CLARIN-IT:
Language Resources and Infrastructures for Digital Humanities

Monica Monachini
CLARIN Italian National Coordinator
ILC-CNR
CLARIN Italian National Representative

Summer School “New trends in quantitative and computational linguistics”
Universidad de Castilla-La Mancha, Ciudad Real, 28 September 2016
1 October 2015

• Italy became a Full Member of CLARIN ERIC, the governance and coordination body of the Common Language Resources and Technology Infrastructure at the European level

• This is a great opportunity for those who are involved in Humanities and Social Sciences
What is meant by RI

- A set of resources, tools, competencies, data, services and functions specialized and dedicated to feed the necessary streams to scientific and technological research on a continuous basis

- CLARIN
  - is an infrastructure of type **ERIC** (European Research Infrastructure Consortium) formally acknowledged
  - has completed the **ESFRI** (European Strategy Forum on Research Infrastructures) **roadmap**
  - has achieved the status of «**landmark**»
What LRs are

- Digital language resources
  - corpora (annotated and not), lexica, ontologies
  - formal grammars, statistical models
  - tools for (written, spoken, multimodal) Natural Language Processing

Ciudad Real, 28/09/2016

Monica Monachini (ILC-CNR)
What LRs are for

• Language resources can be useful in
  – linguistic studies
  – analysis of textual sources
  – language processing
  – automatic annotation of texts
  – …
Current situation and needs

• Language resources often
  – are hosted by the institution that has produced them
  – are intended for the scientific purposes and environments of a project
  – are subject to restrictions

• Building language resources requires a good deal of time and money
  – it is crucial to ensure an easy access to resources to researchers of different institutions and countries
  – it is fundamental to ensure that such resources remain accessible over time

• A change of paradigm is essential
  – sharing, access and replicability of results
• CLARIN is an intangible digital infrastructure aiming at providing
  – a system for the **description** and **cataloguing** of existing language resources
  – a **repository** for the long-term electronic storage
  – an easy and sustainable **access** to scholars of member countries

• CLARIN furthermore provides
  – advanced tools for **identifying** language resources suitable to one’s research wherever they are hosted and **combining** them in order to analyse, explore and annotate digital data
CLARIN services

- **CLARIN portal**: Get an example-based impression of what's currently available.
- **Depositing services**: Store language resources in a sustainable repository at a CLARIN centre.
- **Virtual Language Observatory**: Discover language resources using a faceted browser or a map.

https://www.clarin.eu/content/services
CLARIN services

- Easy access to protected resources
  - Get easy access to protected resources, with your institutional username and password.

- Web services and applications
  - Explore and analyze language data with a wide variety of tools

- Virtual Collections
  - Create your own digital bookmarks, ideal for citing data sets.

https://www.clarin.eu/content/services
Language Resource Inventory
Submit and access information about language resources relevant to your research.

Content Search (prototype version)
Search different corpora with a single search engine

Consulting Services
Searching for a specific data set or application? Wondering how CLARIN can assist your research? Feel free to contact us!

https://www.clarin.eu/content/services
How CLARIN ERIC works

- CLARIN ERIC is organized as a consortium of nations (and supranational bodies)
- Each nation
  - pays an annual membership fee
  - offers in-kind services for the maintenance of the consortium
  - make resources and services available through a CLARIN Centre or more
How a CLARIN National Consortium works

• CLARIN includes 19 member countries (plus a few observer countries)
• CLARIN National Consortia include various institutions (research centres, universities, libraries, institutions supporting research, …) that
  – provide data, tools and services as CLARIN Centres
  – are financially supported by one’s nation through national strategic projects
<table>
<thead>
<tr>
<th>Members</th>
<th>National Consortium (NC)</th>
<th>Leading NC partner</th>
<th>Number NC partners</th>
<th>National coordinator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austria</td>
<td>CLARIN Austria</td>
<td>CLARIN Centre Vienna</td>
<td>4</td>
<td>Karlheinz Mörth</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>CLARIN Bulgaria</td>
<td>Bulgarian Academy of Sciences</td>
<td>10</td>
<td>Kiril Simov</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>LINDAT Clarin</td>
<td>Charles University Prague</td>
<td>4</td>
<td>Eva Hajičová</td>
</tr>
<tr>
<td>Denmark</td>
<td>CLARIN Denmark</td>
<td>University of Copenhagen</td>
<td>4</td>
<td>Bente Maegaard</td>
</tr>
<tr>
<td>Dutch Language Union</td>
<td>Dutch Language Union</td>
<td>Dutch Language Union</td>
<td>4</td>
<td>Jan Theo Bakker</td>
</tr>
<tr>
<td>Estonia</td>
<td>CLARIN Estonia</td>
<td>Center of Estonian Language Resources</td>
<td>3</td>
<td>Kadri Vider</td>
</tr>
<tr>
<td>Finland</td>
<td>FIN-CLARIN</td>
<td></td>
<td></td>
<td>Krister Lindén</td>
</tr>
<tr>
<td>Germany</td>
<td>CLARIN-D</td>
<td>University of Tübingen</td>
<td>12</td>
<td>Erhard Hinrichs</td>
</tr>
<tr>
<td>Greece</td>
<td>CLARIN Greece</td>
<td></td>
<td></td>
<td>Stelios Piperidis</td>
</tr>
<tr>
<td>Italy</td>
<td>CLARIN Italy</td>
<td></td>
<td></td>
<td>Monica Monachini</td>
</tr>
<tr>
<td>Lithuania</td>
<td>CLARIN-Lithuania</td>
<td>Vytautas Magnus University</td>
<td>3</td>
<td>Jurgita Vaičenonienė</td>
</tr>
<tr>
<td>The Netherlands</td>
<td>CLARIN-NL</td>
<td>Utrecht University</td>
<td>24</td>
<td>Jan Odijk</td>
</tr>
<tr>
<td>Norway</td>
<td>CLARINO</td>
<td>University of Bergen</td>
<td>8</td>
<td>Koenraad De Smedt</td>
</tr>
<tr>
<td>Poland</td>
<td>CLARIN Poland</td>
<td>Wroclaw University of Technology</td>
<td>6</td>
<td>Maciej Piasecki</td>
</tr>
<tr>
<td>Portugal</td>
<td>CLARIN-Portugal</td>
<td>University of Lisbon</td>
<td>22</td>
<td>António Branco</td>
</tr>
<tr>
<td>Slovenia</td>
<td>CLARIN.SI</td>
<td>Jožef Stefan Institute</td>
<td>12</td>
<td>Tomaž Erjavec</td>
</tr>
<tr>
<td>Sweden</td>
<td>SWE-CLARIN</td>
<td>Språkbanken</td>
<td>9</td>
<td>Lars Borin</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Observer</th>
<th>National Consortium (NC)</th>
<th>Leading partner NC</th>
<th>Number partners NC</th>
<th>National coordinator</th>
</tr>
</thead>
<tbody>
<tr>
<td>United Kingdom</td>
<td>CLARIN-UK</td>
<td>Oxford University</td>
<td>12</td>
<td>Martin Wynne</td>
</tr>
</tbody>
</table>
CLARIN in Spain

• Centre K
• Services

http://lod.iula.upf.edu/types/Service
CLARIN ERIC pillars

- CLARIN is a network of different centres that are the pillars of the infrastructure

- The fundamental pillars are
  - **documentation**
    - language resources have to be documented with a standard format of metadata (CMDI = Component MetaData Infrastructure)
  - **exposition** of metadata to be harvested in a centralized meta-catalogue (VLO = Virtual Language Observatory)
  - **actual access** to language resources
    - totally free
    - through a single sign-on unified authentication system (which is particularly important for all those digital language resources that are available for academic purposes only)
CLARIN ERIC governance

• **General Assembly**
  – Managers appointed by the Ministry for Research of each nation
  – highest decision-making body
• **CLARIN Office + Board of Directors**
  – executive management
• **National Coordinators’ Forum**
  – Coordinators of CLARIN National Consortia
  + Standing Committee for Technical Centers
  – integration and interoperability at a trans-national level

https://www.clarin.eu/content/govemance
CLARIN National Consortia

- CLARIN ERIC
  - does not impose a model of governance to its various National Consortia
  - requires
    - the creation of a community by involving the main actors of the research sector
    - the establishment of a centre (or more) to make various types of data, tools and services available
    - the participation in centralized activities, committees etc.
CLARIN Conference 2016: Call for Papers

CLARIN is happy to announce the 5th CLARIN Annual Conference and calls for the submission of papers. The series of CLARIN Annual

Read more >

CLARIN-NL in a nutshell

The CLARIN-NL project is a large national project in the Netherlands (2009-2015) which aims to make a significant contribution to the European CLARIN infrastructure.

The CLARIN infrastructure is a research infrastructure intended for humanities researchers that work with language data and tools. The researcher will be able to find data and tools relevant for his/her research via the CLARIN infrastructure and get access to them. We also aim to make it possible to apply tools to data in such a way that no technical background is needed or ad-hoc adaptations to the tools or the data are necessary.

The researcher’s research project may result in new data and tools, which he/she can make available to other researchers via the CLARIN infrastructure. The CLARIN research infrastructure will thus make humanities research related to language easier, faster, better, and in some cases even possible for the first time.
<table>
<thead>
<tr>
<th>CENTRE</th>
<th>FOCUS ON RESOURCES</th>
<th>RESOURCE EXAMPLES</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEERTENS INSTITUTE</td>
<td>relevant for the study of function, meaning and coherence of cultural expressions and resources relevant for the structural, dialectological and sociolinguistic study of language variation within the Dutch language.</td>
<td>typological databases for Dutch dialects, descriptions of Dutch songs, databases of Dutch names, etc. For more examples, see here</td>
</tr>
<tr>
<td>MPI</td>
<td>related to the study of psychological, social and biological foundations of language.</td>
<td>documentation of endangered languages, resources for sign languages, phonetic resources for the study of phone perception, speech error databases, tools for creating and annotating resources, etc. For more examples, see here</td>
</tr>
<tr>
<td>HUYGENS ING</td>
<td>related to the study of history and literature of the Netherlands.</td>
<td>historical and literary manuscripts and their annotations, tools to annotate such manuscripts and create scholarly editions, etc. For more examples see here</td>
</tr>
<tr>
<td>INL</td>
<td>that are relevant to the lexicological study of the Dutch language and on resources relevant for research in and development of language and speech technology.</td>
<td>lexicons, lexical databases, text corpora, speech corpora, language and speech technology tools, etc. For more examples, see here</td>
</tr>
<tr>
<td>DANS</td>
<td>sustained access to digital research data. The DANS data archive contains thousands of datasets in the fields of humanities including oral history, archaeology, geospatial sciences and behavioural and social sciences. The data archive has acquired the Data Seal of Approval. Regarding the conditions under which access to the data is granted, the DANS motto is “Open Access when possible, Restricted Access when necessary”.</td>
<td>text resources, databases, spreadsheets, audio and audio-visual resources. For more examples see here</td>
</tr>
<tr>
<td>CENTRE</td>
<td>FOCUS ON RESOURCES</td>
<td>RESOURCE EXAMPLES</td>
</tr>
<tr>
<td>--------------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>MEERTENS INSTITUTE</td>
<td>relevant for the study of function, meaning and coherence of cultural expressions and resources relevant for the structural, dialectological and sociolinguistic study of language variation within the Dutch language.</td>
<td>typological databases for Dutch dialects, descriptions of Dutch songs, databases of Dutch names, etc. For more examples, see <a href="#">here</a>.</td>
</tr>
<tr>
<td>MPI</td>
<td>related to the study of psychological, social and biological foundations of language.</td>
<td>documentation of endangered languages, resources for sign languages, phonetic resources for the study of phone perception, speech error databases, tools for creating and annotating resources, etc. For more examples see <a href="#">here</a>.</td>
</tr>
<tr>
<td>HUYGENSING</td>
<td>related to the study of history and literature of the Netherlands.</td>
<td>historical and literary manuscripts and their annotations, tools to annotate such manuscripts and create scholarly editions, etc. For more examples see <a href="#">here</a>.</td>
</tr>
<tr>
<td>INL</td>
<td>that are relevant to the lexicological study of the Dutch language and on resources relevant for research in and development of language and speech technology.</td>
<td>lexicons, lexical databases, text corpora, speech corpora, language and speech technology tools, etc. For more examples see <a href="#">here</a>.</td>
</tr>
<tr>
<td>DANS</td>
<td>sustained access to digital research data. The DANS data archive contains thousands of datasets in the fields of humanities including oral history, archaeology, geospatial sciences and behavioural and social sciences. The data archive has acquired the Data Seal of Approval. Regarding the conditions under which access to the data is granted, the DANS motto is “Open Access when possible, Restricted Access when necessary”.</td>
<td>text resources, databases, spreadsheets, audio and audio-visual resources. For more examples see <a href="#">here</a>.</td>
</tr>
<tr>
<td>CENTRE</td>
<td>FOCUS ON RESOURCES</td>
<td>RESOURCE EXAMPLES</td>
</tr>
<tr>
<td>--------------</td>
<td>------------------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>MEERTENS INSTITUTE</td>
<td>relevant for the study of function, meaning and coherence of cultural expressions and resources relevant for the structural, dialectological and sociolinguistic study of language variation within the Dutch language.</td>
<td>typological databases for Dutch dialects, descriptions of Dutch songs, databases of Dutch names, etc. For more examples, see <a href="#">here</a></td>
</tr>
<tr>
<td>MPI</td>
<td>related to the study of psychological, social and biological foundations of language.</td>
<td>documentation of endangered languages, resources for sign languages, phonetic resources for the study of phone perception, speech error databases, tools for creating and annotating resources, etc. For more examples, see <a href="#">here</a></td>
</tr>
<tr>
<td>HUYGENS ING</td>
<td>related to the study of history and literature of the Netherlands.</td>
<td>historical and literary manuscripts and their annotations, tools to annotate such manuscripts and create scholarly editions, etc. For more examples see <a href="#">here</a></td>
</tr>
<tr>
<td>INL</td>
<td>that are relevant to the lexicological study of the Dutch language and on resources relevant for research in and development of language and speech technology.</td>
<td>lexicons, lexical databases, text corpora, speech corpora, language and speech technology tools, etc. For more examples, see <a href="#">here</a></td>
</tr>
<tr>
<td>DANS</td>
<td>sustained access to digital research data. The DANS data archive contains thousands of datasets in the fields of humanities including oral history, archaeology, geospatial sciences and behavioural and social sciences. The data archive has acquired the Data Seal of Approval. Regarding the conditions under which access to the data is granted, the DANS motto is “Open Access when possible, Restricted Access when necessary”.</td>
<td>text resources, databases, spreadsheets, audio and audio-visual resources. For more examples see <a href="#">here</a></td>
</tr>
<tr>
<td>CENTRE</td>
<td>FOCUS ON RESOURCES</td>
<td>RESOURCE EXAMPLES</td>
</tr>
<tr>
<td>--------------</td>
<td>------------------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>MEERTENS INSTITUTE</strong></td>
<td>relevant for the study of function, meaning and coherence of cultural expressions and resources relevant for the structural, dialectological and sociolinguistic study of language variation within the Dutch language.</td>
<td>typological databases for Dutch dialects, descriptions of Dutch songs, databases of Dutch names, etc. For more examples, see <a href="#">here</a></td>
</tr>
<tr>
<td><strong>MPI</strong></td>
<td>related to the study of psychological, social and biological foundations of language.</td>
<td>documentation of endangered languages, resources for sign languages, phonetic resources for the study of phone perception, speech error databases, tools for creating and annotating resources, etc. For more examples, see <a href="#">here</a></td>
</tr>
<tr>
<td><strong>HUYGENS INNG</strong></td>
<td>related to the study of history and literature of the Netherlands.</td>
<td>historical and literary manuscripts and their annotations, tools to annotate such manuscripts and create scholarly editions, etc. For more examples see <a href="#">here</a></td>
</tr>
<tr>
<td><strong>INL</strong></td>
<td>that are relevant to the lexicological study of the Dutch language and on resources relevant for research in and development of language and speech technology.</td>
<td>lexicons, lexical databases, text corpora, speech corpora, language and speech technology tools, etc. For more examples, see <a href="#">here</a></td>
</tr>
<tr>
<td><strong>DANS</strong></td>
<td>sustained access to digital research data. The DANS data archive contains thousands of datasets in the fields of humanities including oral history, archaeology, geospatial sciences and behavioural and social sciences. The data archive has acquired the Data Seal of Approval. Regarding the conditions under which access to the data is granted, the DANS motto is “Open Access when possible, Restricted Access when necessary”.</td>
<td>text resources, databases, spreadsheets, audio and audio-visual resources. For more examples see <a href="#">here</a></td>
</tr>
</tbody>
</table>
How it works

- A researcher is interested in the study of syntax
- In particular, he/she is interested in the Spanish language(s)
- First, he/she checks the availability of syntactically annotated corpora (treebanks)
- The central catalogue: VLO
- The search and PML system for querying annotated treebanks
• The system is available at the Lindat CLARIN Centre repository (Czech Republic)
The tool is freely accessible
• Some corpora are freely available
A researcher can access the system with an institutional login if his/her institution is part of the EduGAIN Federation.
For instance, here is the list of the Italian institutions which are part of the EduGAIN Federation:

- Bocconi University
- University of Genoa
- CSP - innovation in ICT
- Università di Ferrara
- UniServity: CAVLE
- IMT Lucca - Identity Provider
- Università di Napoli Parthenope
- Urbino University
- IZS Sicilia

or search for a provider, such as Example University

- Please help, I cannot find my provider

Locate me and show nearby providers

Show providers in Italy show all countries
After accessing the system, the number of available treebanks grows (because not only those ones freely available are shown)
El gobernante, con ganada fama desde que llegó hace 16 meses a el poder de explotar a el máximo su oratoria y acusado por sus detractores de incontintencia verbal, enmudeció desde el momento en el que el Tribunal Supremo de Justicia (TSJ) declaró suspe el 28 de mayo.
Federated identity

- Making it easier for academic users to get access to password-protected language resources is an important goal of CLARIN
- Academic users should be able to login with their existing institutional credentials
- Accessing CLARIN applications becomes as simple as logging in one’s university’s webmail system
- The user’s stores from universities and academic institutions ("Identity Providers") are connected to password-protected web applications ("Service Providers")
  - These trust networks ("Identity Federations") already exist at the national level in the EU countries
  - CLARIN is working at crossing the country borders when logging in, so that e.g. a Spanish researcher can access language resources hosted in Prague
## Participating Identity Federations

We aim at including all European federations, however CLARIN member countries get the highest priority.

<table>
<thead>
<tr>
<th>Country</th>
<th>Federation</th>
<th>Status</th>
<th>Attribute release</th>
<th>Opt-in</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austria</td>
<td>ACoRe</td>
<td>Fully connected</td>
<td>OK (not required)</td>
<td></td>
</tr>
<tr>
<td>Belgium</td>
<td>Belnet</td>
<td>Fully connected</td>
<td>OK (for Antwerp, Gent, Louven)</td>
<td>OK (not required)</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>eduID.cz</td>
<td>Fully connected</td>
<td>OK (not required)</td>
<td></td>
</tr>
<tr>
<td>Denmark, Iceland</td>
<td>WAYF</td>
<td>Fully connected (via eduGAIN and Kalmar Union)</td>
<td>OK</td>
<td>OK (not required)</td>
</tr>
<tr>
<td>Estonia</td>
<td>TAAT</td>
<td>Fully connected (via eduGAIN)</td>
<td>OK</td>
<td>OK (not required, attribute release mandatory)</td>
</tr>
<tr>
<td>Finland</td>
<td>Haka</td>
<td>Fully connected. Also includes Kalmar Union</td>
<td>OK</td>
<td>OK (not required)</td>
</tr>
<tr>
<td>Germany</td>
<td>DFN-AAI</td>
<td>Fully connected</td>
<td>Problems for some IdPs (e.g. Tübingen)</td>
<td>OK (not required)</td>
</tr>
<tr>
<td>Greece</td>
<td>GRNET</td>
<td>Connection planned since 2015</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Federation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Italy</td>
<td>IDEM GARR</td>
<td>Fully connected (via eduGAIN)</td>
<td>OK (not required)</td>
<td></td>
</tr>
<tr>
<td>Lithuania</td>
<td>LITNET fedi</td>
<td>Connection planned since 2015</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Easy access to protected resources

Thanks to a federated login, the following applications and data sets are available to anyone with an academic computer account from many European countries. 

https://www.clarin.eu/content/easy-access-protected-resources

<table>
<thead>
<tr>
<th>Resource or Tool</th>
<th>Description</th>
<th>Provided by</th>
</tr>
</thead>
<tbody>
<tr>
<td>OpenSoNaR</td>
<td>over 500 million word Dutch reference corpus</td>
<td>INL (Instituut voor Nederlandse Lexicologie)</td>
</tr>
<tr>
<td>Corpus Hedendaags Nederlands</td>
<td>written corpus of contemporary Dutch</td>
<td>INL (Instituut voor Nederlandse Lexicologie)</td>
</tr>
<tr>
<td>VU-DNC</td>
<td>diachronic Dutch newspaper corpus</td>
<td>INL (Instituut voor Nederlandse Lexicologie)</td>
</tr>
<tr>
<td>Tündra</td>
<td>treebank search application</td>
<td>Eberhard Karls Universität Tübingen</td>
</tr>
</tbody>
</table>
CLARIN-IT: who does what

• CNR (Department of SSH) is the Representing Entity
• The National Coordinator
  – builds the National Consortium
  – is the interface between the research community and the ERIC
• ILC-CNR is the Executor
  – builds an Italian centre for documenting and accessing language resources and providing services
CLARIN-IT: Web portal

the Italian Common Language Resources and Technology Infrastructure

CLARIN-IT: Services for the Italian community of Humanities and Social Sciences

Ciudad Real, 28/09/2016  Monica Monachini (ILC-CNR)
CLARIN-IT: consortium

- **IDEM-GARR** Federated Identity
- Institutions that have already joined or are about to join CLARIN-IT
  - Siena University
  - EURAC Bozen
  - FBK Trento
  - Cattolica Milano University
  - Venezia University
  - Pisa University
  - Pisa Scuola Normale Superiore
  - Parma University
  - Roma University
CLARIN ILC centre

- Access to language resources and services of the Institute
- Depositing services by third parties
- National collector of data and tools
- Italian language and classical languages
- Long term preservation
- Single sign-on
Users can document and deposit their language resources in the CLARIN ILC repository.

That guarantees:
- availability
- citation
- usability
Step 3: Select type of your submission

You have initiated a new workflow item. In the next few steps you will provide the details about the item and upload content files. First select the type of the resource you are about to submit.

**Fig 4 Submission info**

Click on one of the type buttons e.g. Corpus. Proceed with filling the basic information such as the title. Click Next to continue the following step.

Step 4: Describe your item

In the following two steps you will provide more details for your item. First describe the people, organization and projects involved with the item.
CLARIN ILC centre: harvesting

- CLARIN ILC metadata, as long as the metadata of the other European centres, are harvested in the central VLO
- Language resources are searchable in the central VLO
# CLARIN ILC centre:
---
**Spanish LRs**

<table>
<thead>
<tr>
<th>Author</th>
<th>Texts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Santa Teresa d'Avila</td>
<td>Cartas - Camino Opemin - Fundacio Castillo</td>
</tr>
<tr>
<td>Cristoforo Colombo</td>
<td>Giornale di bordo - Lettere (Ediz. Varela) Lettere (parte I e II - Ediz. De Lollis)</td>
</tr>
<tr>
<td>Pedro Salinas</td>
<td>Poesias completas</td>
</tr>
<tr>
<td>Neruda</td>
<td>Poesie (volume I e II)</td>
</tr>
<tr>
<td>Miguel de Unamuno y Jugo</td>
<td>San Manuel bueno martir y tres histori</td>
</tr>
<tr>
<td>Gabriel García Márquez</td>
<td>Cronica de una muerte anunciada</td>
</tr>
<tr>
<td>Antonio Machado</td>
<td>Raccolta di Poesie</td>
</tr>
<tr>
<td>Juan del Encina</td>
<td>Raccolta delle opere teatrali (Teatro completo)</td>
</tr>
<tr>
<td>El Pais</td>
<td>El servilismo de las ciencias sociales</td>
</tr>
</tbody>
</table>
Query system

- Santa Teresa’s corpus is morpho-syntactically analysed and queried by a dedicated query system [http://dbtvm1.ilc.cnr.it/lema/](http://dbtvm1.ilc.cnr.it/lema/)
- The system will be available soon through different platforms and devices and also provided as a CLARIN service
The corpus can be browsed by lemma or form.
From the form to the context and from the context to the full sentence.
Search for groups of words (logical operators AND, OR, NOT)

### Definizione famiglia di ricerca

Azzerà definizione famiglia

<table>
<thead>
<tr>
<th>Operatori logici</th>
<th>Parole</th>
<th>Famiglie</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;= AND&quot;</td>
<td>1) liberadas librar V</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2) libertad libertad S</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3) libertades libertad S</td>
<td></td>
</tr>
<tr>
<td>&quot;</td>
<td>OR&quot;</td>
<td>4) libertado libertar V</td>
</tr>
<tr>
<td>&quot;</td>
<td>NOT&quot;</td>
<td>5) libertan libertar V</td>
</tr>
<tr>
<td>&quot;- Apertura&quot;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;</td>
<td>Chiusa&quot;</td>
<td></td>
</tr>
</tbody>
</table>

DBT-2000 Interrogazione testi lemmatizzati

Una famiglia si definisce utilizzando il codice degli operatori logici.

I codici sono W = parola, F = famiglia, n = numero d'ordine associato a una parola o a una famiglia.

Gli operatori logici sono & = "AND", | = "OR", | = "AND NOT", ( ) parentesi che definiscono le gerarchie logiche all'interno dell'espressione.

Esempi di definizione di famiglia:

- **W1W2 =** definisce una famiglia di ricerca che individua i contesti in cui compaiono, entro la distanza stabilita, le parole contrassegnate dai numeri 1 e 2 nel Quadro.
- **W1W2 =** definisce una famiglia che individua i contesti in cui compare la parola 1 oppure la parola 2.
- **W1W2|W3 =** individua i contesti in cui compare la parola 3 associata con la parola 1 o con la parola 2.
- **W1W2|W3W4 =** individua i contesti in cui compare la parola 3 associata con la parola 1 e con la parola 2, purché nella

Ricerca famiglia n. 1

libertados libertad S | OR | liberadas liberar V

1. de mirar que no hagan a las almas con sus libertades ? A Y verdaderamente creo que muchas veces es como
   - FUN CAP 07 07.13 Pag.1116.1
2. por el santo concilio la priora no podia dar las libertades que queria .
   - FUN CAP 17 16.32 Pag.1190.3
3. paternalidad; mas como esta criada siempre en las libertades de esta casa, temo mucho. Digame
   - CARTAS 148 03.5 Pag.0399.5
4. por el consentimiento de la Madre y por el Visitador sortiradas de la carcel . Y cualquiera que en esta
   - CONS 55.11 Pag.1626.1

DBT-2000 Interrogazione testi lemmatizzati

DBT-2000
Query system

Some examples

Search for «casa» and «esta»
Query system
Tagger = analisi morfosintattica = analisi grammaticale
<table>
<thead>
<tr>
<th>Articolo determinativo plurale femminile</th>
</tr>
</thead>
<tbody>
<tr>
<td>sostantivo comune plurale femminile</td>
</tr>
<tr>
<td>Verbo stramazzare. 3° pers. Plur. Indicativo imperfetto.</td>
</tr>
<tr>
<td>Preposizione articolata singolare femminile</td>
</tr>
<tr>
<td>sostantivo comune singolare femminile</td>
</tr>
</tbody>
</table>
Parser = analisi sintattica funzionale
Estrazione candidate
Polirematiche, collocazioni, combinazioni lessicali tipiche

Estrazione candidate
Polirematiche, collocazioni, combinazioni lessicali tipiche
<table>
<thead>
<tr>
<th>Nome (Cognome)</th>
<th>Codice Fiscale</th>
<th>CPF</th>
<th>Data di Nascita</th>
<th>Indirizzo</th>
<th>Sesso</th>
<th>Ruolo</th>
<th>Data di Firma</th>
<th>Telephone</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIGNOR=SIGNORE</td>
<td>155 49</td>
<td>12</td>
<td>1960-01-01</td>
<td>PORTA</td>
<td>M</td>
<td>COUNT</td>
<td>2060-01-01</td>
<td>1234567890</td>
<td><a href="mailto:test@example.com">test@example.com</a></td>
</tr>
<tr>
<td>SIGNORA=SIGNORA</td>
<td>87 6</td>
<td>15</td>
<td>1960-01-01</td>
<td>PORTA</td>
<td>F</td>
<td>COUNT</td>
<td>2060-01-01</td>
<td>1234567890</td>
<td><a href="mailto:test@example.com">test@example.com</a></td>
</tr>
<tr>
<td>COMMISSARIO=SIGNOR</td>
<td>155 10</td>
<td>32</td>
<td>1960-01-01</td>
<td>PORTA</td>
<td>M</td>
<td>COUNT</td>
<td>2060-01-01</td>
<td>1234567890</td>
<td><a href="mailto:test@example.com">test@example.com</a></td>
</tr>
<tr>
<td>COMMISSARIO=SIGNOR</td>
<td>155 10</td>
<td>32</td>
<td>1960-01-01</td>
<td>PORTA</td>
<td>M</td>
<td>COUNT</td>
<td>2060-01-01</td>
<td>1234567890</td>
<td><a href="mailto:test@example.com">test@example.com</a></td>
</tr>
</tbody>
</table>

**Nota:**
- **CPF** sta per Codice Fiscale Patrimoniale.
- **RUOLI** e **SESSO** sono dati personali, potenzialmente sensibili.
- **EMAIL** e **TELEFONO** sono dati di contatto, potenzialmente sensibili.

**Lista candidati, formato testo a campi tabbati**

**Ciudad Real, 28/09/2016**

**Monica Monachini (ILC-CNR)**

69
L'amore. Suerde il vero osi, o regina, io l'oso.
A te recar, poiché il tuo popol fico
Mi tien da lonto, e poiché al soglio intorno
Non è chi voglia a ardita dirlo. In seno
Pian per, cui non son esca umani affetti,
Ma che tanta arde in Dio, libera io nutro.

Non lieve impusso è la licenza vostra
A te sgradito,
Schiavo Argo, o nemico, a me? Che parlì?
E, da te lungi,
Tuoi versi senso interpretar mai due;
E men tu i suoi.

Lungi da me chi l'tene?
Da te in bando l' ten fredda accogienza,
E che? noi desti
Or per anco all'oblio? Straniero vile, in soverchia potei saltò, ea spiacque
Al tuo consorto, e al popol tuo...

Ma farsi
Io tei credo: ma basta. Or degio in breve
Del voce ciclo instigator mendaci,
D'empia setta mintri, udro sempre io
Il favellor nostro appresso? - ah' questo.
Ciudad Real, 28/09/2016
Monica Monachini (ILC-CNR)

Versione elettronica di testo patrimonio dell'Istituto di Linguistica Computazionale

Consoora in formato XML - TEI curata dall'Istituto di Linguistica Computazionale "Antonio Zampolli" CNR-Pisa

Monica Monachini (ILC-CNR)
Another important activity of CLARIN
- user involvement
- facilitation of collaborations
- dissemination
- education to the use of resources and services
- organization of events, seminars
- promotion of exchanges
CLARIN: networking

• CLARIN Annual Conference (Aix-en-Provence, 26-28/10/2016)
• CLARIN PLUS Workshops
• Mobility grants
• Participation in research projects
• Interaction with other infrastructures and networks (DARIAH, …)
CLARIN-IT: survey

- Collecting requirements and desiderata of the community
- Go to the survey
Virtual Language Observatory

From here you can explore the world of language resources and technology from different perspectives:

Go to the Virtual Language Observatory, the CLARIN Language Resource facet browser (with CMDI and OLAC metadata)

- [https://vlo.clarin.eu](https://vlo.clarin.eu)
- Demo
• Italy joined CLARIN
• My message today is about the aims and functioning of CLARIN, its potentialities, what is already in place, what was planned, how it can help advancing the sector
• Through CLARIN you can
  – appropriately document your resources
  – give them international visibility
  – find language resources and combine them with tools for further analysis
  – be in touch with scholars in the field
  – be informed about relevant events
  – apply for mobility grant to visit colleagues for specific joined research
  – be invited at CLARIN related events on strategic themes
  – take part to the CLARIN Annual Conference
• For the research community it is a great opportunity to let CLARIN-IT grow on the basis of concrete needs and expectations
CLARIN-IT: contacts

- **Web:** [http://www.clarin-it.it](http://www.clarin-it.it)
- **Twitter:** [@CLARIN_IT](http://twitter.com/CLARIN_IT)
- **eMail:** [coordination@clarin-it.it](mailto:coordination@clarin-it.it)
Thank you for your attention!