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-IT[2] and the periodically held evaluation campaign of Natural Language Processing - Evalita[3]; 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

[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 conference[6]. 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 institutions[7].

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/

[7]One 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 certification[8]. 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-GARR[9] 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-GARR[10] 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 Infrastructures[12] 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), ILC4CLARIN[13], 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 Observatory[14] (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 services[15] 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 Weblicht[16] (Hinrichs et al., 2010) and the Language Resource Switchboard[17] (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 texts[18] offers also a REST web service[19] 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, Weblicht[22] 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

[17] In 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 Conference[24]. 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 workshops[25] 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 Congress[27].

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 RaDT[28], is part of the beta-test group for the SDL Multiterm and Trados Studio[29] 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 Paisa (Lyding et al., 2014). It also has organized in 2017 the 5th Conference on CMC and Social Media Corpora for the Humanities[30] 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-ÖSS[31] 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 repository[32]. 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 committees[33].

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/

[25]Up 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).

[29]Leading 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/

[33]Members 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 needs[34].

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 project[35] 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 portal[36] 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

---

[34]The 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.

[35]Regione 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 *Gra.fo* 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 registry[38]); (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 Committee[39] 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 chain[40], 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.

---

[38]https://catalog.clarin.eu/ds/ComponentRegistry#/?itemId=clarin.eu\%3Acr1\%3Ap_
1369752611610&registrySpace=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 Humanities[41].

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*

---

[41]CLARIN-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-interviews[48]. 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 imporant 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' requirements[50]. 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.

---

[50]For 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 strenghtening 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 channels[51], 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 LRs 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 Chiocchetti, 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)*.