Tour de CLARIN
ITALY

Edited by Darja Fišer and Jakob Lenardič
Foreword

Tour de CLARIN highlights prominent user involvement activities of CLARIN National Consortia and Knowledge Centres with the aim to increase their visibility, reveal the richness of the CLARIN landscape, and display the full range of activities throughout the CLARIN network that can inform and inspire other consortia and knowledge centres as well as show what CLARIN has to offer to researchers, teachers, students, professionals and the general public interested in using and processing language data in various forms.

The brochure presents Italy and is organized in five sections:

• Section One presents the members of the consortium and their work
• Section Two demonstrates an outstanding tool
• Section Three highlights a prominent resource
• Section Four reports a successful event for researchers and students
• Section Five includes an interview with a renowned researcher from the digital humanities or social sciences who has successfully used the consortium’s infrastructure in their research
CLARIN-IT has established two national centres: ILC4CLARIN, which is hosted by the Institute for Computational Linguistics “A. Zampolli” in Pisa and is a B-certified repository that has been active since 2016, and ERCC, which is hosted by EURAC Research in Bozen and is currently a C-certified repository that has been active since 2018 and aims to become a B-certified centre. Through the two repositories, CLARIN-IT offers a variety of resources and services, such as MERLIN, which is a multilingual learner corpus for German, Italian and Czech, and SIMPLE, which is a multi-layer lexicon of Italian based on the Generative Lexicon Theory. There are also several natural language processing and analysis tools, many of which are offered as web services and integrated into Weblight.

The Italian consortium focuses on the field of Digital Classics, which still suffers from shortage or restricted availability of lexical resources for historical languages such as Ancient Greek, Latin or Sanskrit. To this end, the consortium aims to make some of the existing digitized resources for Ancient Greek and Latin available through its repositories (e.g. an LOD version of the TEI-dict Perseus Liddell-Scott Jones dictionary), as well as to create new ones by enriching existing corpora and lexical datasets with Linked Open Data. CLARIN-IT also specializes in the research of non-standard forms of language as found in learner corpora and computer-mediated communication. Moreover, CLARIN-IT focuses on oral archives which are at the crossroads of speech sciences, digital humanities and digital heritage.

Monica Monachini (second from the right) and some members of ILC4CLARIN: Francesco Frontini, Fahad Khan, Andrea Bellandi, and Federico Boschetti
LexO is a collaborative web editor used for the creation and management of (multilingual) lexical and terminological resources as linked data resources. The editor makes use of Semantic Web technologies (which enrich web data with semantic information in order to make them machine-readable) and the linked data publishing paradigm in order to ensure that lexical resources can be more easily shared and reused by the scientific community. In particular, LexO offers the following functionalities:

- It hides all the technical complexities related to markup languages, language formalities and other technology issues, facilitating access to the Semantic Web technologies to non-expert users who have not yet mastered Semantic Web-based standards and technologies, such as the Resource Description Framework and the Web Ontology Language (OWL).
- It provides the possibility for a team of users, each one with his/her own role (lexicographers, domain experts, scholars, etc.) to work on the same resource collaboratively.
- It adheres to international standards for representing lexica and ontologies in the Semantic Web (such as OntoLex-Lemon and OWL), so that lexical resources can be shared easily or specific entities can be linked to existing datasets (it is based on the OntoLex-Lemon model, currently regarded as a de facto standard for the modelling and publication of lexical resources as linked data).
- It provides a set of services implemented by means of RESTful Web Services that allow software agents access to resources managed by LexO.

**LexO: uses, lexical resources and communities**

LexO has so far been used in several DH research projects, such as:

- DiTMAO, a born-digital multilingual medico-botanical terminology focused on Old Occitan developed by philologists;
- FdS, a multilingual diachronic lexicon of Saussurean terminology in the framework of a lexicographic project;
- Totus Mundus, a bilingual Chinese-Italian resource dealing with Matteo Ricci’s Atlas. LexO has been used by historians to build the linguistic resources related to the map.

In addition, a demo of LexO with a subset of italwordnet adjectives is available as an online service through ILC4CLARIN.
The Interface

The LexO interface is composed of two main sections (Figure 1). Depending on which tab is selected, the left-hand side column will either show the list of lemmas composing the resource, a list of word forms, a list of lexical senses, or a list of concepts belonging to a reference ontology. If the resource is multilingual, then users have the possibility of filtering lemmas, forms and senses by language. Information related to the selected entry is shown in the central panel where the lemma appears in the upper part of the leftmost column on at head of a list of related forms. On the right, the lexical senses are shown.

Figure 1: The main interface of LexO

When selecting the “Dictionary View” tab, the central panel will show a dictionary-like rendering of all the information related to the selected entry (Figure 2). At the top of the central panel, a section can be expanded to query the resource, either by filling a series of fields for advanced searching (Figure 3b) or by composing queries in a controlled natural language style interface (Figure 3c). A team of users can work simultaneously in LexO to create, modify or delete a lexical entry, form, or sense, or to connect an entity to another entity, such as a sense to another sense via the “synonymy” property or a sense to a concept via the “ontological reference” property. The ontology can be imported using a dedicated tab in the left column (Figure 3a). Finally, administrators can monitor the lexicon construction process, for example by adding/removing users to the team, monitoring their productivity, and access the basic statistics of the lexicon (Figure 4).

Figure 2: Dictionary view of LexO. By selecting the “Dictionary View” tab, the central panel shows a dictionary-like rendering of all the information related to the selected entry.

Figure 3a: Imported ontology
References:


The developers of CLARIN-IT created a custom UIMA toolchain in order to enrich the corpus with additional layers of linguistic annotation, such as part-of-speech tagging and syntactic parsing. All in all, the texts were annotated with about 70 different features, covering orthography, grammar and lexicon of the learner language as well as specific sociolinguistic or pragmatic characteristics. This regards features such as the appropriate use of formality/politeness, e.g. the T/V distinction in German, or of idiomatic expressions like greetings or closing formulae.

The corpus has been designed to illustrate the Common European Framework of Reference for Languages (CEFR) with richly annotated authentic learner data. Since its publication in 2001, the CEFR has become the leading instrument of reference for the teaching and certification of languages and for the development of curricula. At the same time, there is a growing concern that the CEFR reference levels are not sufficiently illustrated, leaving practitioners such as teachers, test and curriculum developers, and textbook authors without comprehensive empirical characterizations of the relevant distinctions between the proficiency levels. This is particularly true for languages other than English, where supplementary empirical tools are urgently needed.

The MERLIN corpus was designed to address this demand for the three languages of Czech, German and Italian, by annotating authentic written learner productions and relating them to CEFR in a methodologically sophisticated way. To create the corpus, the partners relied on existing corpus annotation and search tools as much as possible. As no single tool was able to fulfil all the annotation requirements, a combination of tools was required to support the wide range of manual and automatic annotation that had been designed to illustrate the CEFR scales.

The manual annotation, which includes error annotation and the linguistic characteristics of the learner language, was performed using the Falko add-on for Microsoft Excel, which provides an existing framework for annotating learners’ errors, and the MMAX2 multi-level annotation tool, which is a flexible GUI-based tool for creating new annotations as well as visualizing them. Parallel to the manual annotation, the developers of CLARIN-IT created a custom UIMA toolchain in order to enrich the corpus with additional layers of linguistic annotation, such as part-of-speech tagging and syntactic parsing. All in all, the texts were annotated with about 70 different features, covering orthography, grammar and lexicon of the learner language as well as specific sociolinguistic or pragmatic characteristics. This regards features such as the appropriate use of formality/politeness, e.g. the T/V distinction in German, or of idiomatic expressions like greetings or closing formulae.
Since CLARIN-IT was established in 2016, its members have been organizing a series of roadshow events aimed at the Italian Digital Humanities and Social Sciences community. At the roadshows, CLARIN-IT experts present tools and resources deposited in CLARIN repositories (such as the Italian ILC4CLARIN data centre), offer examples of how CLARIN helps to promote novel research with NLP tools, as well as provide guidelines on how to eradicate bottlenecks that hamper the growth of a newly established digital discipline and, ultimately, in which ways scholars can profit from research infrastructures.

One of the most prominent events that the roadshow visited was the 5th Annual Conference of the Association for Humanities Information Sciences and Digital Culture, where the CLARIN-IT Coordinator Monica Monachini gave an invited talk about CLARIN in which she outlined how the Digital Humanities research community in Italy can benefit from the language technologies offered by CLARIN as well as from being involved in a large international research network based on interdisciplinary collaborations in a digital framework.

In order to help train young researchers and thus guarantee a broader utilization of computational tools and methods, a series of roadshow seminars aimed at students was also organized. The first seminars took place in October 2016 at the University of Parma and were intended both to promote CLARIN among students of Ancient Greek and to motivate them to adopt methods and concepts of computational linguistics and Digital Humanities in their studies. A second round took place from November to December 2017 with lectures concentrating on the methods, resources and instruments of a digital approach to philology. This included the encoding of text variants, digital repertoires of multiple editions of the same text and tools for their automatic alignment. During the lessons, CLARIN-IT provided examples of application of the TEI markup and Semantic Web technologies, by annotating geographical and personal references and linking an Ancient Greek lexicon in the Linked Open Data paradigm with a TEI-encoded fragmentary text of the poet Archilochus.

Other roadshow events were aimed at the promotion of CLARIN among graduate students with lectures at the “Master in Digital Humanities” event at Ca’ Foscari University on 3 November 2017 and as part of the “Digital Humanities, Web Resources, and Infrastructures” course at Venice International University on 4 December 2017. In addition, CLARIN-IT information events aimed at introducing the consortium, as well as CLARIN ERIC, to decision-making university figures and researchers.

References:
Abel, A and Wisniewski, K. 2015. MERLIN - die mehrsprachige Plattform für die europäischen Referenzniveaus at the 6th (Österreichische Gesellschaft für Sprachendidaktik) ÖGSD Conference in Salzburg.
Beatrice Nava is a PhD student who uses digital methodologies in Classical Studies.

Please describe your research background.

I have a bachelor’s degree in Greek Philology and a master’s degree in Modern Philology. I have always been interested in text reconstruction, and chose philology because I believe it is a crucial starting point for critical research into texts and the cultural contexts in which they are produced. My first hands-on experience in Digital Humanities was on the modelling of metadata of literary sources, which was made possible through a GARR grant (2017-2018). I am currently pursuing a PhD in Literary and Philological Cultures at the University of Bologna, where I am preparing a critical edition (i.e., a scholarly edition that includes a “critical apparatus” – annotations on the primary source material of a text) of a tragedy written by the famous Italian Romantic poet Manzoni and developing a model for digitizing the critical edition of the tragedy. I am also collaborating on a project funded by the Italian Ministry of Education, Universities and Research which aims to create a web portal dedicated to Manzoni. My principal role in this project is the digitization of Manzoni’s works using XML/TEI encoding. Additionally, I am using the same approach in the DEA project that focuses on Greek philology.

How has getting to know CLARIN influenced your research directions?

CLARIN was the starting point for my interest in Digital Humanities, and has made the opportunity to work in this field more realistic. It has motivated me to apply for a GARR grant with a proposal to model metadata for the description of Alessandro Manzoni’s manuscripts, which became the basis of my PhD project.
Why is the Digital Humanities approach important for classical philology? What kind of new research avenues does it open in the relatively traditional field? <

The application of language technologies and methodologies to solve research questions in Classical Philology is very important in relation to the structural potential of the digital medium. For example, just to mention one of the well-known but essential aspects, which is the option of organizing, storing and managing a substantial amount of data. In our case, we can easily manage and store all the hypotheses of previous editors and the additional useful information linked to the edition in a single place. Therefore, by providing an edition that is richer and much more complete than a paper-based one, it is possible to facilitate new philological studies. In fact, offering all the interpretations of previous editors through a single resource, with the addition of new hypotheses formed by studying the whole corpus of fragments, reopens the debate on some critical points. What is also important is that the digital medium helps scholars to efficiently exchange their ideas and results, as well as accelerates the response to new interventions into the text.

In addition, linguistic annotation allows the development of new teaching methods of Ancient Greek that are aimed at beginners and include the use of language services, such as treebanks and tools like TüNDRA adapted for classics. The annotations also enable an interactive approach to texts that is more inviting and accessible to students.

In addition to making data accessible and interoperable, NLP approaches facilitate and allow for the systematic production of fragment-specific lexica. In our case, having an annotated corpus allows us to develop linguistic services for teaching (e.g., Hyper-Text Archilochus, which is a prototype that provides the learner with a set of resources and tools that ease the critical assessment of ancient texts). It also acts as a stable and immutable sample for automatic translation experiments.

What are the challenges of digitizing and applying NLP techniques to Ancient Greek poetry? How does fragmentary poetry differ from other literary texts and what does this entail for its processing? <

Fragmentary ancient Greek poetry is very different from other literary texts. In fact, its tradition is more complex since it has different kinds of sources (manuscripts, papyrus, epigraphy) with variants and substantial lacunae (i.e., missing parts in a text). Applying NLP techniques to a fragmented tradition, which is complex and has many parts missing, can be particularly challenging, because gaps in the text call for multiple options for its reconstruction. Automatic
linguistic analyses of the whole corpus not only support new readings and interpretations, but also lead us to greater certainty as regards text corrections, integrations and authorship.

> Which tools would you like to see CLARIN Italy develop next that would help researchers interested in classical philology?

I would mostly like to see CLARIN-IT introduce in its repositories an integrated online environment that would support the proof-reading, encoding and enrichment of classical texts. I would also like to see CLARIN-IT experts draft precise guidelines or propose a paradigmatic schema on how to provide metadata specific to digital classics, such as the physical description of the source (papyrus, epigraph, manuscript, etc.), information on the origin, history, publication, and so on. We also need to provide the concordances of different editions of the same text, with correspondences to the fragments that have a different identification number in each edition. Moreover, I would like to see CLARIN-IT develop tools tailored to non-computational researchers that would help them perform linguistic and textual annotation (morphosyntactic, semantic, etc.) without requiring them to possess a great deal of technical know-how. In addition, improving the performance of the existing parsers for Ancient Greek on fragmentary texts would offer very important upgrades for the study and teaching of Ancient Greek.

You have recently visited CLARIN-DK experts on the CLARIN Mobility Grant. How was your research visit beneficial for your work? What knowledge and expertise did you gain from CLARIN-DK experts?

> At the Centre for Language Technology (Department of Nordic Studies and Linguistics, University of Copenhagen) I was given advice on how to better encode the different kinds of sources in which the digital classics are attested. Aside from the practical skills developed during my research stay, I found it inspiring to meet professors and researchers with different backgrounds and research objectives. Before my visit, I had mainly focused on the XML/TEI encoding, but my research stay allowed me to turn my attention to automatic linguistic analysis. I also gained a better understanding of existing tools and the potential of the CLARIN infrastructure as a network of not only language technologies, but also invaluable expertise.

Would you like to continue collaborating with CLARIN-IT after you finish your PhD? Do you have any wishes or plans already?

Yes, of course, I would like to continue working on the DEA project until the new edition of Archilochus' fragments is completed. I believe that further collaboration could improve my research methodology and allow me to gain better skills in applying linguistic analysis tools to my future research in Philological and Literary Studies. Moreover, I think the greater availability of post-PhD research grants, also on a national level, would be useful, as it would support research and, at the same time, aid the development of the national consortium in specific fields of knowledge. In the CLARIN-IT repositories, there currently aren’t many classical texts and resources, so I think that my philological knowledge in combination with my digital skills could make valuable contributions in this direction.
This brochure is part of the ‘Tour de CLARIN’ volume II (publication number: CLARIN-CE-2019-1537, November 2019).

Coordinated by Darja Fišer and Jakob Lenardič

Edited by Darja Fišer and Jakob Lenardič

Proofread by Paul Steed

Designed by Tanja Radež

Online version www.clarin.eu/Tour-de-CLARIN/Publication

Publication number CLARIN-CE-2019-1537 November 2019

ISBN 9789082990911

This work is licensed under the Creative Commons Attribution-Share Alike 4.0 International Licence.

Contact CLARIN ERIC c/o Utrecht University Drift 10, 3512 BS Utrecht The Netherlands www.clarin.eu