

CLARIN-IT: An Overview on the Italian Clarin Consortium After Six Years of Activity

Dario Del Fante¹, Francesca Frontini², Monica Monachini² and Valeria Quochi²

¹*Dipartimento di Studi Linguistici e Letterari, Università degli Studi di Padova*

²*Istituto di Linguistica Computazionale «A. Zampolli», CNR, Pisa*

Abstract

This paper offers an overview of the Italian CLARIN consortium after six years since its establishment. The members, the centres and the repositories and the most important collections are described. Lastly, in order to showcase the visibility and the accessibility of Language Resources provided by CLARIN-IT from a user-perspective, we show how Italian resources are findable within CLARIN ERIC.

Keywords

Language Resources, Data Repositories and Archives, Research Infrastructures, CLARIN

1. Introduction

CLARIN ERIC ¹ is one of the 20 European Research Infrastructure Consortia (ERICs). Its aim is to make digital language resources (henceforth LRs) [1] available to scholars and researchers from all disciplines. CLARIN-IT, the Italian CLARIN consortium was established in October 2015 [2], and it has recently been recognized as "project of international significance" by the Italian government. It benefits from the support of the Ministry of Research and has been listed among the high priority research infrastructures, according to *Decreto Ministeriale No.1082 del 10/09/2021 - Piano Nazionale per le Infrastrutture di Ricerca (PNIR) 2021-2027* ².


CLARIN ERIC has two fundamental objectives. The first one is to maintain digital repositories, where Language Resources (LRs), that is to say data, corpora, lexicons, tools, are catalogued, stored and retrieved in a simple way. The second one concerns the development of technological solutions that can be intuitively accessed by users. Therefore, CLARIN represents a structure where producers of language technologies and users of these technologies are connected and integrated. CLARIN's technical infrastructure is designed in accordance with the FAIR principles and aims at supporting the best practices of Open Science, by facilitating the deposit and long term preservation of data, their persistent identification and citation, by providing standardised and machine readable metadata and clear licenses, and easy access via single sign on³. An important added value provided by CLARIN is that of an increased findability of resources,

IRCDL 2022: 18th Italian Research Conference on Digital Libraries, February 24–25, 2022, Padova, Italy

✉ dario.delfante@unipd.it (D. Del Fante); francesca.frontini@ilc.cnr.it (F. Frontini); monica.monachini@ilc.cnr.it (M. Monachini); valeria.quochi@ilc.cnr.it (V. Quochi)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

¹<http://www.clarin.eu>

²<https://www.mur.gov.it/atti-e-normativa/decreto-ministeriale-n1082-del-10-09-2021>

³<https://www.clarin.eu/fair>

thanks to the harvesting of metadata from various repositories into the central infrastructure. The aim of this paper is twofold:

- To provide a clear overview of the Italian national CLARIN consortium as it currently stands, six years after its creation in terms of members, centres and collections stored
- To illustrate the visibility and the accessibility of Italian LRs within the CLARIN infrastructure

As concerns the second point, for reasons of space, we will not present here the overall architecture of CLARIN ERIC, for which we refer to publications such as [3, 2]. In this contribution we shall concentrate in particular on the Virtual Language Observatory, the CLARIN's meta catalogue where the metadata from all data centres are made visible and searchable from a single point of access.

In section 2, we discuss the current state of affairs of the Italian consortium in terms of members, and centres within the CLARIN federation, with a special focus on what they offer to CLARIN in terms of resources, services and expertise. In section 3, we simulate two queries on the VLO to highlight how this tool can contribute to the visibility of Italian resources from a user's perspective and we discuss about the results.

2. CLARIN-IT in 2021

2.1. Members

The CLARIN-IT⁴ consortium includes a founding member and seven full members. The current full members are the following:

1. the Istituto di Linguistica Computazionale "A. Zampolli" (ILC) of the Consiglio Nazionale delle Ricerche in Pisa is the founding member and host of the ILC4CLARIN repository⁵;
2. The EURAC Research Association (Bolzano) signed the CLARIN-IT Membership Application in 2017. The membership establishes that the organization creates a repository compliant with the CLARIN guidelines in which to deposit the metadata relating to the resources and tools available at its headquarters.
3. The Department of Education, Human Sciences and Intercultural Communication of the University of Siena signed the CLARIN-IT Membership Application in 2016.
4. The Department of Philology and Literary Criticism of the University of Siena signed the CLARIN-IT Membership Application in 2017.
5. The Bruno Kessler Foundation (Trento) signed the CLARIN-IT Membership Application in 2018.
6. The Archival and Bibliographical Superintendence of Tuscany (Firenze) signed the CLARIN-IT Membership Application in 2019
7. The Department of Electrical Engineering and Information Technology and the Interdepartmental Research Center "URBAN/ECO" of the University of Naples Federico II signed the CLARIN-IT Membership Applications in 2020.

⁴For a survey on Language Resources in CLARIN-IT, refer to [4]

⁵<https://ilc4clarin.ilc.cnr.it/>

8. The Catholic University of the Sacred Heart (Milano) joined CLARIN-IT by signing a Scientific Collaboration Agreement with ILC-CNR (Pisa) CLARIN in 2021.

Moreover, thanks to a continuous and focused User Involvement strategy, the Consortium is constantly expanding. Many other institutions have expressed their interest in joining CLARIN-IT or in depositing their data in the CLARIN national repository.

The Italian consortium embraces different research directions. One of those is the field of Digital Classics, which still suffers from shortage or restricted availability of language 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, as well as to create new ones by enriching existing corpora and lexical datasets with Linked Open Data. Another important direction is that of speech and oral archives, which are at the crossroads between speech sciences, digital humanities and digital heritage. CLARIN-IT collaborates with the University of Siena and the Superintendence of Tuscany to coordinate a project aiming at building a model and an architecture for the preservation, enhancement and accessibility of such archives. Finally, the EURAC partners are carrying out research around non-standard forms of language as found in learner corpora and computer-mediated communication.

The founding member and the consortium members are also involved in many international infrastructural projects, which aim to strengthen the cohesion of research across a number of related fields associated with the humanities. Among these we cite ELEXIS⁶ (on e-lexicography), the SSHOC cluster project⁷ (European open cloud ecosystem of data and tools for SSH), the TRIPLE project⁸ (a discovery platform for SSH). CLARIN-IT researchers are also active in standardization initiatives, such as ISO, TEI, W3C, and international academic organizations and networks, such as Learner Corpus Association⁹, Special Interest Groups on Computer Mediated Communication, the COST Action European Network for Combining Language Learning with Crowdsourcing Techniques (enetCollect¹⁰), and the COST Action Nexus Linguarum¹¹, for building an ecosystem of multilingual and semantically interoperable linguistic data at Web scale.

2.2. Centres

The CLARIN network is composed of distributed centres, which can be of three main types¹². Firstly, there are technical centres or B-Centres: these are generally hosted by a university or a public research institution and offer access to resources, services¹³ and/or knowledge. Secondly, there are Metadata Providing Centres or C-Centres: they offer deposit and metadata curation. Lastly, there are Knowledge Centres or K-Centres: centres sharing their knowledge

⁶<https://elex.is/>

⁷<https://sshopencloud.eu/>

⁸<https://project.gotriple.eu/>

⁹<https://www.learnercorpusassociation.org/>

¹⁰<https://enetcollect.eurac.edu/>

¹¹<https://nexuslinguarum.eu/>

¹²A complete list can be found here: <https://www.clarin.eu/content/overview-clarin-centres>

¹³<https://www.clarin.eu/content/services>

and expertise on one or more aspects of a domain covered by CLARIN. Their mission is to ensure that the available knowledge and expertise does not exist as a fragmented collection of informations, but it is made accessible in an organised way to both the CLARIN community and the social sciences and humanities research community at large. Each K-centre has its own specific areas of expertise.

CLARIN-IT comprises two data centres:

- The ILC4CLARIN B-centre, which is hosted and managed by the Institute for Computational Linguistics "A.Zampolli" in Pisa, the founding member of CLARIN-IT .
- The EURAC Research CLARIN Centre (ERCC) C-centre, which is hosted by the Institute for Applied Linguistics (IAL) at Eurac Research in Bolzano, a full member of CLARIN-IT.

Both centres offer the possibility to:

- deposit data, by ensuring that they are stored safely;
- search for data and tools and to download them easily;
- make the citation format as easy and consistent as possible;

Through the two repositories, CLARIN-IT offers a variety of resources (Cf. section 2.3 for an extensive excursus on CLARIN-IT offer). ILC4CLARIN, as the national B centre, has the mission to offer deposit facility to the whole of the national community, and also hosts a number of Natural Language Processing services¹⁴, many of which are offered as web services and integrated into the CLARIN pipeline management system WebLicht¹⁵.

In addition to these, Italy is also currently hosting two K-centers.

- CLARIN Knowledge Centre for Digital and Public Textual Scholarship (DiPText-KC)¹⁶, jointly maintained by University Ca' Foscari in Venice and ILC-CNR
- CLARIN Knowledge Centre for Computer-Mediated Communication and Social Media Corpora (CKCMC)¹⁷, a distributed K-centre jointly hosted by the Institute for Applied Linguistics, Eurac Research (IAL) in Bolzano, the Formal Linguistics Laboratory (LLF), in Paris, the Jožef Stefan Institute (IJS) in Ljubljana and the Leibniz-Institute for the German Language (IDS) in Mannheim

While these centres constitute the backbone of the national in-kind contribution to CLARIN ERIC, other activities are worth mentioning, such as the participation of members of the Italian consortium in various important CLARIN ERIC committees, such as the Legal and Ethical Issues Committee (CLIC) and the Standards Committee, as well as the work carried out to facilitate the deposit of important collections, such as for instance the Archivio della Latinità Italiana del Medioevo (ALIM).

¹⁴<https://ilc4clarin.ilc.cnr.it/services/>

¹⁵<https://weblicht.sfs.uni-tuebingen.de/>

¹⁶<https://diptext-kc.ilc4clarin.ilc.cnr.it/>

¹⁷<https://cmc-corpora.org/ckcmc>

2.3. Collections offered by CLARIN-IT

CLARIN-IT offers seven different digital collections, which are deposited in one of the two centres previously mentioned (section 2.2). As Table 1 shows, each collection includes a number of individual language resources.

Collections			
ALIM Literary Sources	344	ILC4CLARIN : OPEN Data and Tools	9
ILC4CLARIN	58	CIRCSE	8
Alim Documentary Sources	11	ERCC Learner Corpora	8
Eurac: Learner Language	10	ERCC Web Corpora	4

Table 1

Collections in CLARIN-IT. Situation at December 2021.

For example, *ALIM Literary Sources* [5] collection gives access to a vast archive Latin texts produced in Italy during the Middle Ages; its publication is of great importance for providing CLARIN-IT and the CLARIN community, at large, with critically reliable texts for the use of philologists, historians of literature, historians of institutions, culture and science of the Middle Age.

However, the Italian centres do not only host collections by Italian institutions. For example, the Ghent University has deposited three corpora in the ERCC centre:

- ACTER (Annotated Corpora for Term Extraction Research) v1.4
- Beldeko Summary Corpus v1.0.0
- ACTER (Annotated Corpora for Term Extraction Research) v1.3

Indeed, both CLARIN-IT centers offer LRs in a variety of languages, not only Italian, as shown in Table 2.

Languages			
Latin	369	Croatian	1
English	43	Modern Greek	2
Italian	38	Croatian	1
Arabic	32	Ladino	1
German	12	Mòcheno	1
Ancient Greek	10	Sardinian	1
French	4	Saurano	1
Dutch	4	Slovenian	1
Czech	2	Spanish; Castilian	1
Basque	1	Trentino	1
Breton	1	Tyrolean	1
Cimbrian	1	Veneto	1

Table 2

Languages in CLARIN-IT

Since ILC4CLARIN is specialised in ancient texts, Latin and Ancient Greek are particularly represented. However, as discussed in [4], Latin LRs are overrepresented, even with respect to

Italian, because of the metadata of the ALIM corpus: each text of that collection is deposited as a separate resource, while this is not true for other corpora [5].

3. Italian Resources in CLARIN

The most important point of access for CLARIN is the Virtual Language Observatory (VLO)¹⁸ [6] which harvests metadata from all the official CLARIN data providing centres and makes them searchable via a unified interface offering faceted search. Other interesting and useful central discovery services are the Federated Content Search (FCS)¹⁹, the Language Resources Switchboard (SB)²⁰, and the CLARIN Resource Families²¹. In order to investigate the visibility and the accessibility of Italian LRs in CLARIN-IT from a user-perspective, we showcase two simple queries which can be performed using the faceted search functionalities of the VLO. Each query was raised in different formats. Our aim here is to assess the visibility of resources from Italian centres, but also resources that may be relevant to Italian researchers offered by centres outside of Italy, by trying to replicate as much as possible the behaviour of a non-expert user, accessing the VLO for the first time.

Query 1 - Search for Italian Corpora in the VLO by filtering for:

1. *Language = Italian .*
2. *Resource type = Corpora.*

This query returns 159 results. If we filter them by organisation, we can easily assess that the resources from the Italian centres are correctly displayed in the meta catalogue. For instance by adding the filter

- *organisation = Institute for Applied Linguistics, Eurac Research*

all 7 corpora from ERCC are selected. At the same time, it also becomes evident that important resources for the Italian language are also offered by other centres outside Italy, such as, among others, the Universal Dependencies Consortium (20 corpora) and the Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (UFAL) (12 corpora) which host multilingual collections of Treebanks, namely syntactically annotated corpora.

From this query it is thus evident that the participation of Italian centres in CLARIN not only allows resources produced in Italy to gain visibility outside of the national context, but also to make important resources hosted abroad more accessible to Italian researchers. A similar experiment can be attempted for Latin, a language that is strongly represented among CLARIN-IT resources.

¹⁸<https://vlo.clarin.eu>

¹⁹<https://contentsearch.clarin.eu/>

²⁰<https://switchboard.clarin.eu/>

²¹<https://www.clarin.eu/resource-families>

Query 2 - Search for Latin lexical resources in the VLO by filtering for:

1. *Language = Latin*
2. *Resource type = Lexical Resource*

This query returns 22 results. Among these we find the important resources made available on ILC4CLARIN by the CIRCSE collection, which contains the results of the Linking Latin (LiLa) ERC project²², produced by Marco Passarotti and his colleagues. However, an important resource by the CIRCSE lab, namely the Word Formation Latin (WFL), is hosted instead by the LINDAT repository, together with the other resources of the multilingual Universal Derivations collection. Thus the VLO view allows users to easily find links between resources that are hosted on different repositories, and so making them more accessible and reusable in research.

4. Conclusive Remarks and Future Work

The infrastructures such as CLARIN play an important role to overcome future challenges of Open Science, and to deliver on the promise of a digital ecosystem of freely accessible, interconnected resources, where data from different providers, made available according to the FAIR principles, can be reused and combined to produce novel research. Within this context, the role of standards in the archival management is fundamental.

The CLARIN-IT consortium is constantly increasing the number of resources deposited in its centres, and also conducting a regular monitoring of the different collections provided by various partners across the two repositories, so as to verify their visibility in CLARIN.

An important future challenge is that of increasing CLARIN's users base at the national level. CLARIN has developed different methodologies and approaches to measure and evaluate user engagement. Surveys can be used to test the interest in the use of digital resources and related tools [7]. A new survey, which is still on going, shows that while Italian researchers in the domain of Language Resources and Technologies are mostly aware of CLARIN-IT's services, a large number of them is still relying on local repositories or on GitHub to store their data. It is thus crucial that CLARIN-IT provides adequate training to resources such as the VLO, in particular targeting the needs of specific communities. In this sense the newly released tutorial *CLARIN Tools and Resources for Lexicographic Work* [8] is a step in this direction.

References

- [1] J. Godfrey, A. Zampolli, Language resources, in: Survey of the State of the Art in Human Language Technology. *Linguistica Computazionale*, XII-XIII., Cambridge University Press, 1997, pp. 381–384.
- [2] M. Monachini, F. Frontini, CLARIN, l'infrastruttura europea delle risorse linguistiche per le scienze umane e sociali e il suo network italiano CLARIN-IT, *Italian Journal of Computational Linguistics* 2 (2016) 11–30. URL: <http://journals.openedition.org/ijcol/387>. doi:10.4000/ijcol.387.

²²<https://lila-erc.eu/#page-top>

- [3] F. de Jong, B. Maegaard, D. Fišer, D. van Uytvanck, A. Witt, Interoperability in an infrastructure enabling multidisciplinary research: The case of CLARIN, in: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 3406–3413. URL: <https://aclanthology.org/2020.lrec-1.417>.
- [4] D. Del Fante, F. Frontini, M. Monachini, V. Quochi, CLARIN-IT resources in CLARIN ERIC - a bird's-eye view, in: CLARIN Annual Conference 2021, Proceedings CLARIN Annual Conference 2021, 2021, pp. 129–133.
- [5] F. Boschetti, R. Del Gratta, M. Monachini, M. Buzzoni, P. Monella, R. Rosselli Del Turco, “Tea for Two”: The Archive of the Italian Latinity of the Middle Ages meets the CLARIN Infrastructure, in: C. Navarretta, M. Eskevich (Eds.), Proceedings of CLARIN Annual Conference 2020. Virtual Edition, 2020. URL: https://office.clarin.eu/v/CE-2020-1738-CLARIN2020_ConferenceProceedings.pdf.
- [6] D. Broeder, M. Kemps-Snijders, D. V. Uytvanck, M. Windhouwer, P. Withers, P. Wittenburg, C. Zinn, A data category registry- and component-based metadata framework, in: N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias (Eds.), Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, 2010, pp. 19–21.
- [7] M. Monachini, A. Nicolosi, A. Stefanini, Digital classics and CLARIN-IT: What italian scholars of ancient greek expect from digital resources and technology, in: Selected papers from the CLARIN Annual Conference 2017, Budapest, 18-20 September 2017, 2018, pp. 61–74.
- [8] F. Frontini, A. Bellandi, V. Quochi, M. Monachini, K. Mörth, S. Zhanial, M. Ďurčo, A. Woldrich, CLARIN Tools and Resources for Lexicographic Work, 2022. URL: <https://elexis.humanistika.org/en/resource/posts/clarin-tools-and-resources-for-lexicographic-work>, publisher: DARIAH-Campus.