CLARIN Annual Conference Proceedings

2021

Edited by

Monica Monachini, Maria Eskevich

27 – 29 September 2021
Virtual Edition

Please cite as:
CLARIN Annual Conference Proceedings, 2021. ISSN 2773-2177 (online).
Eds. M. Monachini and M. Eskevich.
Virtual Edition, 2021.

Programme Committee

Chair:

- Monica Monachini, Institute of Computational Linguistics “A. Zampolli" (IT)

Members:

- Lars Borin, University of Gothenburg (SE)
- António Branco, Universidade de Lisboa (PT)
- Tomaž Erjavec, Jožef Stefan Institute (SI)
- Eva Hajičová, Charles University Prague (CZ)
- Erhard Hinrichs, University of Tübingen (DE)
- Marinos Ioannides, Cyprus University of Technology (CY)
- Langa Khumalo, North West University (ZA)
- Nicolas Larrousse, Huma-Num (FR)
- Krister Lindén, University of Helsinki (FI)
- Karlheinz Mörth, Austrian Academy of Sciences (AT)
- Costanza Navarretta, University of Copenhagen (DK)
- Jan Odijk, Utrecht University (NL)
- Maciej Piasecki, Wrocław University of Science and Technology (PL)
- Stelios Piperidis, Institute for Language and Speech Processing (ILSP), Athena Research Center (GR)
- Eirikur Rögnvaldsson, University of Iceland (IS)
- Kiril Simov, IICT, Bulgarian Academy of Sciences (BG)
- Inguna Skadiņa, University of Latvia (LV)
- Koenraad De Smedt, University of Bergen (NO)
- Marko Tadič , University of Zagreb (HR)
- Jurgita Vaičenonienė, Vytautas Magnus University (LT)
- Tamás Váradi, Research Institute for Linguistics, Hungarian Academy of Sciences (HU)
- Kadri Vider, University of Tartu (EE)
- Martin Wynne, University of Oxford (UK)

Reviewers:

- Lars Borin, SE
- António Branco, PT
- Tomaž Erjavec, SI
- Eva Hajičová, CZ
- Martin Hennelly, ZA
- Erhard Hinrichs, DE
- Marinos Ioannides, CY
- Nicolas Larrousse, FR
- Krister Lindén, FI
- Monica Monachini, IT
- Karlheinz Mörth, AT
- Costanza Navarretta, DK
- Jan Odijk, NL
- Stelios Piperidis, GR
- Eirikur Rögnvaldsson, IS
- Kiril Simov, BG
- Inguna Skadiņa, LV
- Koenraad De Smedt, NO
- Marko Tadić, HR
- Jurgita Vaičenonienė, LT
- Tamás Váradi, HU
- Kadri Vider, EE
- Martin Wynne, UK

Subreviewers:

- Federico Boschetti, IT
- Christophe Parisse, FR
- Thorsten Trippel, DE
- Valeria Quochi, IT
- Zijian Győző Yang, HU
- Efstathia Soroli, FR
- Enikő Héja, HU
- Bence Nyéki, HU
- Angelo Mario Del Grosso, IT
- Olivier Baude, FR
- Kinga Jelencsik-Mátyus, HU

CLARIN 2021 submissions, review process and acceptance

- Call for abstracts: 19 January 2021, 1 March 2021
- Submission deadline: 28 April 2021
- In total 40 submissions were received and reviewed (three reviews per submission)
- Virtual PC meeting: 16-17 June 2021
- Notifications to authors: 22 June 2021
- 35 accepted submissions

More details on the paper selection procedure and the conference can be found at https://www.clarin.eu/event/2021/clarin-annual-conference-2021-virtual-event.

Table of Contents

Research cases

Resources

Annotation and Acquisition Tools

Research Data Management, Metadata and Curation

Repositories and National CLARIN Centres

Legal Issues Related to the Use of LRs in Research

CLARIN-IT Resources in CLARIN ERIC - a Bird's-Eye View

Dario Del Fante
ILC-CNR - Italy
dario.delfante@ilc.cnr.it

Francesca Frontini
ILC-CNR - Italy
francesca.frontini@ilc.cnr.it

Monica Monachini
ILC-CNR - Italy
monica.monachini@ilc.cnr.it

Valeria Quochi
ILC-CNR - Italy
valeria.quochi@ilc.cnr.it

Abstract

This paper investigates the visibility of CLARIN-IT language resources within the services of the CLARIN ERIC central infrastructure, notably the Virtual Language Observatory, the Switchboard and the Federated Content Search, from a user perspective in order to identify possible issues. While the experiment focused on one national consortium, the ultimate goal is to develop an assessment methodology that can be used by any national consortia aiming to review the accessibility of their resources and tools within the CLARIN central services.

1 Introduction

With a distributed network of over 50 centres, CLARIN ERIC's principal aim is to ensure easy access to their resources and tools by researchers from all over Europe and beyond, independently of their original producers, of the centre or consortium physically hosting them. Ideally, a researcher should not need to know where a given resource is deposited or even be aware of its existence to be able to find, access and use it, thanks to the central functionalities and services available via the CLARIN portal, which is a gateway to the whole network's offerings.

The first and foremost central service, the CLARIN *shop window*, is the Virtual Language Observatory (VLO)[1] (Broeder et al., 2010) 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)[2], the Language Resources Switchboard (SB)[3], and the CLARIN Resource Families[4].

In order to ensure visibility from the CLARIN central services, we argue that national consortia should monitor regularly these four "points of access" and analyse them from a user perspective. We are not talking about automated checks, but rather of a more qualitative assessment aimed at ensuring that any researcher/end-user can easily find the resources she needs and use them as intended. As pointed out by Sugimoto (2016), despite the wide array of useful services for digital research in linguistics and the humanities, it is unclear whether the community is thoroughly aware of the status-quo of the growing infrastructure. At the same time, such an analysis could provide useful instruments in the hands of national coordinators and center managers for bringing to the fore strengths and critical issues of their data providing community.

For these reasons we set out to check and analyse the presence of the LRs available in the CLARIN-IT consortium in the central discovery services with the twofold aim of (i) assessing the Italian consortium presence and of (ii) devising a reproducible qualitative methodology from the user perspective. In this paper, we will present a case-study aimed at investigating the visibility, reliability and searchability of CLARIN-IT LRs in the VLO. In doing so, we shall sketch a proposal for generalising this qualitative assessment procedure to any given consortium.

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/ final paper: en-uk version (to license, a licence)

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

[2] https://contentsearch.clarin.eu/

[3] https://switchboard.clarin.eu/

[4] https://www.clarin.eu/resource-families

2 CLARIN-IT in the VLO

CLARIN-IT is actively involved in the sectors of documentation, digitization and language technologies for the Humanities and it is focused on Language Resources (LRs) both data and tools (Monachini and Frontini 2016,). Currently the consortium offers two data centres, ILC4CLARIN[5], the national B centre, and the EURAC Research CLARIN centre (ERCC)[6]; both host repositories for the preservation of LRs, which thus become visible from the Virtual Language Observatory (VLO) [7]. The VLO represents the principal means of exploring LRs in CLARIN. By using a facet browser that allows for the filtering of metadata records according to previously specified categories - the facets - the VLO makes it possible to carry out targeted searches.

However, given the variability of the CMDI metadata framework (Haaf et al., 2014), the VLO represents an improvable asset. In this work we rely on and extend a previous attempt to assess the searcheability of tools (Odijk, 2019; Odijk, 2014).

The CLARIN-IT resources in the VLO can be easily extracted thanks to a faceted query that uses the national project as filter[8]. The search query returns 490 different LRs, of which 51 are hidden because of duplicate naming, which leaves 439 distinct resources as shown in Table 1. The presence of duplicate naming represents a quite common occurrence in the VLO search. The VLO browser automatically removes all the duplicates from the search results on the basis of the naming. However, the presence of duplicates is specified under each problematic record.

CLARIN-IT - a birds eye view	
Total Number of LR	439
Monolingual	388
Multilingual resources	46
Format	12
Languages	10
Organisations	8
Collections	7
Resource type	6
Data providers	2

Table 1. CLARIN-IT on VLO

This basic query can be further narrowed down by using other facets, which allows not only for a systematic classification of the national resources by language, resource type, collection, but also for the verification of other metadata such as subject, format, availability, and so on. The aim is to determine the extent to which the resources are correctly described, in particular by verifying the difference between the number of LRs which are expected to occur and the actual number of LRs under each facet. The idea is to explore and test an assessment procedure that may assist repository managers, national coordinators or even the central office in harmonising the content of each repository and consequently of the VLO. In what follows we discuss the results of some of the most interesting filters.

2.1 Languages of CLARIN-IT

Filtering by *Language* identified ten different languages as indicated in Table 2.

These results show that CLARIN-IT offers LRs in a variety of languages, not only Italian. Due to the specialisation of the ILC4CLARIN, Latin and Ancient Greek are particularly represented. However, at a closer look, the over-representation of Latin LRs even with respect to Italian ones is due to the choice of metadata description of the ALIM corpus, which reflects the internal organisation of the original archive.

[5] `https://ilc4clarin.ilc.cnr.it/`

[6] `https://clarin.eurac.edu/`

[7] `https://www.clarin.eu/content/virtual-language-observatory-vlo`

[8] `https://vlo.clarin.eu/search?fqType=nationalProject:or&fq=nationalProject:CLARIN-IT`

Languages			
Latin	366	Italian	30
English	40	German	8
Arabic	32	Czech	2
Ancient Greek (to 1453)	6	Breton	1
Ancient Greek	8	Basque	1

Table 2. Languages in CLARIN-IT

As described in (Boschetti et al., 2020), every text of that corpus is deposited as a separate resource, while this is not true for other corpora. Such a finding may indicate a need for harmonization of the depositing guidelines for specific resource types. *Thus, the store of ALIM corpus in such a way, might be problematic in terms of availability and search-ability because its structures lacks of interconnections.*

2.2 Organisations and Collections of CLARIN-IT

By checking the filter results of the *Organisation* and *Collections* facets we can easily verify that the organisations and consortium members are actively contributing to the repository and correctly represented in the VLO. Table 3 indicates the number of LRs which each organisation is responsible for:

Organisations			
Archivio della Latinità Italiana del Medioevo (ALIM)	354	CIRCSE - Università Cattolica Sacro Cuore	8
Istituto di Linguistica Computazionale - CNR	39	Ghent Universities	2
Institute for Applied Linguistic Research - EURAC	9	Università di Parma	2
Università di Salerno	8	Basque	1

Table 3. Organisations in CLARIN-IT

This query confirms that almost all of the Latin LRs are indeed items of the ALIM collection, but also that most of the CLARIN-IT LRs are produced by Italian consortium members. Looking at Table 3, only two LRs deposited in CLARIN-IT centres were produced by a foreign institution (Ghent University and Basque).

Collections			
ALIM Literary Sources	344	ILC4CLARIN : OPEN Data and Tools	7
ILC4CLARIN	54	ERCC Learner Corpora	8
Alim Documentary Sources	11	ERCC Web Corpora	4
CIRCSE	8	ERCC	1

Table 4. Collections in CLARIN-IT

2.3 Resource Types and Data Providers of CLARIN-IT

By combining *Resource type* and *Data Provider* it is possible to check the number and types of resources offered by each of the two CLARIN-IT centers, and thus get some information on their specialisation, as appears in Table 5.

2.4 CLARIN-IT Formats and Subjects

A very useful query concerns the available Formats and Subjects for CLARIN-IT resources. By checking this we assess whether all resources are correctly typed and whether they have been further described with suitable and harmonised subject keywords. In the Italian case, the coverage for the latter seems to

ILC4CLARIN		**Eurac Research**	
Corpus	368	Corpus	13
Lexical Resource	43		
Software, webservice	12		
Webservice	2		
Text	1		

Table 5. Resource type for each Data provider

be incomplete (only 18 LRs are mapped onto VLO subjects keywords, whereas many of the keywords present in the national repositories are not visible in the VLO) and harmonisation could be increased by using controlled vocabularies.

2.5 CLARIN-IT LRs availability

One important final check concerns Availability, which indicates “degree to which resources and tools are publicly accessible”. In the case of CLARIN-IT, most of the LRs are publicly available; however, the filter also returned 25 resources with unspecified availability. A closer inspection shows that these correspond to corpora from the ERCC repository and webservices from ILC4CLARIN. This finding might lead to amendments of the records.

3 The Methodology

During this work, we assessed the presence of two unexpected issues which gave us the opportunity of learning two relevant lessons in terms sustainability and usability of our methodology.

3.1 Two Lessons

- *Granularity*
  Regarding the CLARIN-IT consortium, we found that there are some cases when LRs are stored in the VLO as single records belonging to the same collection. This is due to the fact that some repository does not support a nested archive. For example, the collection *ALIM Lietrary Sources*, does not appear in the VLO as a single entry. It is stored as a collection of different entries categorised as corpora despite the fact that are composed of only one single texts. As a result, the high level of granularity in relation to data storage might work against the availability and the accessibility of the resources caused by the dispersive collection methods. However, different levels of granularity respectively correspond to different analytical possibilities. A nested archive represent a more solid sample of data which can be representative of a specific language and might give different options for analysis by using the Switchboard. Differently, a non-nested archive represents a more agile option for comparing records from different collections and national consortia by saving these records as virtual collections.

- *Duplicate namings*
  The second issue encountered is related to the presence of duplicates which are automatically removed from the results produced by the search in the VLO. This issue has attracted our attention and after a careful examination, most of them resulted in being false duplicates. Within the *ALIM Literary Sources* for instance, all 50 (out of 394) hidden items, are in fact different critical editions of the same texts by different editors. Having the exact same title, the system considers them as duplicates. For example the *Summa Dictaminis* corresponds to three records, one for each editor (Matteo de’Libri , M. Thumser, Emil Polak). While a possible strategy for avoiding such texts to be treated as duplicates could be to add ”by EDITOR” in the ’title’ metadata, this conflicts with the collection praxis. This may be another issue for discussion for the Standing Committee on CLARIN

Technical Centres. In order to enhance the availability and the accessibility of the data within the VLO, we propose that greater attention should be payed in relation to these aspects.

3.2 Extending the methodology

Based on the results and observations stemming from the analysis we have just described, a more general methodology for qualitative assessment can be thus generalised. We suggest the following checks should be carried out by new consortia, after the registration of at least one B or C centre, but also, periodically, buy existing national consortia, especially when new centres are registered or large collections are injected.

1. Select in the National Project tab, the project of interest in order to select only the LRs which are provided by CLARIN centre of the national consortium

2. Check which LRs are shown and how are presented in the VLO, filtering for: (a) Languages, (b) Organisations and Collections, (c) Resource type

3. Check the presence of duplicates

4. Check the status of activation for a sample of links to the original place

5. Register all the inconsistencies in terms of accessibility and availability

4 Conclusions

With the growth of the CLARIN-IT consortium, a thorough check of the LRs contributed by various partners across two repositories has become necessary. This exercise of qualitative assessment of the visibility of the consortium resources in the VLO has proven extremely useful and might become a model for other projects. To this end, more checks could be added, including a template search for important resources such as reference corpora and lexicons, to ensure that they correctly appear as expected.

References

Boschetti, F., Del Gratta, R., Monachini, M., Buzzoni, M., Monella, P., and Rosselli Del Turco, R. 2020. "Tea for Two": The Archive of the Italian Latinity of the Middle Ages meets the CLARIN Infrastructure. In C. Navarretta and M.Eskevich (Eds.) *Proceedings of CLARIN Annual Conference 2020.* . Virtual Edition.

Haaf, S., Fankhauser, P., Trippel, T., Eckart, K., Hedeland, H., Herold, A., Knappen, J., Schiel, F., Stegmann, J., and Van Uytvanck, D. 2014. CLARIN's virtual language observatory (VLO) under scrutiny - the VLO taskforce of the CLARIN-D centres. In *CLARIN annual conference 2014.*

Monachini, M. and Frontini, F. 2016 CLARIN, l'infrastruttura europea delle risorse linguistiche per le scienze umane e sociali e il suo network italiano CLARIN-IT. *IJCoL - Italian Journal of Computational Linguistics*, 2(2),11-30.

Odijk, J. 2014. *Discovering Resources in CLARIN: Problems and Suggestions for Solutions.* Unpublished paper.

Odijk, J. 2019. *Discovering software resources in CLARIN.* In *Selected Papers Clarin Conference 2018*, 159: 121-132.

Sugimoto,G. 2016 Number game -Experience of a European research infrastructure (CLARIN) for the analysis of web traffic. In *CLARIN Annual Conference 2016*,ArXiv: abs/1706.05089.

Broeder, D., Kemps-Snijders, M., Van Uytvanck, D., Windhouwer, M., Withers, P., Wittenburg, P., and Zinn, C. 2010. A Data Category Registry and Component-based Metadata Framework. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).* European Language Resources Association (ELRA).