CLARIN-IT: State of Affairs, Challenges and Opportunities

Lionel Nicolas[a], Alexander König[a], Monica Monachini[b], Riccardo Del Gratta[b], Silvia Calamai[c], Andrea Abel[a], Alessandro Enea[b], Francesca Biliotti[c], Valeria Quochi[b]

[a] Institute for Applied Linguistics, Eurac Research, Bolzano, Italy
{lionel.nicolas, alexander.koenig, andrea.abel}@eurac.edu

[b] Istituto di Linguistica Computazionale "A. Zampolli", Consiglio Nazionale delle Ricerche, Pisa, Italy
{monica.monachini, riccardo.delgratta, alessandro.enea, valeria.quochi}@ilc.cnr.it

[c] Dipartimento di Scienze della formazione, scienze umane e della comunicazione interculturale,
University of Siena, Italy, {silvia.calamai, francesca.biliotti}@unisi.it

Abstract

This paper gives an overview on the Italian national CLARIN consortium and the status of CLARIN-IT in general. It thus discusses the current state of affairs of the consortium and provides information on the 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.

1 Introduction

Italy has joined CLARIN at the end of 2015 and since then, the national consortium is being set up while planning and implementing the first building blocks for the national CLARIN infrastructure. At present time, there are three members: the *Istituto di Linguistica Computazionale "A. Zampolli"* (ILC) of the *Consiglio Nazionale delle Ricerche* in Pisa, which is leading the consortium, the *Institute for Applied Linguistics* (IAL) of *Eurac Research* in Bolzano, and, both from the *University of Siena*, the *Dipartimento di Scienze della formazione, scienze umane e della comunicazione interculturale* (DISFSUCI) and the *Dipartimento di Filologia e critica delle Letterature antiche e moderne* (DFCLAM).

This paper is organized as follows: in Section 2 , the current state of affairs is discussed while in Section 3 the current members are shortly presented and their synergies with the overall CLARIN initiative are outlined. In Section 4, an overview of the next steps is provided right before concluding.

2 Current State of Affairs

As it stands, the current consortium includes only few participating organizations. Nonetheless, a noticeable number of other Italian institutions from a wide range of disciplines in the digital humanities expressed their interest in participating. Among those, we can cite the *Fondazione Bruno Kessler* (Trento), the *Università degli Studi di Parma, Dipartimento di Discipline Umanistiche* (Parma), the *Università Cattolica del Sacro Cuore* (Milano), the *Università "Ca' Foscari"* (Venezia), the *Università "Tor Vergata"* (Roma), the *Università di Pisa, Dipartimento di Linguistica* (Pisa).

One reason for the limited number of members are the still ongoing negotiations regarding an Italian national funding of the CLARIN-IT consortium with the Research Ministry. Consequently, while other institutions have put on hold their membership until a viable context for their participation can be arranged, the current 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.

On the technical perspective, the CLARIN-IT consortium closely cooperates with the *Consortium GARR*, the Italian University and Research Network, in particular with the IDEM-GARR[1] office that supports federated authentication in CLARIN. The CLARIN-IT consortium is also in contact with the CLOUD-GARR[2] office so as to allow members to safely and securely deposit data in the cloud.

[1] https://www.idem.garr.it/en

[2] https://cloud.garr.it

3 CLARIN-IT centers

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.

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 center 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 reality. The Institute is part of the Department of Social Science and Humanities, Cultural Heritage (DSU) of the *Consiglio Nazionale delle Ricerche* (CNR) and 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[3] 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 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 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-Center (aiming for type B certification in 2017), **ILC4CLARIN**[4], along with a *CLARIN DSpace* repository, where some of the above mentioned language resources are described according to the CMDI model, and made available in CLARIN VLO[5].

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; as well as preservation and persistent identification of research data and results is fundamental, since they provide users and researchers 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[6] (Hinrichs et al., 2010) 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 center available to a wider audience in the DH community while the CMDI model ensures a high quality in terms of metadata.

[3] http://lari.ilc.cnr.it

[4] https://ilc4clarin.ilc.cnr.it/en/

[5] https://vlo.clarin.eu/

[6] https://weblicht.sfs.uni-tuebingen.de

3.2 Synergies between the IAL and CLARIN

3.2.1 The IAL in few words

The *Institute for Applied Linguistics*[7] is part of *Eurac Research*, a private non-profit research center located in Bolzano, South Tyrol. The IAL is an international research environment where around 20 to 25 Junior and Senior Researchers with heterogeneous backgrounds are performing research on terminology, translation, variety linguistics, corpora, sociolinguistics, lexicography and computational linguistics

3.2.2 The IAL as an asset for CLARIN

With a majority of its workforce dedicated to linguistically-related or terminology-related research questions, the IAL is an active producer of manually created high-quality datasets of interest for the research community and beyond. For example, it is a known figure with regards to Learner Corpora (Abel et al., 2014), a type of datasets composed with textual data written by language learners that is then linguistically annotated and analyzed, and with regards to legal terminology, for which it has produced several terminological datasets (Chiocchetti et al., 2013). The IAL is also an active member of the ISO Committee TC/37 for ”Terminology and other language and content resources”. With the rest of its workforce providing assistance on automatic processing for their colleagues, the IAL is as well an active figure in the domain of computational linguistics, especially with regards to the automatic processing of the South Tyrolean German Dialect. Last but not least, because of its diversity in terms of research subjects and member profiles, the IAL relies on a varied set of workflows. As such, it can provide a wide range of expertise for devising and testing flexible solutions of interests for a larger scope of stakeholders.

Except for specific cases, the IAL intends to integrate as many resources as possible. It also intends to be involved in the CLARIN DSpace initiative, the General Assembly, the Scientific Advisory Board, the Standards Committee and the Standing Committee for CLARIN Technical Centres.

3.2.3 CLARIN as an asset for the IAL

As outlined earlier, the main added value from the IAL’s participation to CLARIN is the number of opportunities in terms of sustainability, an aspect that became key in the IAL’s strategy as it kept on growing over the years. Such opportunities thus constitute the main added value for the IAL. Apart from that, because the research profile of the IAL is varied but the one of CLARIN is even more, a number of CLARIN-related initiatives directly address needs of the IAL. A good example is the Language Resource Switchboard[8] (Zinn, 2016) which allows non expert stakeholders to seamlessly use advanced natural language processing tools. Indeed, computational linguists at the IAL assist linguists and terminologists to the best of their availability and capacity. Nonetheless, the more independent researchers are, the better and fluider can they develop their own research ideas, while relying on their computational linguist colleagues for the later stages (e.g. for the fine tuning of the automatic tools). In that perspective, technologies such as Switchboard directly tackle a need of the IAL.

3.3 Synergies between the DSFUCI/DFCLAM and CLARIN

3.3.1 The DSFUCI and DFCLAM in few words

Two departments of the *University of Siena* are part of the CLARIN-IT Consortium. On the one hand, the *Department of Scienze della Formazione, Scienze umane e della Comunicazione interculturale* (DSFU-CI) which is interested in building a digitization and cataloging system aiming at creating a regional network for the management of sound archives (Calamai et al., 2013), in the dissemination of oral archives to high school students and also in building cultural trail via Mobile APPs (Pozzebon and Calamai, 2015). On the other hand, the *Department of Philology and Criticism of Ancient and Modern Literature* (DFCLAM) focusses on the philological and literary competences that lie at the very heart of the study of literary texts, from the ancient world to modernity and for each literary genre: from poetry to theater, from novel to treatism, to history and to historiography. The interaction between philology and literature is central in the long-standing European humanistic tradition.

[7] `http://www.eurac.edu/en/research/autonomies/commul/Pages/`

[8] `http://weblicht.sfs.uni-tuebingen.de/clrs/`

3.3.2 The DSFUCI/DFCLAM as an asset for CLARIN

The inclusion of DSFUCI would give several advantages to CLARIN. For example, the Gra.fo speech archive (Calamai et al., 2013) is an unique and exemplary accomplishment in the Italian panorama. Gra.fo not only constitutes a precious repository of Tuscan memory and a first-hand documentation of Tuscan language varieties from the early 1960s, but it also represents a model that can be reproduced by others. Gra.fo covered the entire workflow with respect to the managing of oral archives: from digitization to long-term preservation, cataloging and description, ethical and privacy issues managing, and dissemination, also in terms of public history and general public involvement (Calamai et al., 2016).

The Dept. of *Filologia e critica delle Letterature antiche e moderne* committed itself to offering data and free online access to the ALIM (the Archive of the Italian Latinity of the Middle Ages) digital library which includes Latin texts and documents produced in Italy during the Middle Ages. Strategies for importing the metadata of ALIM in the CLARIN-ILC repository are under study, as well as procedures for delivering dedicated tools for textual and linguistic analysis through the CLARIN channels.

3.3.3 CLARIN as an asset for the DSFUCI

Being part of CLARIN would benefit the Gra.fo archive 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); (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; (3) the possibility to offer a proper reuse of research data (license agreement, ethical and legal issues).

4 Next steps and Conclusion

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. In case it cannot be, current members will provide assistance to potential new members for reaching a least a C certification. Regarding the ILC, the next steps are to complete the set of linguistic resources freely accessible through its online portal and achieve a CLARIN-B certification in 2017. For the IAL, the next steps are to set up a CLARIN DSpace repository and achieve a CLARIN-C certification in 2017. As regards the DSFUCI, the nexts steps are to make Gra.fo archives accessible via CLARIN DSpace and ensure their long term preservation.

In this paper, we have presented the current Italian CLARIN consortium and explained why it has a lot to offer to CLARIN and vice-versa. Despite a slow start, a pragmatical and flexible strategy is being implemented to steadily integrate CLARIN-IT into the CLARIN landscape.

References

Andrea Abel, Aivars Glaznieks, Lionel Nicolas, and Egon Stemle. 2014. Koko: an l1 learner corpus for german. In *Proceedings of the LREC Conference*.

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.

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*.

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.

Alessandro Pozzebon and Silvia Calamai. 2015. Smart devices for intangible cultural heritage fruition. In *Digital Heritage, 2015*, volume 1, pages 333–336. IEEE.

Claus Zinn. 2016. The clarin language resource switchboard. In *Proceedings of the CLARIN Conference*.