie podziękować Bogu), tak nas powrócisz cudem na Ojczyzny łono.

Analyse

Clear

Figure 1. Web-based User Interface for the morpho-syntactic tagger.

```

<chunkList> <chunk id="ch1" type="p">
<sentence id="s1">
<tok> <orth>Ciemne</orth>
    <lex disamb="1"> <base>ciemny</base>
    <ctag>adj:sg:nom:n:pos</ctag></lex> </tok>
<tok> <orth>włosy</orth>

```



```

<lex disamb="1"> <base>włos</base>

<ctag>subst:pl:nom:m3</ctag></lex>    </tok>

<tok> <orth>jej</orth>

<lex disamb="1"> <base>on</base>

<ctag>ppron3:sg:gen:f:ter:akc:npraep</ctag> </lex>

</tok>

```

Listing 1. An example of the XML-based format for the output of a tagger from Figure 1.

Despite such rich description, this output is useless for those users who are not familiar with XML or are unable to convert this input to a desirable form, e.g. a list of lemmas. Users from Humanities and Social Sciences, without the background in Computer Science, do not possess such skills, which dramatically reduces usefulness of this service as a research tool and effectively excludes it from their workflows. Furthermore, the results of the 2015 survey into digital methods and practices, conducted by the DARIAH DiMPO Working Group, seem to corroborate these observations: scholars clearly articulate a need for technological support and guidance concerning the existing tools (Dallas et al., 2017: 7).

To address these challenges we have developed a web-based system, called *Literary Exploration Machine* (LEM),³ which:

- does not require installation,
- aggregates existing language tools for Polish,
- has a modular architecture and can be expanded through the addition of new components;
- allows for the preprocessing of texts to make them compatible with tools developed for other languages, e.g., CLUTO (Zhao & Karypis, 2005), or Mallet (McCallum, 2002),
- offers a simple workflow not requiring programming skills,
- provides elaborated descriptions of tools, outputs, and parameters, and aims at supplementing them with rich use-case descriptions.

This approach makes LEM similar to other ‘one-stop-shop’ initiatives which make sophisticated tools accessible to less experienced users. A good example is DARIAH-DE/Topics⁴ package, which provides interface and tailored workflows for topic modeling. However, LEM offers a wider variety of tools and techniques. Popular *Voyant*⁵ allows for quick analysis of the word forms and their relative frequencies across texts, supplemented by a range of NLP tools based on the Stanford CoreNLP, e.g. Proper Nouns recognition. However, *Voyant* is dedicated to texts written in English, so its applications remain limited. Nevertheless, *Voyant* visualisation methods and a friendly GUI remains a gold standard for LEM.

2 Design of the system

LEM was developed in a user-driven paradigm through interdisciplinary cooperation of computer scientists, linguists, literary scholars, and sociologists. All functionalities, options, and output formats were defined and described in a series of case studies, dedicated to particular research problems in Digital Humanities:

1. *Literary studies:*

³ <http://ws.clarin-pl.eu/lem.shtml>

⁴ <https://github.com/DARIAH-DE/Topics>

⁵ *Voyant*: <http://docs.voyant-tools.org>, CoreNLP: <https://nlp.stanford.edu/software/>

- a. research on secondary literature, i.e. academic journal articles. A comparative study of the transformation in the Polish literary scholarship 1989-2014 (Maryl 2016a).
 - b. Text preprocessing for advanced stylometric analysis of authorship, genre, and chronology in a corpus of novels (Rybicki, 2017).
 - c. analysis of an online genre and its evolution overtime on the example of a web portal (Maryl 2016b).
2. *Sociology*: multi-feature classification of teachers' attitudes on the basis of students' comments. Complex preprocessing (e.g., counting co-occurrences of selected lemmas and PoS) served as an input for statistical data analysis, e.g. (Bryda & Tomanek, 2015), cf (Brosz et al., 2017).
 3. *Social psychology*: sentiment analysis in the studies of depression and emotions in Polish texts, inspired by a workflow of Linguistic Inquiry and Word Count (LIWC) (Tausczik & Pennebaker, 2010),⁶ e.g. (Rohnka et al., 2015).

Each of these case studies produced the following content:

- a. research questions which could be answered through a computational analysis;
- b. relevant textual source material;
- c. design of tools capable of answering those questions;
- d. development of actual tools;
- e. rich description of a tool with a hands-on guide for other scholars.

LEM is built on the processing engine WebSty,⁷ an open stylometric system, which provides tools for the statistical description of text corpora, texts and subcorpora comparison. However, LEM is designed for a variety of research uses, so in addition to WebSty it allows for an analysis of other linguistic levels:

- basic description of words
 - segments (e.g., length of documents, paragraphs, and sentences),
 - morphology (word forms, punctuation marks, pseudo-suffixes, and lemmas),
- combined morpho-syntactic information
 - grammatical classes and categories (based on the Polish National Corpus tagset (Przepiórkowski et al., 2012)), as well as their n-grams,
- lexical semantics
 - proper names and their semantic categories,
 - word senses and their selected lexico-semantic relations like synonymy or hypernymy.

All of the LEM **processing paradigms** are designed to fit into this general workflow:

1. *Uploading* a corpus of documents together with their metadata (thanks to the compatibility with CLARIN repositories, Component Metadata format is supported, but also simple metadata can be read from the file names).
2. *Text extraction* and cleaning. OCR-ed documents usually contain many language errors which can be corrected at this stage.
3. *Selection of features* for the description of documents is done manually by users or available by default, depending on the processing paradigm. This step is based on the hands-on guide.

⁶ <https://liwc.wpengine.com/>

⁷ WebSty: <http://websty.clarin-pl.eu/>, Mallet: <http://mallet.cs.umass.edu/>

Users are not expected to have advanced knowledge of Natural Language Engineering or Data Mining.

4. *Setup of parameters* for processing is also customised, but some default settings of parameters are provided. More advanced users will be able to tune the tool to their needs.
5. *Text preprocessing* with language tools provided by CLARIN-PL. Each text is analysed by a PoS tagger, e.g. *WCRFT2* (Radziszewski, 2013), and eventually piped to a Name Entity Recognizer, i.e. *Liner2* (Marciniczuk et al., 2013), a temporal expression recognizer and a word-sense recogniser, e.g., *WoSeDon* (Kędzia et al., 2015), etc.
6. *Feature values calculation*. Selected features of the preprocessed texts are extracted together with their frequencies and annotations by comparing patterns defining the features with every position in a document.
7. *Filtering and/or transforming* the original feature values. Most filtering and transformation functions are provided by WebSty and its components. Further data-analysis features allowing for advanced comparison of corpora will be added.
8. *Data mining*. Several processing paradigms are employed to allow for gathering more complex information about the data, namely *topic modelling* (representing a document in terms of subsets based on word co-occurrences), *unsupervised clustering* (grouping documents on the basis of the document-feature vectors similarity, and *supervised classification* (a prototypical application of Machine Learning based on *Weka* (Witten et al., 2017),⁸ *scikit-learn* (Pedregosa et al., 2017), and *SciPy*⁹ packages (Jones et al., 2018), trained on documents manually classified by users).
9. *Presentation of results*. The results are presented either as interactive visualisation, or as downloadable files in formats compatible with external exploratory tools and programs (e.g., spreadsheets or *Gephi*).

3 Current functionality of LEM

From the user's perspective, the complex workflow described above could be translated into a simple, three-step procedure:

1. upload texts to be processed,
2. choose the task and its parameters,
3. browse or download the results.

In the first step, users need to prepare a ZIP archive with texts they want to analyse. LEM accepts most of the popular formats: TXT, RTF, DOC, DOCX, ODT, XLSX, PDF. Files could be uploaded directly from the hard drive, from the URL, or from CLARIN Cloud storage (based on the NextCloud technology and maintained by CLARIN-PL)¹⁰. LEM was designed for efficient processing of large volumes of data. However, the size and the number of files to be processed is limited for common users, due to processing workload and limitations of some of the output formats (e.g., XLSX). Larger datasets can be processed with the assistance of the project team.

The results of processing depend on the quality of the corpus. Optionally, users can provide a 'stoplist', i.e. a list of words or characters which should be excluded from the analyses. It is especially convenient when users want to filter out OCR mistakes or words overrepresented in the corpus. For instance, in the corpus of academic articles we used for a case study discussed below, the numerals were overrepresented because of the footnote numbers. Knowing that we could easily exclude them from further processing.

⁸ Weka: <http://www.cs.waikato.ac.nz/ml/weka/>, scikit-learn: <http://scikit-learn.org/stable/>, SciPy: <https://www.scipy.org>, Gephi: <https://gephi.org>

⁹ <https://www.scipy.org>

¹⁰ <https://nextcloud.clarin-pl.eu/>

CLARIN-PL

[SERVICE INDEX](#)
[ABOUT APP](#)
[ABOUT](#)

PL

Literary Exploration Machine

Used tools ^

Apache Tika- files to text converter

Morfeusz 2 with SGJP dictionary (for morphological analysis)

WCRFT2 tagger, WSD

Liner2, Polish Wordnet, WebSty

Instructions ^

Drag and drop ZIP package with files in specific formats: txt, doc, docx, pptx, xlsx, odt, pdf, html, rtf - they will be automatically formatted to text format

Next press "Analyse" button and wait for the results to be displayed. The bigger files for rendering, the longer waiting time for loading results (progress bar will be displayed)

After completion of choosen task the Result will appear below - download the result and proceed as you need.

Options

TYPE OF TAGGER

Morphodita

☒ USER STOPLIST

CHOICE OF TASK

Lemmatisation

Lemmatization – a process of deriving the base form of inflected word. The base form is always grammatically correct word, e.g. nominative singular for nouns. In linguistics the base form of a word represents all correct inflected forms, and is called a lemma, thus the name of the process. Lemmas always have meaning. The process is complex and relatively slow but offers very high accuracy. Closely connected to stemming, the process of reducing inflected word to its root, lemmatization is however more complicated, as its success depends on successful interpretation of the morphological, syntactical and semantical properties of the given token

Input data

ZIP FILE

URL

NEXTCLOUD

URL of zip file

http://ws.clarin-pl.eu/public/teksty/2mini.zip

Example corpora

mini korpus

Analyse

Clear

Authors: Maciej Maryl, Maciej Piasecki, Tomasz Walkowiak webserwis@clarin-pl.eu

Figure 2. LEM web-based User Interface – the main screen.

Users can also choose between different morphological analysers: two versions of the morphological analyser *Morfeusz*¹¹ (Woliński, 2014) combined with the WCRFT tagger and a new MorphoDiTa-pl morphosyntactic tagger (Piasecki and Walentynowicz, 2017) including a morphological analyser. *Morfeusz 1* is the older of the two and its dictionary is smaller, and the words which do not appear in it are left unlemmatized. *Morfeusz 2*, on the other hand, has a significantly larger built-in dictionary, and provides some additional features (e.g., classification of proper names).¹² The third tagger, MorphoDiTa, was developed by CLARIN-PL and is based on neural networks. Its model is frequently modified through a constant learning process, hence the processing results for the same text may differ over time. The use of *Morfeusz 1* can result in a higher accuracy of the tagger's output, *Morfeusz 2* recognises more word forms on the basis of its dictionary, but the tagger may still make some mistakes when choosing the correct form, while MorphoDiTa-pl was trained already on the NCP automatically transformed to the annotation compatible with *Morfeusz 2* and uses its own morphological analyser and guesser based on SGJP dictionary (Saloni et al., 2015). Researchers can select the tagger which suits their research questions or source materials best. MorphoDiTa-pl should produce the best results in most cases for the cost of lower efficiency. However, the effect of the lower efficiency can be visible only for a larger amount of text to be processed.

In the second step, users choose one of the processing tasks, which are described in the following subsections in more detail. Once the version of a morphological analyser and a task are selected, users click the “Process” button to perform an analysis and may observe the progress of the analysis on the percentage bar. When the processing is done, users can access the results through an interactive visualisation or a downloadable output file. What follows is a detailed description of tasks.

3.1 Lemmatisation

Lemmatisation is a task of converting inflected word forms – words as they appear in sentences – to *lemmas* (basic morphological forms, dictionary forms). It is closely connected to *stemming*, which was introduced in Information Retrieval as a process of reducing inflected words to their pseudo-roots (not always proper words), cf Manning et al. (2008). Lemmatisation is, however, more complicated, as its success depends on the successful interpretation of the morphological, syntactic and semantic properties of a given token. Manning et al. (2008, p. 29) provide the following example:

“If confronted with the token *saw*, stemming might return just *s*, whereas lemmatization would attempt to return either *see* or *saw* depending on whether the use of the token was as a verb or a noun”.

The task is especially significant in highly inflected languages, such as Polish that has, for instance, more than 100 possible word forms for an adjective. Proper lemmatisation is a necessary prerequisite for almost any further textual analysis, even as basic as counting word frequencies.

Lemmatisation is performed by a selected morpho-syntactic tagger, one of the three possible configurations. Word forms are either assigned by a morphological analyser or guessed. Contextually appropriate lemmas are selected, and disambiguated lemmas are assigned by the selected tagger.

For each text file from the corpus, LEM returns its lemmatized version, in which each word is replaced by its lemma, i.e., different word forms of the same meaning are replaced by the same basic morphological form representing them. Due to the reduced complexity, such files can be used as the input to many statistical tools that work on the level of words and were originally designed for English.

For instance, for the input text (from the novel *Szczęśliwa* by Eliza Orzeszkowa):

Wysoka, kształtna, z twarzą myślącą, zimną nieco, ale pięknie zarysowaną i bardzo świeżą, w stroju pełnym smaku i powagi, siedzi pod rozłożystymi drzewami wspaniałego parku i myśli o tem, jaki ten park jest piękny, jaki ten dzień letni jest pogodny i jaka ona sama jest szczęśliwa.

LEM returns:

¹¹ <http://sgjp.pl/morfeusz/morfeusz.html.en>

¹² Nevertheless the existence of the two analysers and the necessity of the choice is an additional burden put on the user that should be removed in the future by exchanging the morpho-syntactic tagger with a new one, better compatible with *Morfeusz 2*, e.g. a new CLARIN-PL tagger MorphoDiTa-pl with a guessing module (<http://ws.clarin-pl.eu/morphoDiTa.shtml>) or even a better one (in development). This transition has not been fully completed yet, due to the lower efficiency of MorphoDiTa-pl in comparison to WCRFT and on-going works in CLARIN-PL on taggers that should be even better.

wysoki , kształtny , z twarz myśleć , zimny nieco , ale pięknie zarysować i bardzo świeży , w strój pełny smak i powaga , siedzieć pod rozłożystemi drzewo wspaniały park i myśleć o tema , jaki ten park być piękny , jaki ten dzień letni być pogodny i jaki on sam być szczęśliwy .

A couple of lemmatisation errors, that can be noticed (e.g. *rozłożystemi* instead of *rozłożysty*) are caused by the archaic word forms in the original text written in the XIXth century, while the tagger was trained on the contemporary texts and is using a contemporary morphological analyser.

3.2 Part of Speech Tagging

PoS tagging refers to both manual and automatic attribution of tokens to word classes. In a simplified version, PoS tagging is taught at school under the form of attribution of word classes such as nouns, verbs, adjectives etc. to given morphological forms. As in the case of lemmatization, successful tagging depends on the correct interpretation of morphologic and syntactic properties of a word, as the same word form may represent different classes, e.g. 'drink', depending on the context, can function as a noun and as a verb.

LEM encodes the PoS Tagging results with the tagset developed for the National Corpus of Polish (NCP, cf. Przepiórkowski et al., 2012). For each file in the corpus, LEM returns a CSV file with columns containing tokens, their lemmas and tags¹³. Table 1 contains an encoded excerpt from *Szczęśliwa* by Eliza Orzeszkowa.

WORD	LEMMA	PoS
Nie `not`	nie	qub
była `was`	być	praet
już `already`	już	qub
młoda `young`	młody	adj
,	,	interp
lecz `but`	lecz	conj
twarz `face`	twarz	subst
jej `her`	on	ppron3
zachowała `had kept`	zachować	praet
delikatność `delicacy`	delikatność	subst
rysów `of countenance`	rys	subst
i `and`	i	conj
cery `complexion`	cer	subst
,	,	interp
kibić `figure`	kibić	subst

Table 1. LEM Part-of-Speech tagging. English translations added by the authors.

¹³ Simplified tagset table is available here: <http://nkjp.pl/poliqarp/help/ense2.html>

3.3 Verb characteristics

LEM allows for the further exploration of the corpus through the verb analysis. Verb characteristics feature returns numeric data about the occurrences of the verbs depending on their tense, number, person and gender, using the grammatical categories from the NCP tagset. The resulting table is delivered in an XLSX¹⁴ file and contains the number of tokens and verbs in the corpus, together with aggregated counts for the following verb forms: infinitive; 1st, 2nd, 3rd person singular; 1st, 2nd, 3rd person plural. Table 2 contains sample results.

SINGULAR								PLURAL					
Tokens	Verbs	1Pers	2pers	3pers	3pers_s_m	3pers_f	3pers_n	1Pers_s	2pers_s	3pers_s	3pers_m	3pers_nm	inf
11242	1299	100	100	84	151	465	0	0	0	0	0	0	150

Table 2. LEM verb form characteristics for an excerpt from Eliza Orzeszkowa's novella *Kto winien*

3.4 Lemmas and PoS statistics

Basic descriptive statistics of lemmas and PoS occurrence in the corpus is aggregated on the basis of lemmatisation and PoS tagging, described in A and B above. Resulting tables, delivered in an XLSX spreadsheet, contain absolute counts of respective lemmas and tags, together with their share in the entire corpus. Tables 3 and 4 contain sample results.

<i>być</i> `to be'	201	1.7673%
<i>który</i> `which/who'	97	0.8529%
<i>to</i> `this'	92	0.8089%
<i>oko</i> `eye'	35	0.3077%
<i>ręka</i> `hand/an arm'	32	0.2814%
<i>czy</i> `whether'	31	0.2726%

Table 3. LEM lemma statistics for Eliza Orzeszkowa's novella *Kto winien*

interp	2607	22.9227%
qub	673	5.9175%
conj	497	4.3700%
adv	254	2.2334%
praet:sg:f:imperf	211	1.8553%
prep:gen:nwok	189	1.6618%
subst:sg:gen:f	172	1.5124%

¹⁴ The data here have a more complex form, and that is why we exchanged CSV to XLSX. This was done after we observed that our users had had a lot of problems with importing a CSV encoded in UTF-8 in a proper way into Microsoft Excel.

adv:pos	163	1.4332%
---------	-----	---------

Table 4. LEM frequencies of the top most frequent morpho-syntactic tags in Eliza Orzeszkowa's novella *Kto winien*

3.5 Named-Entity Recognition and Statistics

Corpora usually contain proper names, which may be relevant for a given research question (e.g. names of scholars quoted in a collection of academic papers). The Named Entity Recognition feature extracts Named Entities from the corpus and returns a sorted XSLX table with all proper names and their frequencies (See Table 5.).

NAMED ENTITY	LEMMA	FREQ
<i>Rzym</i>	<i>Rzym</i> `Rome'	19
<i>Palatynie</i>	<i>Palatyn</i> `Palatinus'	13
<i>Kapitolu</i>	<i>Kapitol</i> `Capitol'	7
<i>Forum</i>	<i>forum</i> `forum'	6
<i>Konstantyna</i>	<i>Konstantyn</i> `Constantine'	4
<i>Koloseum</i>	<i>Koloseum</i> `Colosseum'	3
<i>Piotra</i>	<i>Piotr</i> `Peter'	3
<i>Słońce</i>	<i>słońce</i> `Sun'	3
<i>Via Sacra</i>	<i>via sacrum</i>	3
<i>Baedeker</i>	<i>Baedeker</i>	2
<i>Grecji</i>	<i>Grecja</i> `Greece'	2
<i>Kastora</i>	<i>Kastor</i> `Castor'	2
<i>Marka Aureljusza</i>	<i>Marek aureljusza</i> `Marcus Aurelius'	2

Table 5. LEM Named-Entities recognized in Jerzy Żuławski's short story *Veneri et Romae*. English translations provided by the authors.

Proper names are extracted with *Liner2*, which was built through Machine Learning¹⁵ performed on *KPWr Corpus*, manually annotated with more than 28,000 proper name occurrences (Broda et al., 2012, Marcińczuk et al., 2016). *Liner2* can recognise the beginning and end of a proper name occurrence in text and also classify it semantically into up to 82 classes organised in a shallow hierarchy. *Liner2* expresses very high accuracy concerning the recognition of the proper name occurrences (above 90%) and very good accuracy of their classification. In the example above, one can notice that *Liner2* had problems with generating proper lemmas for multi-word proper names. Although the problem posed by proper names consisting of common words was almost solved by Marcińczuk (2017) and his solution will be implemented to LEM, recognition of multi-word proper names containing foreign words not listed in the dictionary still poses a challenge. The components of proper names do not provide clues for its internal morpho-syntactic structure, and it is hard to recognise components that should be inflected.

¹⁵ Conditional Random Fields algorithm was used to learn contextual probabilities (depending on the left and right contexts) of tokens starting, occurring in the middle and outside proper names. Contexts are represented by more than 30 features referring to, e.g., word forms, dictionaries of characteristic words and semantic properties of words acquired from a very large wordnet of Polish, namely plWordNet (<http://plwordnet.pwr.edu.pl>).

For instance, *Marka Aureljusza* in Table 5 is the genitive form and was not properly processed due to the old-fashioned orthography of the second component. The only solution here would be to first determine the nominative form of a proper name on the basis of the tagger output (or its most frequent form), and then recognise its inflected forms.

3.6 Disambiguated Word-Senses and their Relations

Word-sense disambiguation (WSD) is the task of identifying the meaning of a semantically ambiguous word used in the text. Successful disambiguation is necessary for a semantic analysis of a text, especially if the ambiguities are strictly semantic (not morpho-syntactic): e.g., the word “paper” can mean either the material, a scientific article, or a newspaper.

WSD in LEM is performed with *WoSeDon* tool, which identifies the most characteristic word senses for a given text by first mapping the tokens onto the *plWordNet*¹⁶ (Polish name: SłowoSieć; Maziarz et al., 2016), a very large and comprehensive wordnet, expanded with some other connected knowledge resources like *SUMO* ontology¹⁷ (Pease, 2011). WSD is automatically preceded with necessary lemmatisation and PoS tagging (see sections A and B above) and returns CSV files with columns containing the token, its lemma, PoS, and representation of its meaning in the form of a synset. A synset is a set of synonyms or near-synonyms, i.e. words that share the same selection of lexico-semantic relations (called constitutive relations) and, thus, possess exactly the same semantic description according to a given wordnet. Such words can be considered as semantically equivalent, and can be interchanged in certain linguistic contexts (cf. Maziarz et al., 2013).

WORD	LEMMA	PoS	plWordNet 3.0 SYNSET
<i>niespokojny</i>	<i>niespokojny</i> ‘uneasy’	adj	niespokojny.3(42:jak)
<i>sen</i>	<i>sen</i> ‘sleep’	subst	spoczynek.2(23:st), sen.1(23:st)
<i>jakieś</i>	<i>jakiś</i> ‘some’	adj	jakowyś.1(42:jak); który.1(42:jak) jaki.1(42:jak); jakiś.1(42:jak) jakowy.1(42:jak); któryś.2(42:jak)
<i>jednej</i>	<i>jeden</i> ‘one’	adj	pewien.1(42:jak) jeden.3(42:jak)
<i>nocy</i>	<i>noc</i> ‘night’	subst	noc.2(25:czas)
<i>jesiennej</i>	<i>jesienny</i> ‘autumnal’	adj	jesienny.1(43:rel)

Table 6. LEM WSD-based analysis of Jerzy Żuławski’s *Veneri et Romae*. English translations provided by the authors.

3.7 Hypernyms & hyponyms

This feature identifies hypernyms and hyponyms of the disambiguated word senses, described in the previous section. Information about hypernyms and hyponyms sheds more light on the senses identified in the previous step and may help in the interpretation by enriching text representation with micro semantic fields (e.g., hypernyms and hyponyms can be shared by some words in the corpus).

While synonymy, the basic relation in *plWordNet*, is the relation of semantic equivalence, hypernymy and hyponymy entail meanings which are, respectively, more general or more specific. For instance, the hypernym of ‘plant’ is ‘organism’, while ‘flower’ is its hyponym. In *plWordNet*, the relations of hypernymy/hyponymy are among constitutive relations that shape a hierarchical structure of the lexicon (Piasecki et al., 2009, Maziarz et al., 2013).

¹⁶ <http://plwordnet.pwr.wroc.pl/wordnet/>

¹⁷ This also opens a possibility of collecting statistics of the concepts matching the given text.

For each file in the corpus, the feature returns a CSV file similar to the one returned by the WSD (presented in the previous section), containing columns for the token, its lemma, PoS, synset (semantic representation) and the lemma's hypernym(s)¹⁸ and hyponym(s).

3.8 Stylometric analysis

Stylometric analysis is “the study of measurable features of (literary) style, such as sentence length, vocabulary richness and various frequencies (of words, word lengths, word forms, etc.)” (Eder and Rybicki, 2012). The best-known use of computational stylometric analysis is the authorship attribution, which provides “taxonomies of features to quantify the writing style, the so-called style markers, under different labels and criteria” (Stamatatos, 2009) to identify the author of a text. It also allows for discovering patterns in the corpus by grouping texts according to their stylistic features.

LEM provides a simplified interface to the CLARIN-PL WebSty¹⁹ and allows for various visualizations of its results. The detailed description of the tool could be found in (Eder et al., 2017).

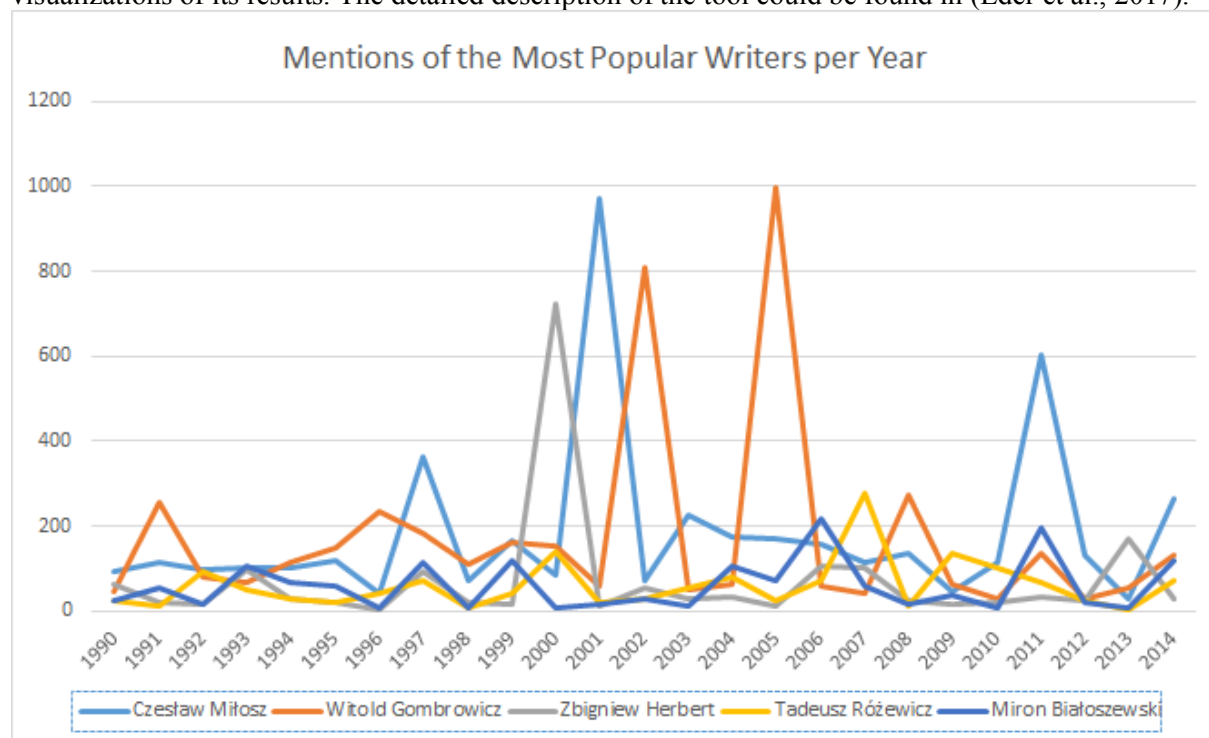


Figure 3. Patterns of interest in Polish writers based on “Teksty Drugie”.

3.9 Topic modelling

This feature allows to discover of abstract ‘topics’ that occur in a given corpus. LEM uses Latent Dirichlet Allocation (LDA) (Blei et al., 2003) method to generate word collections that are significant and specific for a given corpus on the basis of word frequency and the scope of their presence in samples. In order to control the overrepresentation of certain words in particular texts, LEM uses only lemmatized nouns that appeared at least in 80% of the documents (a fraction of the total corpus size). Users can choose the number of topics they wish to explore (the default is set at 20) and whether they want to split input files into chunks of roughly 20,000 bytes each, to ensure better processing of longer texts.

The results (i.e. words in particular topics and topic distribution across the text) could be explored through the interactive interface, or downloaded in XLSX or JSON format.

¹⁸ As hypernymy is defined in plWordNet in a linguistic way, there can be more than one hypernym for a word sense.

¹⁹ <http://ws.clarin-pl.eu/websty.shtm>

4 Use Case

LEM prototype was developed by the team working with a particular textual corpus of 2,553 Polish texts, published in *Teksty Drugie*,²⁰ an academic journal dedicated to literary studies. The corpus consists of the two parts: OCR-ed scans (1990 – 1998) and digital-born files (1999 – 2014). Given the aim of this paper (software presentation), we will treat the results only as examples of the method, without getting into too much detail. For a more extensive interpretation of the results of *Teksty Drugie* analysis with LEM see Maryl (2016a) and Maryl and Eder (2017).

The work on the prototype was divided into several stages, conceived as feedback loops for the developing team: on every stage, a new service was added to the application, and the test run was performed. After the analysis of its result, either the step was repeated, or the team moved to the next phase.

Phase 1. The OCR-ed corpus was cleaned (e.g., word breaks and headers were removed).

Phase 2. The corpus was lemmatized and PoS-tagged. Frequency lists were created, which enabled searching for patterns in the textual output. This allowed for a simple discovery of interest patterns in the journal over time. For instance, Figure 3. shows a pattern of interest in most popular Polish writers based on the number of times their names were mentioned per year (the figure contains only the authors who reached the threshold of 1,000 total mentions).

Another temporal insight to be gained from a lemmatised corpus concerns the uptake of critical approaches in literary studies. Figure 4 presents the reception of postcolonial studies in the Polish academia on the example of mentions of words related to “postcolonial” (i.e., *postkolonializm* ‘postcolonialism’, *postkolonialny* ‘postcolonial’, *postkolonialność* ‘postcolonialness’, *postkolonialista* ‘postcolonialist’) in a literary-studies journal.

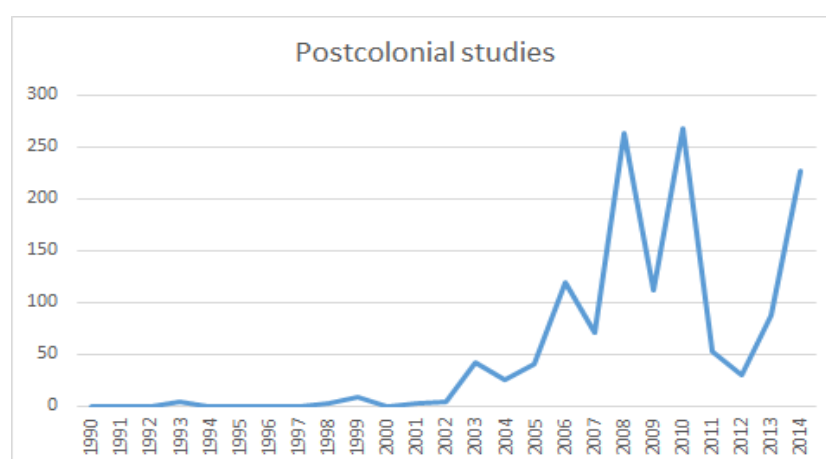


Figure 4. Mentions of words related to ‘postcolonial’ in “Teksty Drugie” per year.

Phase 3. The analysis of word frequencies, especially the ML-based training of classifiers revealed problems with the word list, e.g., occurrences of numbers, years and city names, which were preserved in the bibliographic references. This led to the addition of a user-defined stopwords list feature. The exclusion of corpus-specific problematic words and general function words (e.g., Polish words corresponding to ‘this’ *to*, *ta*, *ten*, ‘that’ *tamto*, *tamta*, *tamten*, ‘if’ *jeśli*, *jeżeli*) allowed for the visualisation of the most frequent words in *Teksty Drugie* (Fig. 5)

²⁰ <http://tekstydrugie.pl/en/>

5 Further Development

Currently, LEM's GUI is being continuously developed in cooperation with users: mostly literary scholars working on various types of texts (fiction, journal articles, blog posts), sociologists and social psychologists.

LEM prototype was fully implemented and made available as a web application²¹ to the scholarly audience working on texts in Polish. Next, it will be extended with tools for other languages (e.g. English and German), in a similar way to WebSty. Thanks to LEM's modular architecture, it will require mostly linking new processing Web Services and adding converters. LEM is available on an open licence, and the authors will be happy to share their tools, code and *know-how*. Export options to other formats will be added, so researchers can easily create the output in a particular format (list, text, table) and upload it to other applications for further processing.

References

- [Blei et al., 2003] Blei, D.M., Ng, A.Y. and Jordan, M.I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*. 3 (4–5): pp. 993–1022.
- [Broda et al., 2012] Broda, B., Marcińczuk, M., Maziarz, M., Radziszewski, A., and Wardyński, A. 2012. KPWr: Towards a Free Corpus of Polish. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugür Dogan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, May 2012*. European Language Resources Association (ELRA), pp. 3218–3222.
- [Bryda and Tomanek, 2015] Bryda G., Tomanek K. 2015. Odkrywanie wiedzy w wypowiedziach tekstowych. Metoda budowy słownika klasyfikacyjnego. In Niedbalski J. (Ed.) *Metody i techniki odkrywania wiedzy. Narzędzia CAQDAS w procesie analizy danych jakościowych*, Wydawnictwo UŁ, pp. 51–81.
- [Brosz et al., 2017] Brosz M., Bryda G., Siuda P. 2017. Od redaktorów: Big Data i CAQDAS a procedury badawcze w polu socjologii jakościowej. *Przegląd Socjologii Jakościowej* [Big Data, CAQDAS and research procedure in the field of qualitative research], t. 13, nr 2, pp. 6–23 [Access 30.01.2018, URL: www.przegladsocjologiijakosciowej.org].
- [Calle-Martin and Miranda-Garcia, 2012] Calle-Martin, J. and Miranda-Garcia, A. 2012. Stylometry and Authorship Attribution: Introduction to the Special Issue. *English Studies*, 3(93): 251–258.
- [Calzolari et al., 2014] Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S. (eds.) 2014. *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014. Reykjavik, Iceland*, ELRA.
- [Dallas et al., 2017] Dallas, C., Chatzidiakou, N., Benardou, A., Bender, M., Berra, A., Clivaz, C., ... Zebec, T. 2017. European Survey on Scholarly Practices and Digital Needs in the Arts and Humanities – Highlights Report. Zenodo.
- [Eder et al., 2017] Eder, M., Piasecki, M. and Walkowiak, T. 2017. An Open Stylometric System Based on Multilevel Text Analysis. *Cognitive Studies | Études cognitives*, No. 17, <https://doi.org/10.11649/cs.1430>
- [Eder and Rybicki, 2012] Eder, M. and Rybicki, J. 2012. Introduction to Stylometric Analysis using R. *Digital Humanities 2012 Conference. Hamburg*.
- [Jones et al., 2018] Jones E, Oliphant E, Peterson P, et al. 2018. SciPy: Open Source Scientific Tools for Python. <http://www.scipy.org/> [Online; accessed 2018-03-27].
- [Kędzia et al., 2015] Kędzia, P., Piasecki, M. and Orlńska, M. J. 2015. Word Sense Disambiguation Based on Large Scale Polish CLARIN Heterogeneous Lexical Resources. *Cognitive Studies | Études cognitives*, (15), 269–292.

²¹ <http://ws.clarin-pl.eu/lem.shtml>

- [Manning et al., 2008] Manning, C., Prabhakar, R. and Schütze, H. 2008. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.
- [Marcinićzuk, 2017] Marcinićzuk, M. 2017. Lemmatization of Multi-word Common Noun Phrases and Named Entities in Polish. In (ed.) Ruslan Mitkov and Galia Angelova *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017 Varna, Bulgaria, Sep 4–6 2017*, INCOMA Ltd., pp. 483–491, https://doi.org/10.26615/978-954-452-049-6_064
- [Marcinićzuk et al., 2016] Marcinićzuk, M., Oleksy, M., Maziarz, M., Wieczorek, J., Fikus, D., Turek, A., Wolski, M., Bernaś, T., Kocoń, J., Kędzia, P. 2016. Polish Corpus of Wrocław University of Technology 1.2, CLARIN-PL digital repository, <http://hdl.handle.net/11321/270>
- [Marcinićzuk et al., 2013] Marcinićzuk, M., Kocoń, J. and Janicki, M. 2013. Liner2 – A Customizable Framework for Proper Names Recognition for Polish. In Bembenik, Robert and Skonieczny, Lukasz and Rybinski, Henryk and Kryszkiewicz, Marzena and Niezgodka, Marek, *Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions*. Springer, vol. 467, pp. 231–253.
- [Marcinićzuk and Radziszewski, 2013] Marcinićzuk, M. & Radziszewski, A 2013. WCCL Match – A Language for Text Annotation. In Kłopotek, A., M., Koronacki, Jacek, Marciniak, Małgorzata et al (editors), *Language Processing and Intelligent Information Systems*, pages 131–144. Springer Berlin Heidelberg.
- [Maryl, 2016a] Maryl, M. 2016a. Tekstów świat. Przyczynek do makroanalitycznej monografii czasopisma literaturoznawczego [World of Texts. Take on a Macroanalytical Monograph of a Scholarly Journal] In Nasiłowska, A. & Łapiński, Z. (Eds.), *Projekt na daleką metę. Prace ofiarowane Ryszardowi Nyczowi*, Warszawa: Wyd. IBL, pp. 443–462.
- [Maryl, 2016b] Maryl, M. 2016b. Cyberwspólnota sądów żalu w perspektywie makroanalitycznej [Cybercommunity of regret statements in the macroanalytical perspective]. In *3rd Congress of the Polish Society for Cultural Studies, Adam Mickiewicz University of Poznań, 21–23 September 2016*.
- [Maziarz et al., 2016] Maziarz, M., Piasecki, M., Rudnicka, E., Szpakowicz, S., and Kędzia, P. 2016. plWordNet 3.0 – a Comprehensive Lexical-Semantic Resource. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, The COLING 2016 Organizing Committee pp. 2259–2268, 2016, <http://www.aclweb.org/anthology/C16-1213>
- [Maziarz et al., 2013] Maziarz, M., Piasecki, M., and Szpakowicz, S. 2013. The Chicken-and-egg Problem in Wordnet Design: Synonymy, Synsets and Constitutive Relations. *Language Resources and Evaluation*, 47(3):769–796.
- [McCallum, 2002] McCallum, A.K. 2002. *MALLET: A Machine Learning for Language Toolkit*. Web page of the system. URL: <http://mallet.cs.umass.edu>.
- [Pease, 2011] Pease, A. 2011. *Ontology: A Practical Guide*. Articulate Software Press, Angwin, CA,.
- [Pedregosa et al., 2011] Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, pp.2825–2830.
- [Piasecki et al., 2009] Piasecki, M., Szpakowicz, S. and Broda, B. 2009. *A WordNet from the Ground Up*. Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej, <http://www.dbc.wroc.pl/dlibra/docmetadata?id=4220&from=publication>
- [Piasecki and Walentynowicz, 2017] Piasecki, M. and Walentynowicz, W. 2017. MorphoDiTa-based Tagger Adapted to the Polish Language Technology. In Z. Vetulani and P. Paroubek, editors, *Proceedings of Human Language Technologies as a Challenge for Computer Science and Linguistics*, pp. 377–381, Poznań, 2017. Fundacja Uniwersytetu im. Adama Mickiewicza w Poznaniu.

- [Piasecki et al., 2016] Piasecki, M.; Walkowiak, T. & Eder, M. 2016. WebSty — an Open Web-based System for Exploring Stylometric Structures in Document Collections. In Eder, M. & Rybicki, J. (Eds.) *Digital Humanities 2016 Conference Abstracts*, Jagiellonian University and Pedagogical University, 2016, pp. 859–861.
- [Przepiórkowski et al., 2012] Przepiórkowski, A., Bańko, M., Górski, R. L. and Lewandowska-Tomaszczyk, B. (eds) 2012. *Narodowy Korpus Języka Polskiego*. Warszawa: PWN.
- [Radziszewski, 2013] Radziszewski, A. 2013. A Tiered CRF Tagger for Polish. In Bembenik, Robert and Skonieczny, Lukasz and Rybinski, Henryk and Kryszkiewicz, Marzena and Niezgodka, Marek, *Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions*. Berlin: Springer, vol. 467, pp. 215–230.
- [Rohnka et al., 2015] Rohnka, N., Szymczyk, B., Rusanowska, M., Holas, P., Krejtz, I., Nezlek, J. 2015. Właściwości języka osób cierpiących na zaburzenia emocjonalne i osobowości - analiza treści opisów codziennych wydarzeń [Language characteristics of individuals with emotional, and personality disorders: content analysis of daily events]. *Psychiatria i Psychoterapia*, Vol. 11, No. 3, pp. 3–20.
- [Rybicki, 2017] Rybicki, J. 2017. Reading Novels with Statistics: What Numbers of Words Tell Us about Authorship, Genre, or Chronology. In J. A. Dobelman (Ed.) *Models and Reality: Festschrift For James Robert Thompson*, Chicago: T&NO Company, pp. 207–224.
- [Saloni et al., 2015] Saloni, Z., Woliński, M. Wołosz, R., Gruszczyński, W., and Skowrońska, D. 2015. *Słownik gramatyczny języka polskiego*. [Grammatical dictionary of Polish]. SGJP, 3rd edition.
- [Stamatatos, 2009] Stamatatos, E. 2009. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, 3(60): 538–556.
- [Tausczik and Pennebaker, 2010] Tausczik, Y.R., and Pennebaker, J.W. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29, 24–54.
- [Witten et al., 2017] Witten, I.H., Frank, E., Hall, M.A., Pal, C.J. 2017. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufman.
- [Woliński, 2014] Woliński, M. 2014. Morfeusz Reloaded. In (Calzolari et al., 2014), pages 1106–1111.
- [Zhao and Karypis, 2005] Zhao, Y. and Karypis, G. 2005. Hierarchical Clustering Algorithms for Document Datasets. *Data Mining and Knowledge Discovery*, 10(2): 1.

Open Stylometric System WebSty: Towards Multilingual and Multipurpose Workbench

Maciej Piasecki

Faculty of Computer Science
and Management
Wrocław University of
Science and Technology
maciej.piasecki@pwr.edu.pl

Tomasz Walkowiak

Faculty of Electronics
Wrocław University of
Science and Technology
tomasz.walkowiak@pwr.edu.pl

Maciej Eder

Institute of Polish Language
Polish Academy of Sciences
and Pedagogical University
of Kraków
maciej.eder@ijp.pan.pl

Abstract

WebSty is an open, web-based stylometric system designed for Social Sciences & Humanities (SS&H) users. It was designed according to the CLARIN philosophy: no need for installation, minimised requirements on the users' technical skills and knowledge, and focus on SS&H tasks. In the paper, we present its latest extension with several visualisation methods, techniques for the extraction of characteristic features, and support for multilinguality.

1. Introduction

Stylometry is based on the analysis of language features extracted from texts and aimed at tracing similarities between texts. It is used to identify groups of texts that exhibit subtle similarities hidden to the naked eye but traceable by multidimensional statistical techniques. A classical type of such an analysis is authorship attribution or an experimental setup in which anonymous (or disputed) texts are compared against a set of texts of known authorship, to identify the nearest neighbour relations (Stamatatos 2009). In Social Sciences & Humanities (SS&H) text analysis is becoming an interesting methodological proposition to assess textual similarities beyond authorship. In the study of literature, one might be interested in distant reading techniques to pinpoint genre characteristics, literary period, intertextuality, etc. In sociology, one might want to analyse textual biases in press materials from different press agencies, in psychology one might trace a change of the style as a function of the authors' age or correlations between a text and mental diseases (Le et al. 2011).

An application of the stylometric methods can be difficult for SS&H researchers, mostly because the combination of the variety of data formats, language tools, and data analysis tools is not straightforward, but also an application of the tools usually requires specialised knowledge and technical skills. Moreover, the entire NLP workflow is controlled by a large number of hyperparameters whose influence on the overall results of the stylometric analysis is complex.

*WebSty*¹ is an open stylometric system with the web-based user interface designed to be used without any installation, and which offers a variety of dedicated language processing tools, provides ready to use processing chains, and assists users in setting up the processing parameters. It was initially focused on processing texts in Polish² and offered a limited number of the visualisation and data analysis methods (Eder et al. 2017). Below we present a new version which has been expanded with a more flexible and efficient processing architecture, several visualisation methods and techniques for the extraction of characteristic features. The modular architecture of WebSty enabled adding support for more languages, namely English, German, Russian, Hungarian and Spanish in a relatively easy way.

Stylometric techniques are based on converting text documents or fragments into vectors of numerical values and next on processing the resulting vectors by data analysis method. The goal is to find similarities in the input data. This is often achieved by applying clustering algorithms that divide the vectors into similarity classes, e.g., documents of the same author.

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

¹ <http://websty.clarin-pl.eu>

² <http://ws.clarin-pl.eu/websty.shtml>

The paper is structured as follows. First, related solutions are analysed. Next, the architecture of the language processing is presented. It is followed by a detail description of the data analysis and visualization techniques.

2. Related Work

In spite of the long tradition of stylometry, there is only a limited number of online systems. The well known *Voyant*³ (Sinclair & Rockwell, 2012) is an online tool for the limited statistical analysis of texts supplemented with a good GUI and several visualisation methods. A range of NLP tools was added on the basis of the Stanford CoreNLP (Manning et al., 2014), e.g., PN recognition. However, the functionality of Voyant is based mostly on tracing word forms and their relative frequencies across text and limited to English. Only simple statistical measures: tf.idf and Z-score are available to compare word forms vs. documents. Popular *Stylo* (Eder et al. 2016) is a library in the R programming language for different stylometric tasks. It is designed to analyse shallow morphological features (function words and letter n-grams) harvested from the locally stored plain text files, but it can also be used to analyse preprocessed corpora. The package offers both selected exploratory methods, and supervised Machine Learning (ML) algorithms. It needs to be locally installed. *Mallet* (McCallum, 2002) is an off-line document classification system working on the basis of machine learning, but it is mostly used for the topic analysis.

Also, we can find on the Web a couple of simple online applications for the authorship attribution⁴, e.g. *Signature* (only word-level features) and *AICBT* (limited number of feature types for English). There is a number of off-line applications, like *JGAAP* (an entire processing workflow), *JStylo* (McDonald et al., 2012) (rich set of feature types, recognition of obfuscation), and *StyleTool* (Maurer, 2017) (quite rudimentary). Neither of the discussed systems supports parallel processing of large amounts of data, nor they use multiple language tools and processing methods, and an advanced extraction of characteristic features.

3. Language Processing Architecture

A multi-user, web-based system generates problems related to the system availability and performance. The system should be *scalable*, *responsive* and *available* all the time. Language tools (LTs) have excessive CPU/memory consumption. Needless to say, the number of users and/or tasks at a given time is fairly unpredictable, which makes the allocation of resources even more problematic. WebSty architecture is presented in Fig. 1. It is based on a *service-oriented software* idea (Bell, 2010), that has gained great popularity, according to which each LT is implemented as a *microservice* (Wolff, 2016) and run as a separate process with the pre-loaded data models. The number of microservices run in parallel is limited by hardware. Each type of LT has its own queue. NLP microservice collects tasks from a given queue and sends back messages when the results are available.

The usage of microservices communicating via lightweight mechanisms solves the problems of a variety of programming languages used, and complexity of the tight integration. As the number of microservices run in parallel is limited by hardware, the queuing system is crucial for the system performance and effective scalability. The most required and most frequently used LT microservices have to be run in several instances, and the queuing system acts as a load balancer. AMQP⁵ protocol for the lightweight communication mechanisms and RabbitMQ broker are applied. An additional server grants access from the Internet and works as a proxy for the core system delivering REST API to WebSty. This allows for easy integration with almost any kind of applications. The exchange of data between microservices, i.e., input/output of the LT tools is done via a network file system. It makes the integration of new LT tools easier since they are mostly designed in a manner in which they expect a file at their input and produce files on their output.

³ Voyant: <http://docs.voyant-tools.org>, CoreNLP: <https://nlp.stanford.edu/software/>, Mallet: <http://mallet.cs.umass.edu>

⁴ Signature: <http://www.philocomp.net/humanities/signature.htm>, JGAAP: <https://github.com/evllabs/JGAAP>, JStylo: <https://psal.cs.drexel.edu/index.php/JStylo-Anonymouth>, StyleTool: <https://github.com/lnmaurer/StyleTool>

⁵ AMQP: <https://www.amqp.org>, RabbitMQ: <https://www.rabbitmq.com>

The most common approach to the communication with web applications, namely REST API (REST), is synchronous. A thread producing the request is blocked until the response has not been returned to the client. In the case of requests that can be served in short response time, it is a very useful solution. However, when the response time is increasing, this can cause errors. First of all, the number of threads on the server side is limited, so the increasing response time could result in approaching this limit. Secondly, a response longer than the HTTP client timeout (usually equal to 189 s) causes the timeout limit error on the client side and breaks the connection. Thus, the receiving of the results fails (Walkowiak, 2014). That is why we use asynchronous way of communication in a polling-like way in WebSty. The client (written in JavaScript) keeps checking, whether the server has already finished the processing. To keep the user informed about the processing status, HTTP API provides the information concerning the advancement of processing in percentages. The other problem with REST API can be caused by a large input data volume. To prevent this, we implemented different methods for uploading corpora. WebSty allows for loading text files one by one, in one ZIP file (for texts a ZIP file is usually ten times smaller than original documents). In the case of really very large corpora, it is possible to download them first into a CLARIN-PL Cloud storage⁶ and next to inform WebSty (via REST API) about the corpus name in the cloud.

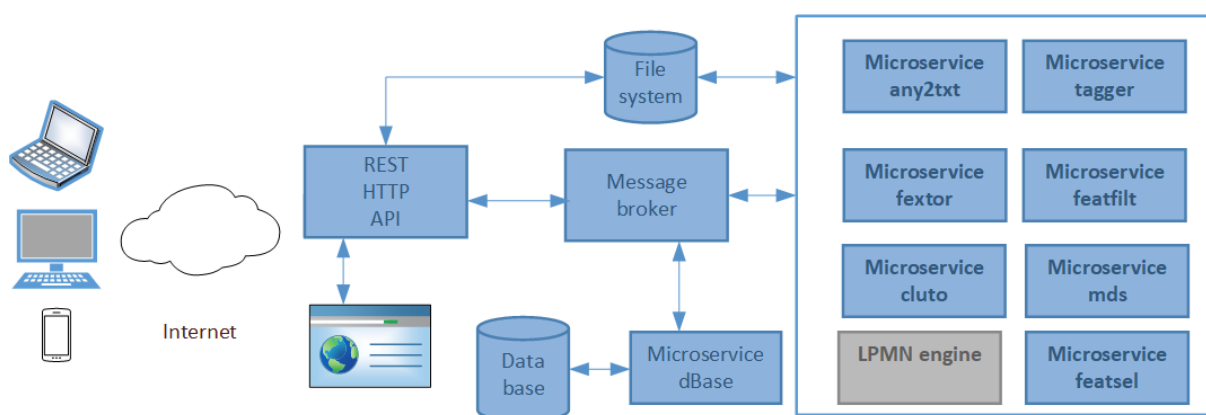


Figure 1. Language processing architecture

To achieve high availability, the system was deployed on a scalable hardware and software architecture that forms a private cloud (ten Blade Servers, connected by the fast fiber channel with highly scalable midrange virtual storage designed to consolidate workloads into a single system for the simplicity of management). XENServer controls each machine and forms a private cloud. Each frequently used NLP microservice is deployed on a separate virtual machine (Walkowiak, 2016).

WebSty requires to run NLP and ML tools in a specific order. Very often this is not a simple chain of tools, but a workflow of tools (Walkowiak, 2017). Therefore, we developed (Walkowiak, 2018) a human-readable orchestration (Peltz, 2003) language, that allows for the description of the WebSty processing tasks. It is called *Language Processing Modelling Notation* (LPMN). An exemplar of the LPMN statement for WebSty is presented in Fig. 2. Each microservice is defined by its name (for example: *any2txt*, *tagger*, *fextor* in Fig. 2). Data can be loaded from different sources, which are defined by dedicated LPMN statements. For example *urlzip* in Fig. 2 defines the URL of a ZIP file with input corpus. A dedicated microservice (LPMN engine in Fig. 1) is used to process LPMN tasks and acts as an orchestrator for other microservices.

The LPMN engine processes tasks of different sizes and computational complexity. Therefore, it is needed to prevent every tasks from blocking the small ones. To achieve this, we added a dedicated scheduling algorithm that prevents large files and large corpora (including a large number of files) from blocking the NLP microservice queues in the message broker. The engine checks the queue size, and if it has exceeded a predefined threshold (different for large files and for large corpora), the processing (sending tasks to the queue) is delayed for a given amount of time. As a result, simple tasks (for example: processing of one, small text file) are processed by LPMN engine in the time shorter than 6 s,

⁶ <http://nextcloud.clarin-pl.eu>

even if the LTC is busy with processing very large corpora. The experiments, showed that the delay caused by the scheduling algorithm is smaller than 1% of overall processing time.

The LPMN engine registers tasks in the database (see Fig. 1). The recorded information allows for collecting statistics about the WebSty usage (see also the conclusions).

```
urlzip("http://ws.clarin-pl.eu/public/teksty/2mini.zip")|any2txt|div(20000)|tagger({"lang":"polish"})|fextor2({"features":"base interp_signs bigrams","base_modification":"startlist","orth_modification":"startlist","lang":"ud","filters":{"base":[{"type":"lemma_stoplist","args":{"stoplist":"@resources/fextor/ml/polish_base_startlist.txt"}}]})|dir|out("output_fextor")|featfilt({"similarity":"cosine","weighting":"all:tf","filter":"min_tf-1 min_df-1"})|cluto({"no_clusters":2,"analysis_type":"plottree"})
```

Figure 2. Exemplar LPMN for WebSty analysis

4. Data Analysis

4.1 Basic options

WebSty development follows the CLARIN recommendations to make language tools available on Internet and develop research web based applications (Wittenburg et al., 2010), as a way to the elimination of the problems caused by the necessity of installing language tools (LTs) and possessing the required computational power. WebSty allows users to process data online without a need to bother about technicalities. However, some level of understanding the processing mechanisms is required to fully operate the application on the level of its user interface.

The user interface has been developed in HTML5 and JavaScript technology, using REST web service to run and control the language processing workflow on the server side. Firstly, the user must select the language of the text to be analysed (Fig. 3). Next, the number of groups into which the input corpora will be divided by the assumed clustering algorithm (see Sec. 4.6). Moreover, the input texts can be automatically divided into smaller parts of the approximately equal size set up by the user.

The key issues in the stylometric analysis are: definitions of features for the description of texts and methods for their further processing. WebSty offers a large set of features and weighting methods (presented in Sec. 4.3 and 4.5). Therefore, the four predefined sets of features and weightings are available for users, namely for the analysis of: Authorship, Grammatical style, Content and Classical Authorship (the last one is based on the most frequent words only). By selecting each of them the appropriate set of features and weighting methods is automatically set up for processing.

The screenshot shows the 'Options' page of the WebSty application. It is divided into two columns: 'Basic options' and 'Initial settings'. In the 'Basic options' column, there are three settings: 'LANGUAGE' with a dropdown menu showing 'pl', 'NUMBER OF GROUPS' with a text input field showing '2', and 'DIVISION OF INPUT FILE INTO CHUNKS OF SIZE (CHARS)' which is checked with a green box and has a text input field showing '20000'. In the 'Initial settings' column, there is a 'CHOICE OF FEATURES' dropdown menu showing 'Authorship'.

Figure 3. Basic WebSty options

4.2 Input data

Documents can be uploaded in many formats, e.g. MS Word, PDF, plain texts etc. The format of each document is automatically detected and the text content extracted. For larger data sets, a connection

between WebSty and the D-Space-based⁷ public repository of CLARIN-PL was built: data sets deposited in the repository can be selected for processing in WebSty. However, due to the users' demands, it is also possible to upload documents from a ZIP file identified by its URL or ZIP files that are locally stored on the user's disk (the last option is limited concerning both the data volume and the number of files). In the case of very large corpora or private corpora (that are less convenient to be stored in the repository), users can also use the CLARIN Cloud – a NextCloud-based⁸ private storage provided by CLARIN-PL. To use it, the user has to first log into the CLARIN-PL single authorization system (Pol, et al. 2018), and next he can select files from the storage (Fig. 5). It is worth to emphasize that logging into CLARIN-PL is necessary only for accessing the CLARIN Cloud and some restricted resources in the repository, WebSty as a system is completely open.

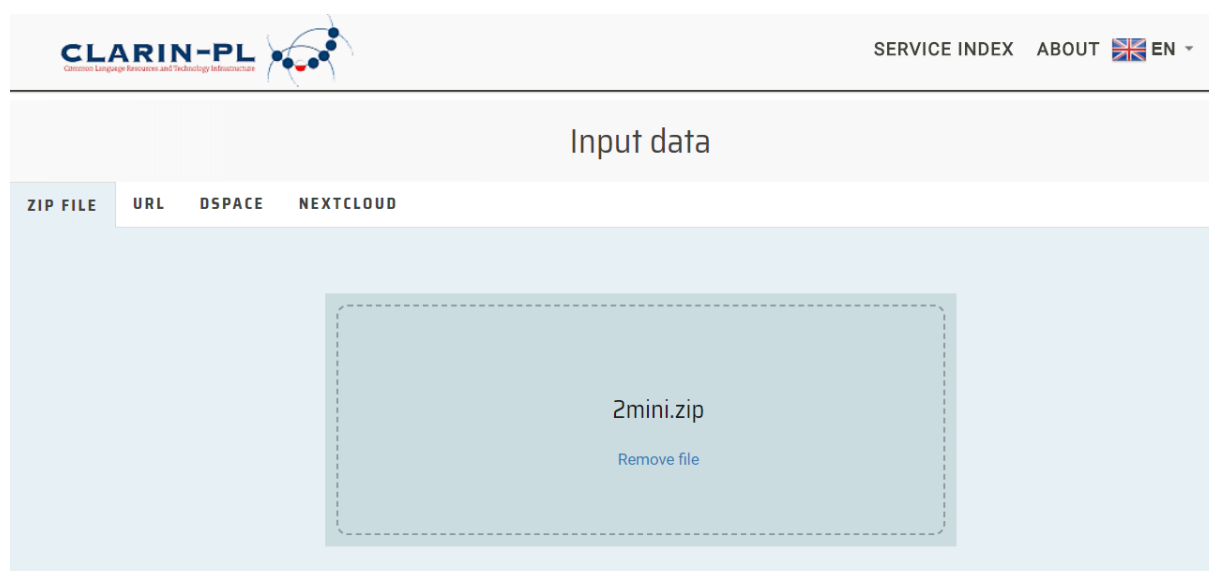


Figure 4. ZIP file input data interface

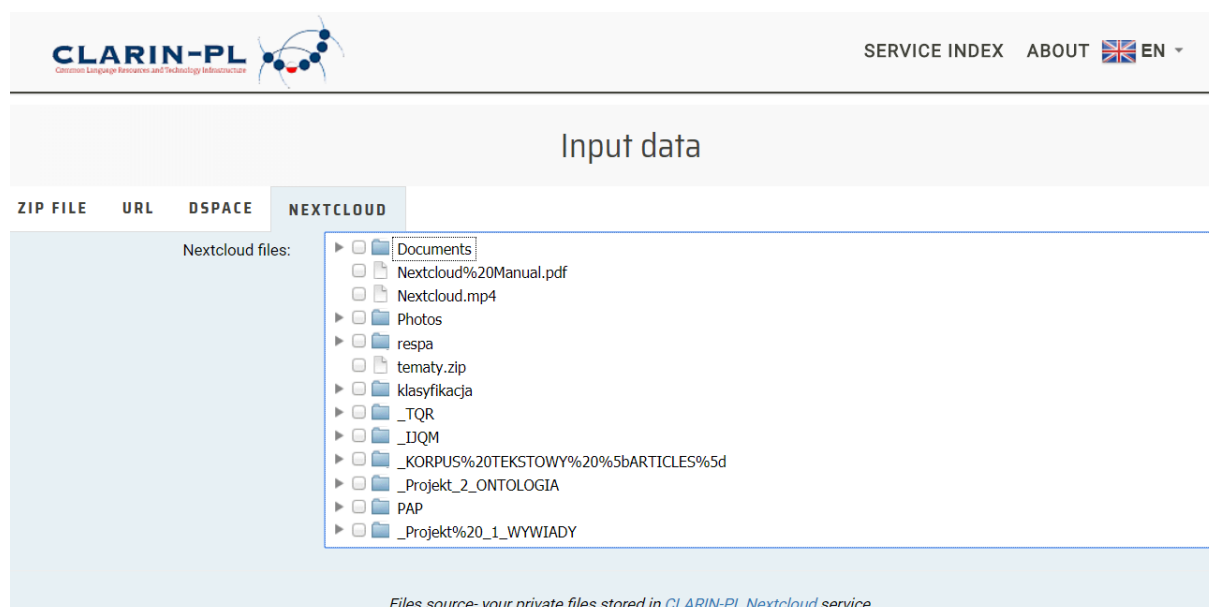


Figure 5. Selecting input data from CLARIN Cloud research storage

⁷ <https://clarin-pl.eu/dspace/>

⁸ <https://nextcloud.clarin-pl.eu>

Elements

☐ LEMMAS

☒ WORD FORMS

Punctuation

☐ ALL MARKS ☐ SELECTED MARKS

Word classes

☐ VERBS ☐ ADJECTIVES ☐ PREPOSITIONS

☐ NOUNS ☐ ADVERBS ☐ PROPER NOUNS

Figure 6. Feature selection panel

4.3 Document features

Text documents or their fragments (if the automatic division option has been switched on, see Fig. 3) are first converted into *feature* vectors of numerical values, which are next filtered, transformed and finally processed by the various data analysis methods. The ultimate goal is to group the vectors into similarity classes by the clustering algorithms (i.e. an unsupervised approach).

The features are defined on the basis of the characteristic elements of the text linguistic structure and also of the document. The initial values of the features are frequencies of these elements. Next processing techniques are used to clean the vectors from noise and finally to compare them.

The **features** should reveal properties of a text that are characteristic for its author and his style, and they should not be correlated with its semantic content. A feature can refer to any level of the language analysis but should be based only on LTs that express a relatively small error. WebSty (Fig. 6), implements features based on the frequency of: word forms (words in text), punctuation, lemmas (basic morphological forms), grammatical classes (a rich tagset), Parts of Speech (as sets of grammatical classes), grammatical categories, bigrams and trigrams of grammatical classes, and semantic types of proper names. The lemmas and grammatical classes are obtained by the application of a morpho-syntactic tagger, occurrences and types of proper names come from a Named Entity Recogniser.

4.4 Multilinguality

WebSty was developed for Polish (Eder et al. 2017), and, initially, all grammatical features were based on the Polish National Corpus tagset⁹ (Przepiórkowski et al., 2012). Next, it has been expanded with support for English¹⁰. For this, we used spaCy¹¹ package for PoS tagging and lemmatization (Honnibal, Johnson, 2015). This process was continued by converting WebSty architecture into a **multilingual system**. It was quite simple, because even in the original monolingual version of the system the phases of text preprocessing and feature extraction were separated and performed by the two different modules. In the multilingual version, a third module was introduced in between, as it has been assumed that the annotated text delivered on the input of the feature extraction module has the same annotation format

⁹ <http://nkip.pl/poligarp/help/ense2.html>

¹⁰ <http://ws.clarin-pl.eu/webstyen.shtml?en>

¹¹ <https://spacy.io>

independently from the language, see below. So the task of the third added module – a transformation module – is to convert the automatically annotated text obtained from the preprocessing into the format expected by the extraction module.

As a result, the contemporary version of WebSty is capable of analysing texts in Polish, English, German, Russian, Hungarian and Spanish. The extension depends on the existing taggers and Named Entity Recognizers for the supported language. Due to the different tagsets used by the taggers, we selected Universal Tagset¹² (Petrov et al., 2012) as the input format for the feature extraction. The transformation module converts the original tagset of a tagger (a native format) into Universal Tagset. As the size of the latter is limited, it expresses quite coarse-grained classification (especially in comparison to the PNC format), so the conversion is a lossy process which introduces some generalisation¹³.

As a default tagger we use UDPipe¹⁴ (Straka & Straková, 2017) which has models trained for a large set of languages. In fact, we initially hoped to use UDPipe as the only and universal tool for preprocessing. However, the accuracy of tagging and lemmatization is significantly smaller than the one expressed by solutions dedicated to the individual languages. That is why, finally, whenever possible we use taggers dedicated for different languages and of better accuracy. For instance, for Polish WCRFT tagger (Radziszewski, 2013) with a converter from the NCP to Universal Tagset was selected and PurePos¹⁵ tagger for Hungarian (Orosz & Novák, 2013).

4.5 Filtering, weighting and similarity

The features which are suspected to introduce too much noise or not be relevant to the goal of the analysis, can be **filtered** out on the basis of: their raw value (e.g., minimal number of documents), weighted value (after preprocessing) or their type (e.g., specified lemmas, grammatical classes, bigrams, etc.).

Raw frequencies are often skewed, e.g., by the document length, document content, or by the general properties of a given very frequent lemma. WebSty (Fig. 7) offers several **weighting methods**: *tf* (normalised text frequency), *tf.idf*, vector normalisation, PMI (Pointwise Mutual Information) simple and discounted, and *tscore*. As the number of features can be very high, a few dimensionality reduction techniques were included: *SVD* (Singular Value Decomposition), *LSA* (Latent Semantic Analysis) (Landauer & Dumais, 1997) and Random projection.

Similarity calculation ^

Filtering method		Determining probability	
WITH SMALLER NUMBER OF OCCURRENCE THAN	<input type="text" value="4"/>	FEATURE WEIGHING METHOD	<input type="text" value="mi"/> ▼
OCCURRING WITH NUMBER OF DOCUMENTS THAN	<input type="text" value="21"/>	DIMENSION REDUCTION METHOD	<input type="text" value="none"/> ▼ <input type="text" value="10"/>
		SIMILARITY/DISTANCE MEASURE	<input type="text" value="Jaccard"/> ▼

Figure 7. Filtering, weighting and similarity interface

Text similarity is computed from transformed vectors by several measures: *cosine*, *Dice*, *Jacquard*, *ratio* (a heuristics measuring the average ratio of commonality), *shd* (a heuristics measuring the

¹² <http://universaldependencies.org/u/pos/>

¹³ Following an interesting suggestion of one of the reviewers, we can consider whether it is possible to process and compare texts in several languages in the same corpus, at the same time. Technically this is possible if only we limit the text representation only to the grammatical features. However, it should be noted that as stylometric methods can also be used to identify the source language of the translated texts, so we can expect that texts would be first clustered into groups corresponding to their languages.

¹⁴ <https://github.com/ufal/udpipe/>

¹⁵ <https://github.com/ppke-nlpg/purepos>

precision of mutual rendering of the two vectors). Several data analysis algorithms are based on the distance measure between vectors:

- *Manhattan, Canberra, Euclidean*,
- *Simple* (L1 on vectors normalised by a square root function) (Eder, 2016)
- *Burrows's Delta*,
- *Argamon* (Euclidean distance combined with Z-score normalisation),
- and *Eder's delta* (Eder, 2016).

WebSty also provides psychologically motivated conversion of similarity to a distance measure by arc cosine function.

4.7 Clustering

For **clustering** vectors, the combined *agglomerative-flat* clustering method from Cluto (Zhao & Karypis, 2005) was selected as it merges two perspectives: a pairwise hierarchy of similarity and a flat division into a predefined, expected number of clusters. The clustering is controlled by the three parameters: a number of clusters, a similarity measure and a clustering criterion function. A proper selection of the criterion function is a complicated issue. Thus, some ready to use defaults are provided in WebSty.

5. Data Visualisation and Exploration

The data analysis process produces (Fig. 8): similarity/distance values between texts (documents or parts) represented by a 2D matrix and the created clusters. The latter can be downloaded as an XLSX file or presented in a graphical form as a **dynamic dendrogram** – an interactive binary tree, whose nodes (individual texts or subtrees) can be collapsed or unfolded (JavaScript and *D3.js*¹⁶ library). The **similarity results** are presented as: a *heatmap* (a matrix showing similarity by colours, see Fig. 9) and a *schemaball*. In the schemaball plot (Fig. 10) the user can select a file name and analyse the similarity of texts by presented connections, their colour, and thickness. For larger text collections, a multidimensional scaling can be very helpful to visualise a set of multidimensional vectors in 2D or 3D space (interactive presentation). WebSty offers four methods of the multidimensional scaling:

- *metric* – preserving distances,
- *non-metric* – preserving orders in distances, *t-distributed Stochastic Neighbor Embedding* (Maaten et al., 2008) – preserving similarities,
- and *spectral embedding* (Belkin et al., 2003) – preserving the local neighbourhood.

Results after scaling are presented (Fig. 11) as points in the 2D space (*3D.js library*) or in the 3D space. The interactive 3D plot utilises the *three.js* library, based on *WebGL* and using the user's graphic card 3D acceleration.

In response to the frequent users' questions: what features are responsible for determining a given cluster, we added a module for the **selection of important features** (Fig. 12). It is based on a set of statistical and ML-based methods. It also assumes that enough training data is provided. The implemented methods include: *statistical tests* (for example Mann-Whitney), *information metrics* (for example InfoGain), *recursive feature elimination* using supervised classifiers (like Naive Bayes) and *feature importance* assessment implemented by the tree-based classifiers, e.g. Random Forest.

¹⁶ D3.js: <https://d3js.org/>, Three: <https://threejs.org/>, WebGL: <http://www.khronos.org/registry/webgl/specs/latest/>

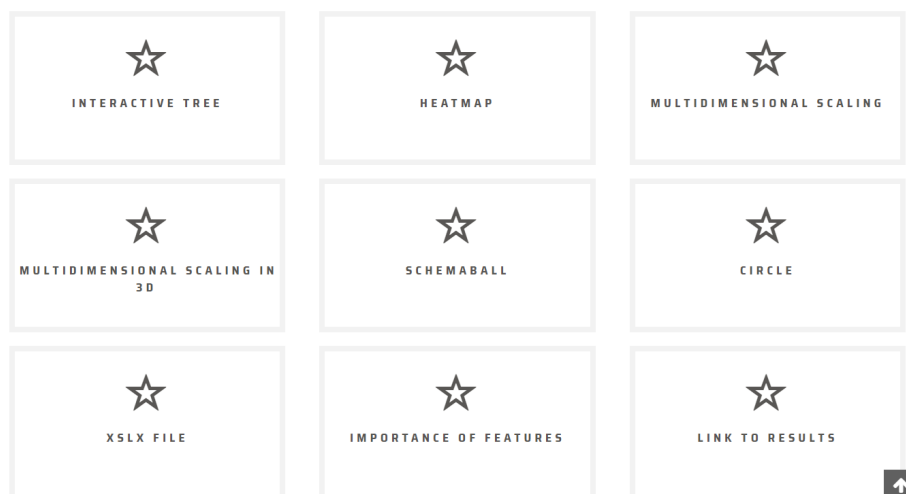


Figure 8. WebSty results interface

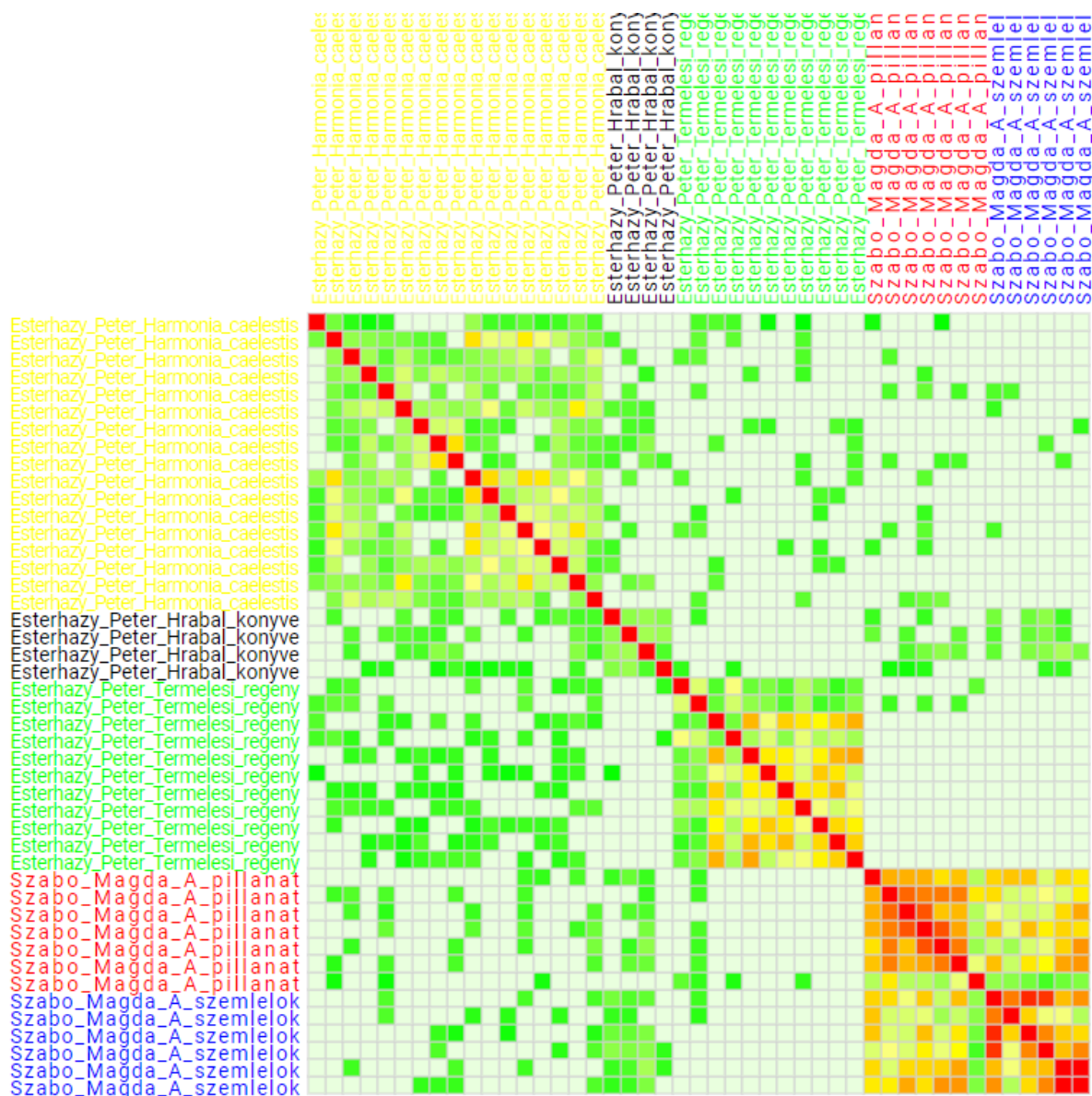


Figure 9. Similarity results in the form of a heatmap (5 books of 2 Hungarian authors, divided into chunks of size 20kB, coloured according to the book titles)

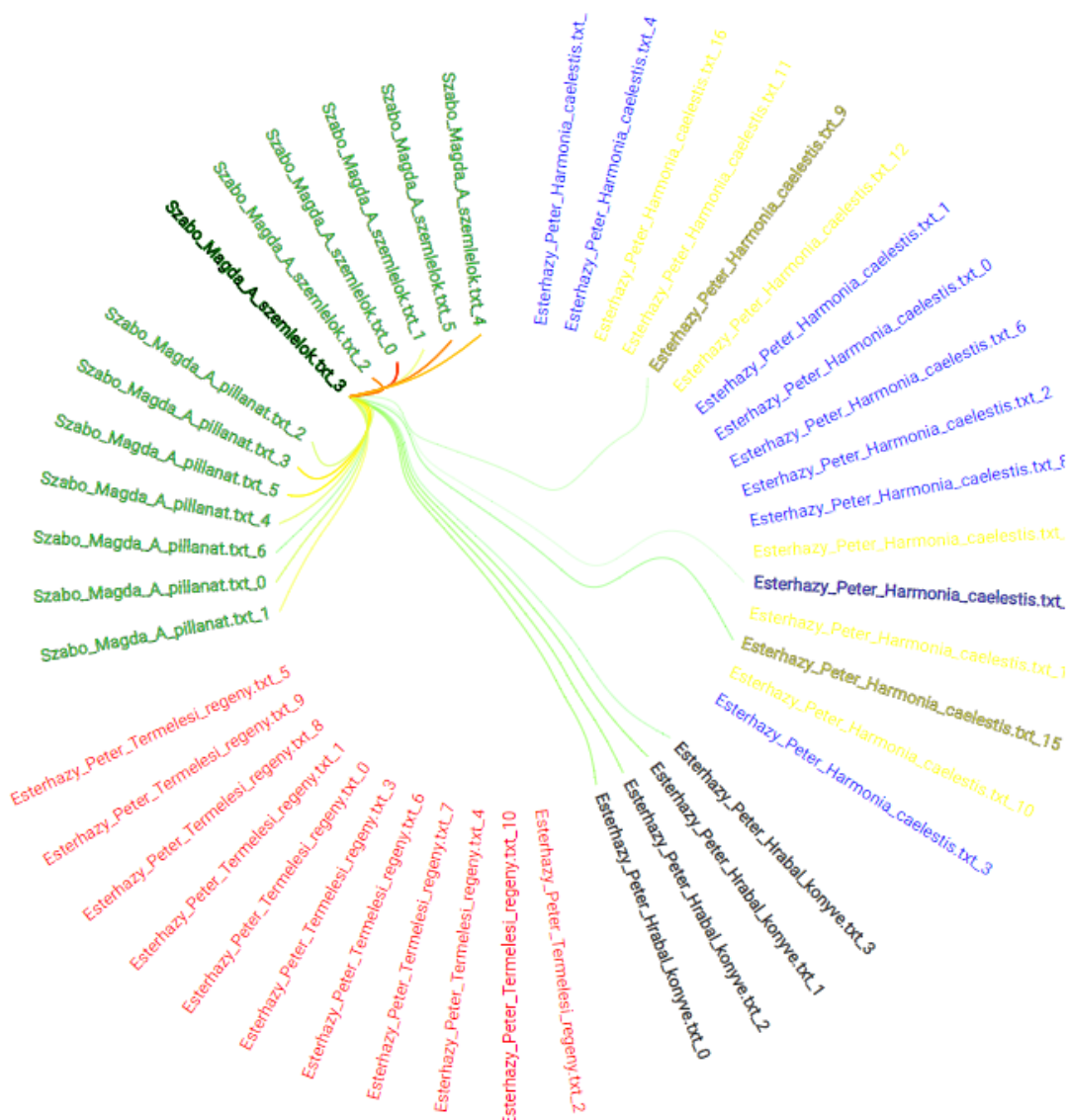


Figure 10. Similarity results in the form of a schemaball (the same corpus as in Fig. 9)

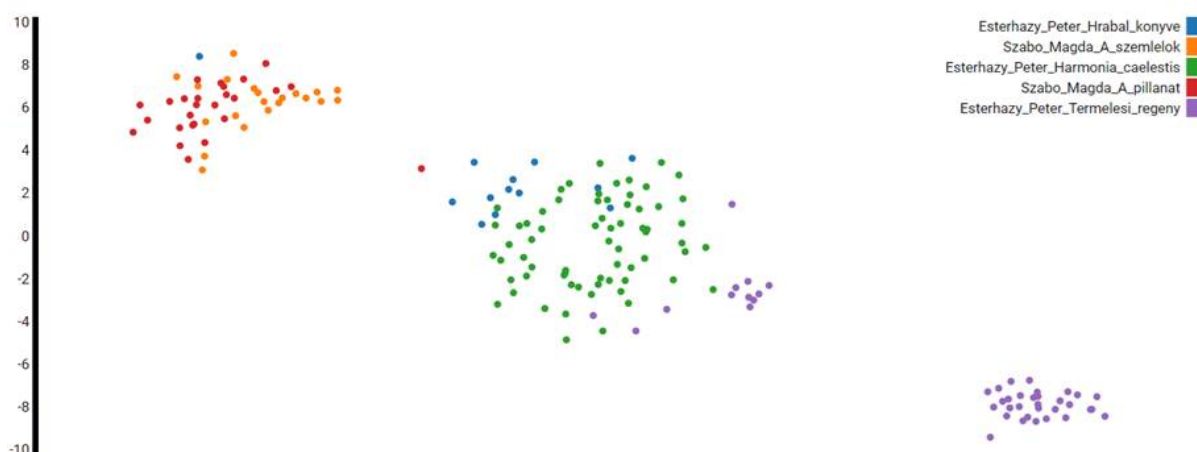


Figure 11. Distance results in the form of 2D plot (5 books of 2 Hungarian authors, divided into chunks of size 20kB, visualised using t-distributed Stochastic Neighbor Embedding)

Example results of the selection of the important features for the corpora of 5 books of 2 Hungarian authors (*Peter Esterhazy* and *Magda Szabo*), divided into chunks of the size 80kB) by the Mann-Whitney statistical test are presented in Tab. 1.

Feature	p-value	Mean in	Std in	Mean out	Std outside
base:csak	6.35E-10	0.008345	0.030443165	0.425654	0.102972377
base:míg	9.13E-10	0.02729	0.087679375	0.725347	0.241345316
bigrams:SCONJ_PROPN	2.62E-09	0.019812	0.047598872	0.557377	0.284573009
base:mikor	3.05E-09	0.007198	0.032770479	0.780015	0.378869513
bigrams:SCONJ_VERB	3.67E-09	0.017319	0.045162571	0.370154	0.15555791
base:aki	6.01E-09	0.015045	0.039769617	0.430457	0.174866081
base:ha	1.9E-08	0.033247	0.0841063	0.340515	0.127223237
base:közül	2.86E-08	0.013428	0.064323646	0.754823	0.4528463
bigrams:PRON_PROPN	4.79E-08	0.022123	0.093301757	0.505833	0.305402958
base:fel	5.34E-08	0.059155	0.111206063	0.609583	0.27429157
base:vele	6.19E-08	0.021742	0.073760873	0.527269	0.316427109

Table 1 Most important features that differentiate *Peter Esterhazy* from *Magda Szabo* (identified by the Mann-Whitney test)

6. Conclusions and further development

WebSty was implemented as a part of the CLARIN-PL infrastructure and made publicly available. Some features, like the connection to the repository and the cloud storage, are dedicated to the CLARIN-PL users. WebSty (different versions) has been already applied to several research tasks from the area of SS&H, as well as used in teaching. Among its applications, it is worth mentioning the literary analysis of the styles of web blogs (Maryl et al., 2016).

Automatic selection of features

Used tools ▼

Options

SELECTION METHOD

Statistical tests ▼

METHOD

Mann-Whitney ▼

NUMBER OF FEATURES

100

GROUPING METHOD

last level ▼

ⓘ

ANALYZE

Figure 12. Importance of features interface

During the last seven months (from September 2017 to March 2018) WebSty processed ca. 2900 sets of text documents. This was 9.8 GB of texts in total, and more than 840,000 files. The largest data set consisted of ca 19,000 files and the largest one concerning its size included ca. 170 MB.

Due to the asynchronous communication and the scalable architecture based on microservices, WebSty allows for processing corpora of really large sizes. There are three factors that limit the size of an input dataset. First of all, the maximum file size (in practice 2GB of zipped text corpus) that can be uploaded to CLARIN-PL file repository¹⁷. Secondly, the patience of a user can determine the maximal processing time. Up till now the longest processing in WebSty was equal to 18 hours. Thirdly, the memory required to keep raw frequencies matrix. This is the most limiting factor. Now, we have assigned 64 GB of memory to virtual machines that host microservice instances responsible for processing raw frequencies matrix. The size of such a matrix depends on the number of files in the given dataset and the number of features used for the documents. That is why the most extensive feature types (all word forms or all word lemmas) are not available in the GUI. In the case of choosing recommended features (selected by default in GUI) the limit is ca. 500,000 files in one corpus.

So far, a proper usability evaluation has not been performed for WebSty. Its fast and continuous evolution was one of the reasons for this. However, only during the last year, WebSty was used during three large CLARIN-PL training workshop by more than 100 users (the vast majority of them were scholars from SS&H) under the careful observation of the WebSty developers. The experience collected during such classes was the main source of inspiration for the changes and expansions introduced. Nevertheless, proper usability evaluation is planned to be conducted.

WebSty has been so far focused on unsupervised processing by clustering. We are working on an extended version offering support for applications of the supervised approach in which classifiers are trained by ML on the basis of manually annotated data sets, e.g., sets of texts annotated with the authors' names. We also working on extending WebSty to a system enabling unsupervised and supervised semantic analysis of text data sets, e.g., identification of text fragments that are related to situations of specific types or specific phenomena. Topic analysis will also be included into WebSty as a tool of the double roles: a tool by itself and also as a tool for data preprocessing.

Bibliography

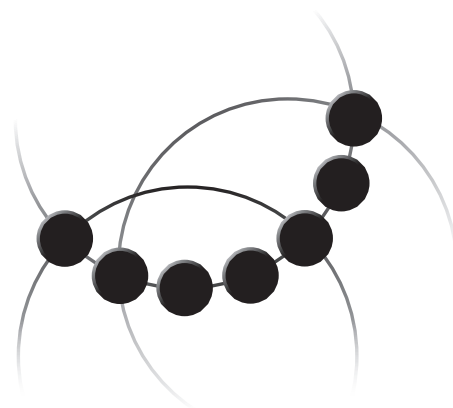
[Bell, 2010] Bell, M. (2010). *SOA Modeling Patterns for Service-Oriented Discovery and Analysis*. Wiley & Sons

¹⁷ <http://nextcloud.clarin-pl.eu>

- [Belkin & Niyogi, 2003] Belkin, M., Niyogi, P. (2003). Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, 15(6): 1373–1396.
- [Eder et al., 2017] Eder, M., Piasecki, M. and Walkowiak, T. (2017). An open stylometric system based on multilevel text analysis. *Cognitive Studies | Études cognitives*, 2017(17), <https://doi.org/10.11649/cs.1430>.
- [Eder et al., 2016] Eder, M., Rybicki, J. and Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. *R Journal*, 8(1): 107–121, <http://journal.r-project.org/archive/2016-1/eder-rybicki-kestemont.pdf>.
- [Honnibal & Johnson, 2015] Honnibal, M. and Johnson, M. (2015). An Improved Non-monotonic Transition System for Dependency Parsing. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal, 1373-1378
- [Landauer & Dumais, 1997] Landauer, T. and Dumais, S. (1997) A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition. *Psychological Review*, 1997, 104, pp. 211-240.
- [Le et al., 2011] Le, X., Lancashire, I., Hirst, G. and Jokel, R. (2011). Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists. *Literary and Linguistic Computing*, 26(4): 435–461.
- [van der Maaten & Hinton, 2008] van der Maaten, L.J.P.; Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9 (Nov), pp.: 2431–2456.
- [Maryl et al., 2016] Maryl, M., Piasecki, M. & Młynarczyk, K. (2016) Where Close and Distant Readings Meet: Text Clustering Methods in Literary Analysis of Weblog Genres. In Eder, M. & Rybicki, J. (Eds.) *Digital Humanities 2016 Conference Abstracts*, Jagiellonian University and Pedagogical University, pp. 273-275.
- [Maurer, 2017] Maurer, Leon (access Apr. 2017) Web page of the StyleTool program URL: <https://github.com/lnmaurer/StyleTool>
- [McCallum, 2002] McCallum, A.K. (2002) MALLET: A Machine Learning for Language Toolkit. Web page of the system. URL: <http://mallet.cs.umass.edu>.
- [McDonald et al., 2012] McDonald, A., Afroz, S., Caliskan, A., Stolerman, A. and Greenstadt, R. (2012) Use Fewer Instances of the Letter "i": Toward Writing Style Anonymization. *PETS 2012*
- [Manning et al., 2014] Manning, Ch. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J. and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Association for Computational Linguistics (ACL) 2014 – System Demonstrations*, ACL.
- [Orosz & Novák, 2013] Orosz, G. and Novák, A. (2013) PurePos 2.0: a Hybrid Tool for Morphological Disambiguation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2013)*, page 539–545, Hissar, Bulgaria, 2013. INCOMA Ltd. Shoumen, BULGARIA.
- [Peltz, 2003] Peltz, Ch. (2003). Web services orchestration and choreography. *Computer*, vol. 36, no. 10, pp. 46–52
- [Petrov et al., 2012] Petrov, S., Das, D., & McDonald, R. (2012) A Universal Part-of-Speech Tagset. In *Proceedings of LREC 2012*.
- [Pol et al., 2018] Pol M., Walkowiak T., Piasecki M. (2018). Towards CLARIN-PL LTC Digital Research Platform for: Depositing, Processing, Analyzing and Visualizing Language Data. In *Reliability and Statistics in Transportation and Communication*. Lecture Notes in Networks and System, Springer International Publishing, vol. 33.
- [Przepiórkowski et al., 2012] Przepiórkowski, A., Bańko, M., Górski, R. L. and Lewandowska-Tomaszczyk, B. (eds) (2012). *Narodowy Korpus Języka Polskiego*. Warszawa: PWN.
- [Radziszewski, 2013] Radziszewski, A. (2013). A Tiered CRF Tagger for Polish. In *Intelligent Tools for Building a Scientific Information Platform*. Studies in Computational Intelligence, vol. 467, pp. 215–230.
- [Sinclair et al., 2012] Sinclair, S., Rockwell, G. and the Voyant Tools Team (2012) Voyant Tools (web application). URL: <http://docs.voyant-tools.org>
- [Stamatatos, 2009] Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3): 538–556.
- [Straka & Straková, 2017] Straka, M. and Straková, J. (2017) Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. ACL, pp. 88–99.

- [Walkowiak, 2014] Walkowiak, T. (2014), Behavior of Web Servers in Stress Tests. In *Advances in Intelligent Systems and Computing* Vol. 286, Springer, pp. 467–476.
- [Walkowiak, 2016] Walkowiak, T. (2016). Asynchronous System for Clustering and Classifications of Texts in Polish. In: *Proceedings of the Eleventh International Conference on Dependability and Complex Systems DepCoS-RELCOMEX*, 2016, Springer International Publishing, pp. 529–538.
- [Walkowiak, 2018] Walkowiak, T. (2018). Language Processing Modelling Notation – orchestration of NLP microservices. In: *Advances in Dependability Engineering of Complex Systems*, Springer International Publishing, pp. 464–473.
- [Wittenburg et al., 2010] Wittenburg, P., Bel, N., Borin, L., Budin, G., Calzolari, N., Hajicová, E., Koskenniemi, K., Lemnitzer, L., Maegaard, B., Piasecki, M., Pierrel, J., Piperidis, S., Skadina, I., Tufis, D., van Veenendaal, R., Váradi, T., and Wynne, M. (2010) Resource and Service Centres as the Backbone for a Sustainable Service Infrastructure. In Nicoletta Calzolari et al. (ed.) *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*, European Language Resources Association (ELRA), pp. 60--63.
- [Zhao & Karypis, 2005] Zhao, Y. and Karypis, G. (2005). Hierarchical Clustering Algorithms for Document Datasets. *Data Mining and Knowledge Discovery*, 10(2): 1.

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/>