CLARIN-IT

the Italian Common Language Resources and Technology Infrastructure

1° October 2015

CLARIN-IT: CNR Leading institution

- CLARIN-IT Representative body (in the General Assembly):

  Dipartimento Scienze Umane e Sociali, Patrimonio Culturale – Consiglio Nazionale delle Ricerche (DSU-CNR)

  Prof Gilberto Corbellini, Director of Dept.

- CLARIN-IT National Coordinator:

  Monica Monachini, Institute for Computational Linguistics "A. Zampolli" - Italy (ILC-CNR)

- CLARIN-IT Executive Centre

  Institute for Computational Linguistics "A. Zampolli" - National Research Council of Italy (ILC-CNR)

- CLARIN-IT Italian technological node: ILC4CLARIN

CLARIN-IT: consortium first nucleus

Members, Institutions:

- Institute for Applied Linguistics, Eurac Research, IAL, Bozen.. Dott. A. Abel
- Siena University
  1. Dipartimento di Scienze della Formazione, Scienze Umane e della Comunicazione Interculturale,. Prof. S. Calamai
  2. Centro di Studi Comparati, Dipartimento di Filologia e Critica delle Letterature Antiche e Moderne.. Prof. F. Stella
- Fondazione Bruno Kessler, Trento;. Dott. S. Tonelli

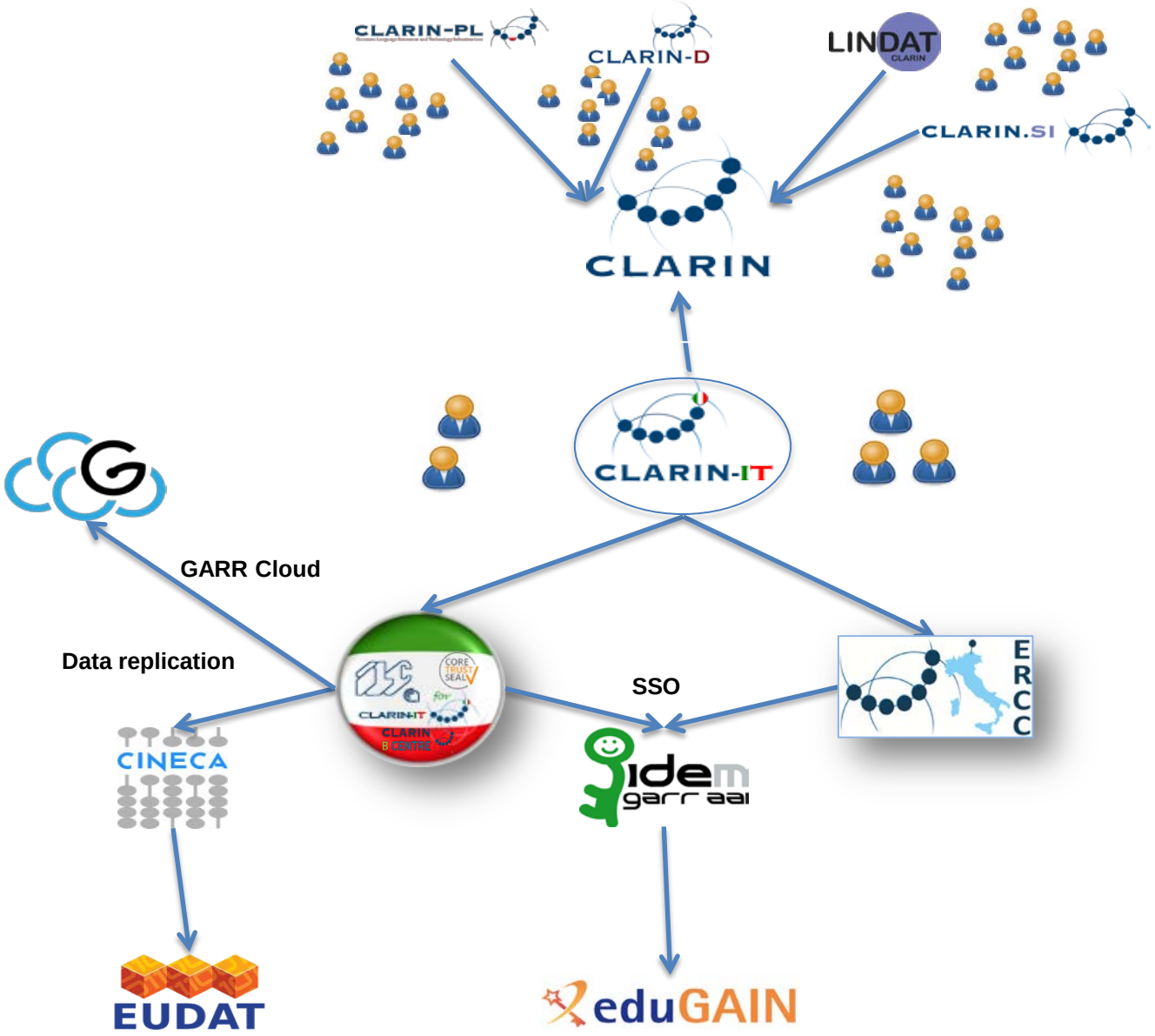
Members, Person:

- Dipartimento di Discipline Umanistiche, Sociali e delle Imprese Culturali, University of Parma, Prof. Anika Nicolosi

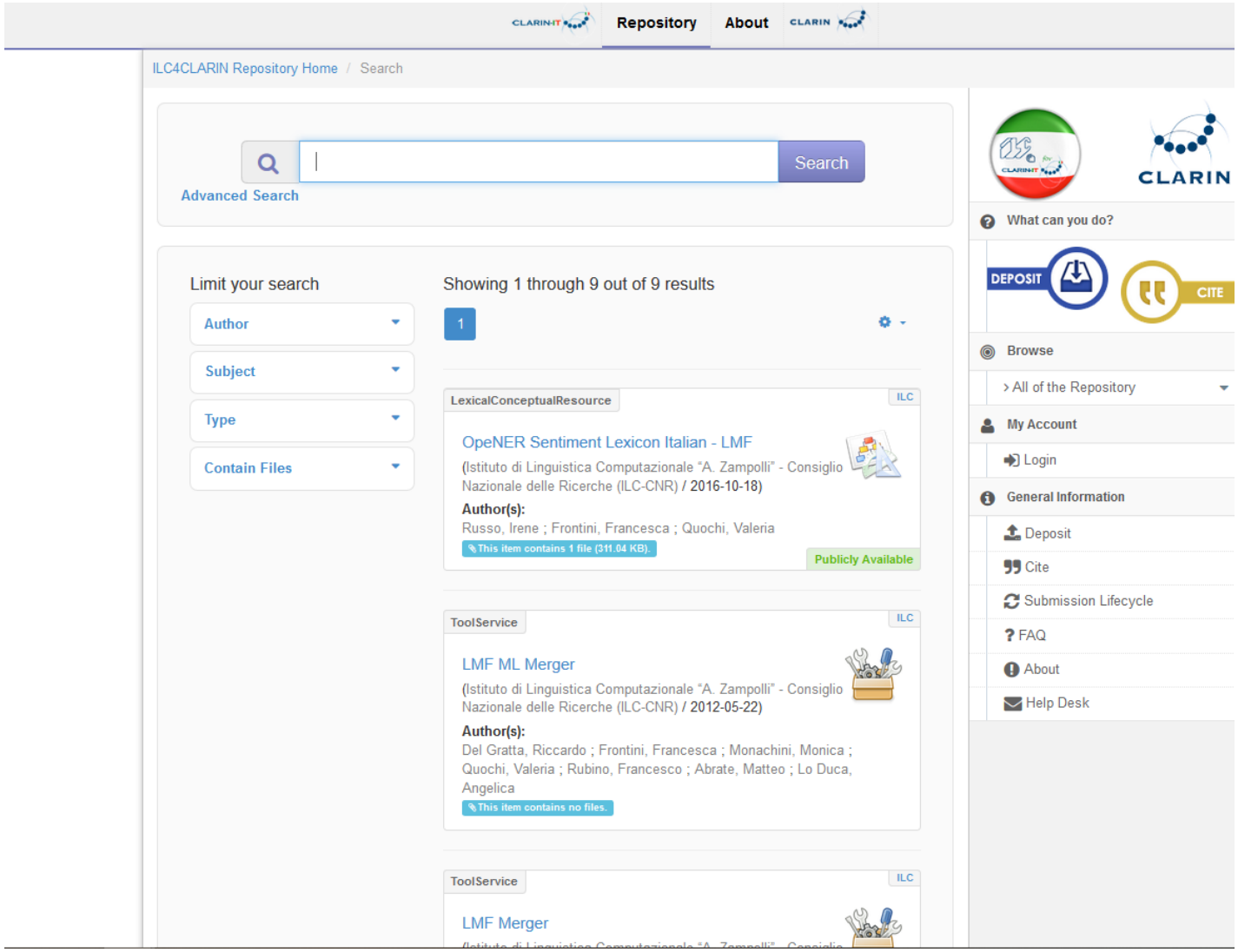
Expression of interest for membership:

- Pisa University; Prof A. Lenci
- Dipartimento di Studi Linguistici e Culturali Comparati, Università "Ca' Foscari" Venezia, Prof. Marina Buzzoni
- Scuola Normale Superiore

CLARIN-IT Technology Infrastructure



ILC4CLARIN repository



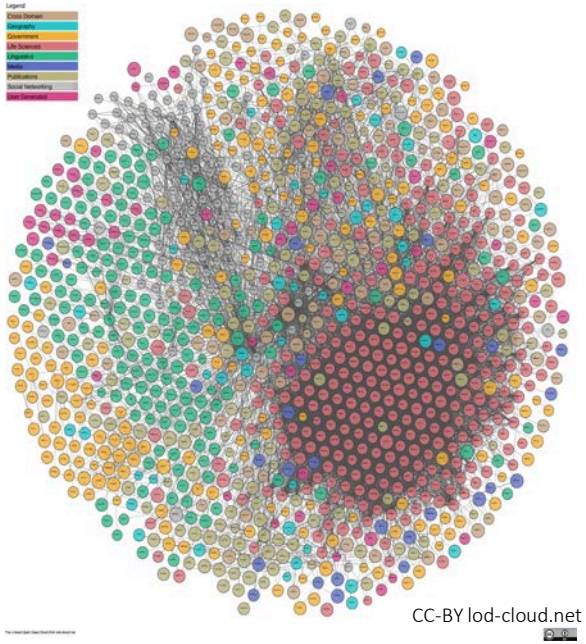
ILC4CLARIN goes LLOD

From CLARIN Component Metadata to Linked Open Data

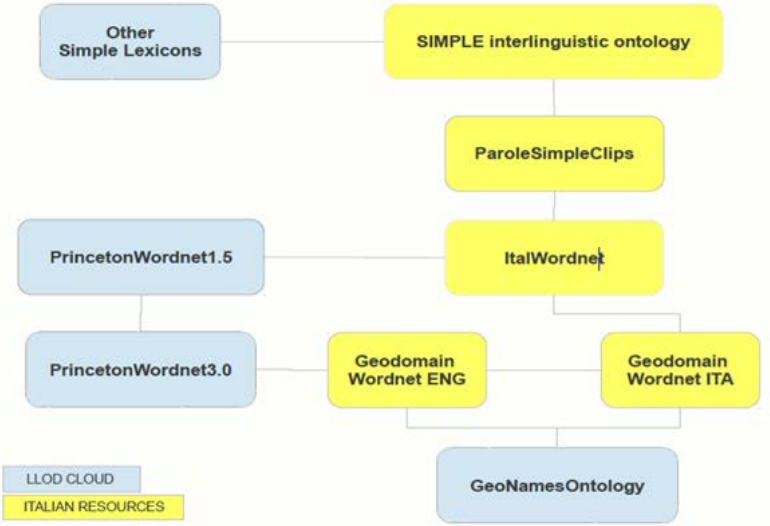
Matej Durco
Institute for Corpus Linguistics and Text Technology
matej.durco@oeaw.ac.at

Menzo Windhouwer
The Language Archive - DANS
menzo.windhouwer@dans.knaw.nl

LDL@LREC2014
Reykjavik, Iceland



CC-BY lod-cloud.net



LLOD and scholarly dictionaries: The Middle Liddell-Scott

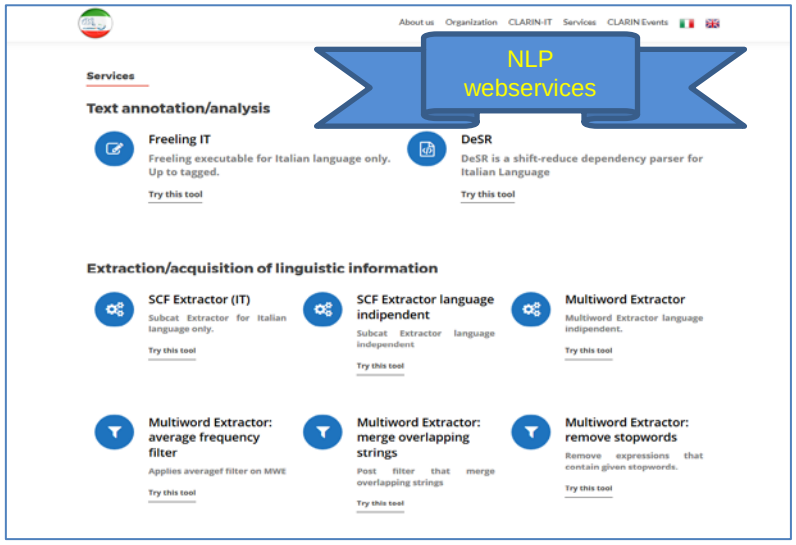
```
:mlEntry_n57 a ontolex:LexicalEntry ;
  ontolex:canonicalForm :mlEntry_n57_lemma ;
  ontolex:sense :mlsense_n57_0,
    :mlsense_n57_1 .

:mlsense_n57_0 a ontolex:LexicalSense ;
  lsj-lemon:senseGloss """immortal, divine, holy, νὺξ ἄβρότη, either holy Night, as a divinity,
(like ἱερὸν κνέφας, ἱερὸν ἦμαρ), or never failing (like ἄφθιτος ἡώς), Il.; ἕπη ἄβροτα holy hymns, Soph."""@en ;
  lsj-lemon:usage <http://dbpedia.org/resource/Iliad>,
    <http://dbpedia.org/resource/Sophocles> ;
  polyLemon:senseSibling :mlsense_n57_0,
    :mlsense_n57_1 ;
  lexinfo:translation "divine"@en,
    "holy"@en,
    "immortal"@en,
    "never failing"@en ;
  ontolex:isSenseOf :mlEntry_n57 .
```

Khan et al. 2017

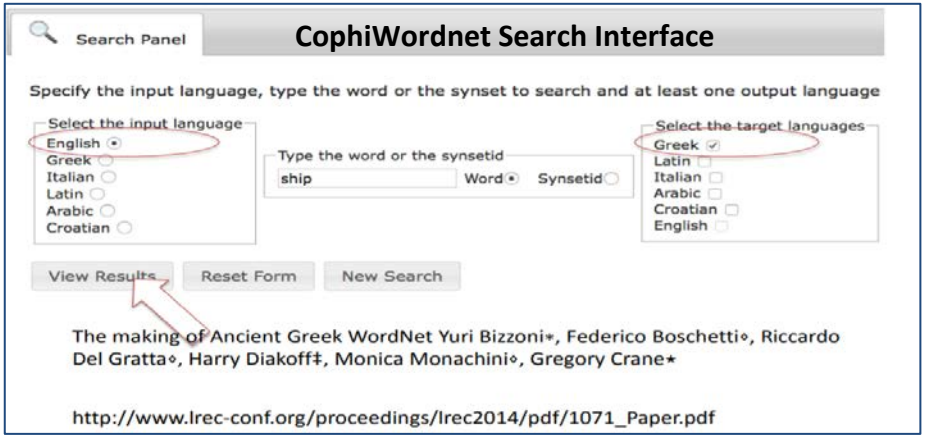
ILC4CLARIN services

Access to Text resources

Access to Lexical resources



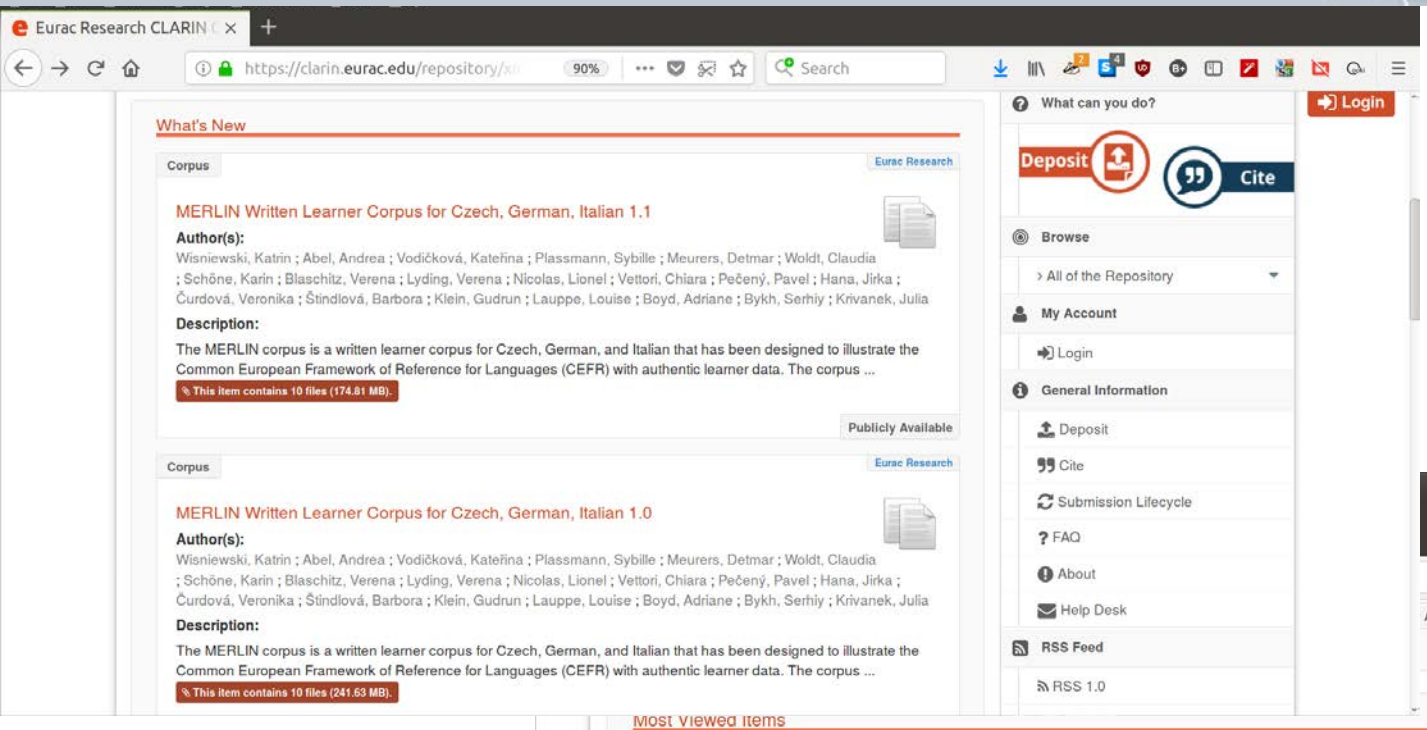
NLP webservices



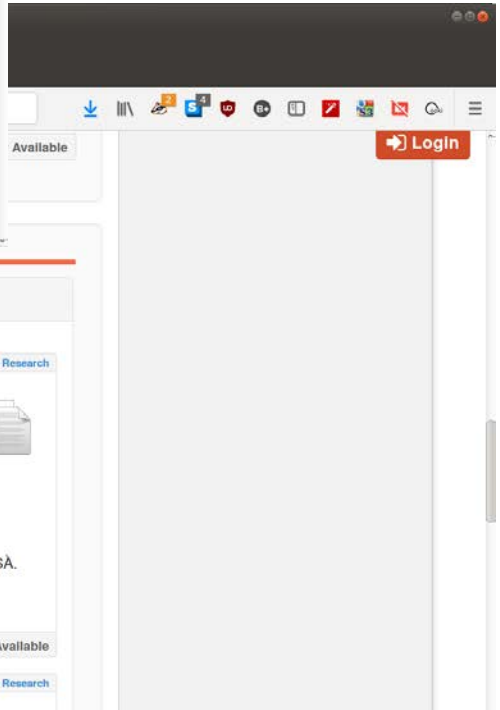
ERCC –
Institute for Applied Linguistics (IAL)
Eurac Research Bozen

- The EURAC Research CLARIN Centre (ERCC) the IAL of EURAC Research
- hosts corpora mainly on Italian or the regional variant of German.
- has strong focus on the region and especially on the local multilinguality
- has a research focus on non-standard language

ERCC –
The CLARIN-IT data center



learner corpora



corpora on computer-mediated communication

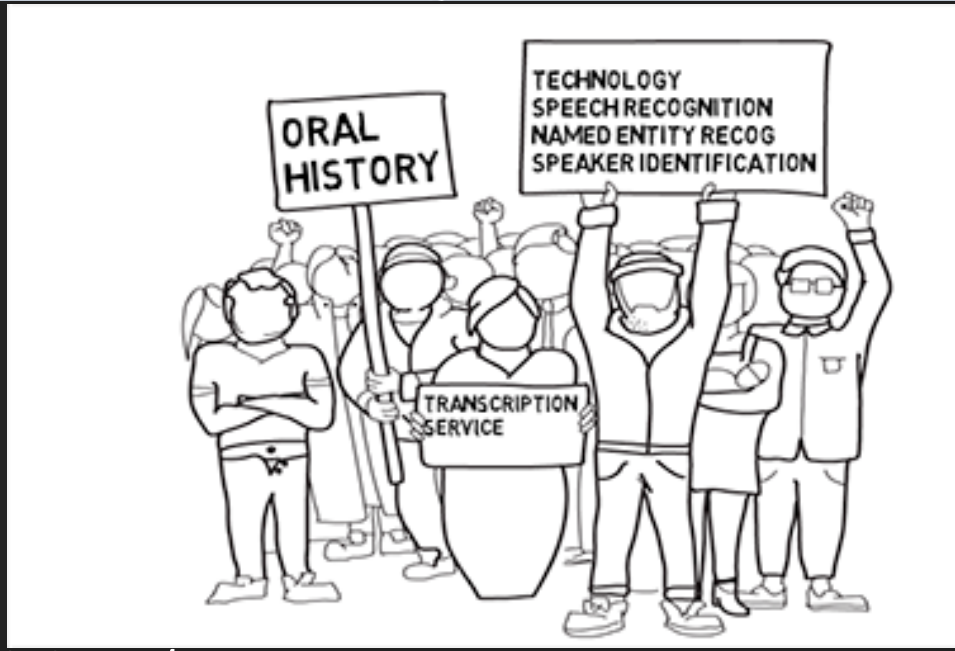
UniSI – Dept. Education and Human Sciences DSFUCI

- Grammo-foni (Gra.fo) project: DSFUCI (Silvia Calamai) & Scuola Normale Superiore Pisa (P.M. Bertinetto)
- a unique speech and oral archive of the past in the Italian panorama:
  - over 3,000 hours of digitized recordings and
  - vast range of documents and topics covered
- Gra.fo provides a first-hand documentation of Tuscan language varieties from the early 1960s
- a model for other research groups to deal with oral archives.

UniSI – Dept. Education and Human Sciences DSFUCI

- ✓ CLARIN-IT supports the use and take up of resources and technologies among social scientists, oral historians and digital humanists

- ✓ Silvia Calamai has participated in and organized a CLARIN Workshops series

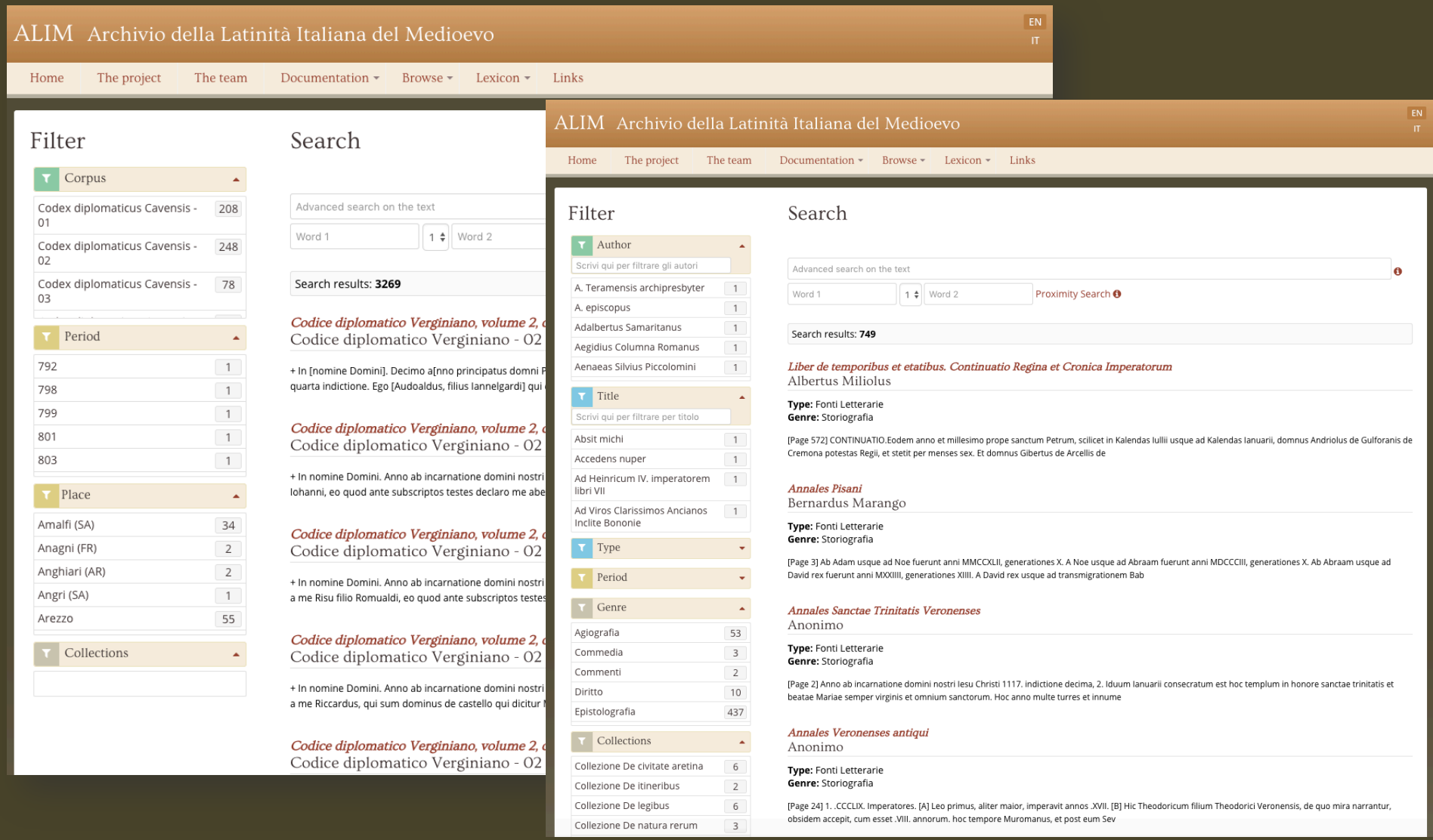


- ✓ Oxford 2016
- ✓ Utrecht 2016
- ✓ Arezzo 2017 - http://oralhistory.eu/workshops/arezzo
- ✓ Munich 2018 - http://oralhistory.eu/workshops/munich

- ✓ lower barriers and diversity of practices in the two communities
- ✓ create automatically generated transcripts.
- ✓ develop a 'Transcription Chain' prototype bringing together CLARIN technologies for speech retrieval and alignment from different countries

UniSI –
Dept. Phylology and Literary Criticism
Centro Studi Comparati DFCLAM

- Free on-line access to digital archives of literary and historical texts
- The Archive of the Italian Latinity of the Middle Ages (ALIM), the largest digital library in Italy
- Latin never lost its role and prestige as transnational language – until the end of the Middle Ages and beyond
- Resp. Prof. Francesco Stella
- http://en.alim.unisi.it



- Latin texts and documents, encoded in XML-TEI
- textual analysis tools and a medieval-Latin lemmatizer

FBK – Fondazione Bruno Kessler
Digital Humanities Unit

The “Digital Humanities” Unit (resp. S. Tonelli) mainly focuses on the development and implementation of techniques for semantic analysis of texts.

They make available:

- annotated corpora of historical texts and political texts (Italian and English)
- NLP software for italian
- Text digitalization and trascription tools
- web tools for advanced navigation of texts

The field of application is:

- study of language complexity
- automated text simplification.

The target community:

- political scientists, history scholars, education and cultural heritage institutions

FBK – Fondazione Bruno Kessler
Digital Humanities Unit

Italian text tool

Simpatico

DH DIGITAL HUMANITIES

FONDAZIONE BRUNO KESSLER

Original text

DICHIARA

1) di presentare la presente richiesta in qualità di: □ Proprietario □ Altro titolo (specificare) 2) che l'intervento da realizzare rientra tra quelli previsti dall'art 64, comma 4, della L.P. 15/2015 e ss.

mm. e precisamente si tratta di:

□ lettera a) - recinzioni

□ lettera b) - muri di sostegno e di contenimento fino a tre metri di altezza

□ lettera c) - pavimentazioni stradali

□ lettera d) - interventi previsti dall'art 78, comma 2, lettera b) (interventi volti all'eliminazione delle barriere architettoniche che non comportano la realizzazione di nuovi volumi esterni all'edificio o comunque la modificazione della sagoma dell'edificio);

□ lettera d) - interventi previsti dall'art 78, comma 2, lettera d) (opere di pavimentazione e di finitura degli spazi esterni nelle aree pertinenziali degli edifici, comprese le sistemazioni del terreno dell'area pertinenziale che non comportano modificazioni delle quote superiori a 50 cm.

di altezza, non incidono sugli indici urbanistici dell'area e risultano raccordate alle quote dei terreni adiacenti il perimetro dell'area);

□ lettera d) - interventi previsti dall'art 78, comma 2, lettera g) (strutture mobili e attrezzature installate per lo svolgimento di manifestazioni culturali, sportive, religiose e simili di carattere temporaneo);

□ lettera d) - interventi previsti dall'art 78, comma 2, lettera h) – appostamenti di caccia realizzati secondo le disposizioni provinciali vigenti in materia di protezione della fauna selvatica ed esercizio della caccia, realizzati in difformità rispetto ai criteri e alle tipologie approvati dalla sottocommissione della CUP con riferimento alle relazioni con il contesto, alle forme e ai materiali da impiegare nella realizzazione;

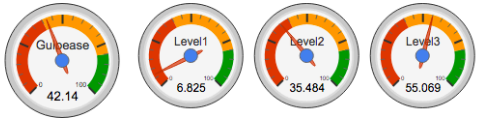
□ lettera d) - interventi di manutenzione straordinaria previsti dall'art 78, comma 3, lettera a) (le opere di manutenzione straordinaria, quando non riguardano le parti strutturali dell'edificio), che riguardano parti esterne dell'edificio, fermo restando quanto stabilito dall'art 78, comma 3, lettera b);

□ lettera d) - interventi previsti dall'art 78, comma 3, lettera c) (installazione di pannelli solari o fotovoltaici e dei relativi impianti collocati negli edifici o nelle relative pertinenze, nel rispetto dei criteri e dei limiti stabiliti dal regolamento urbanistico-edilizio provinciale);

□ lettera d) - interventi previsti dall'art 78, comma 3, lettera d) (legnaie pertinenziali degli edifici che rispettano le tipologie e i limiti dimensionali stabiliti dal PRG);

Sentences:	24
Tokens:	914
Words:	723
Content words:	434

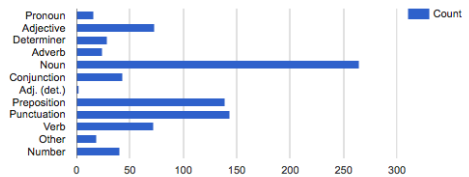
Difficulty levels



These numbers are normalized values showing how difficult is the comprehension of a text depending on the user (from 0 to 1, 0 meaning difficult, 1 meaning easy).

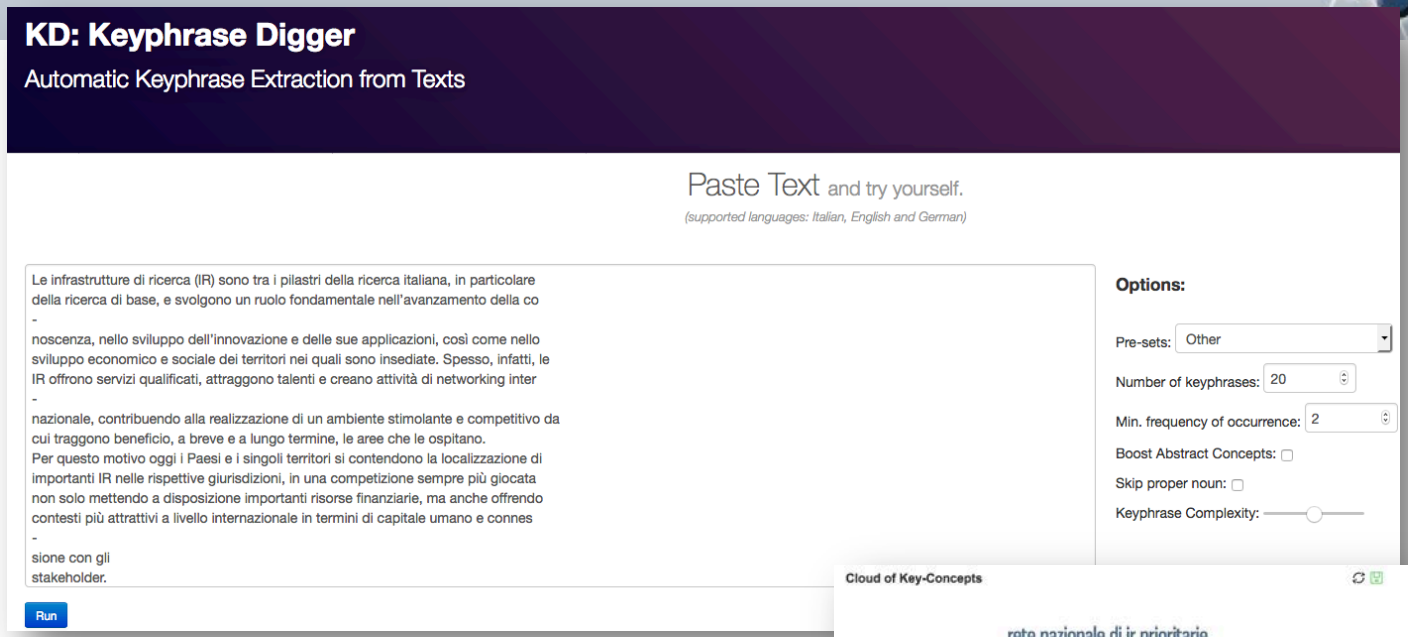
- The **gulpease** is a readability index for the Italian language. It does not take into consideration the content of the text, but technical information such as the length of sentences and words:
  - less than 80 means difficult for a person at elementary school;
  - less than 60 means difficult for a person at medium school;
  - less than 40 means difficult for a person at high school.
- The **Level1** gauge shows how difficult the text is for a user knowing only the 500 easiest words in Italian (elementary school).
- The **Level2** gauge shows how difficult the text is for a user knowing only the 2500 easiest words in Italian (middle school).
- The **Level3** gauge shows how difficult the text is for a user knowing only the 5000 easiest words in Italian (high school).

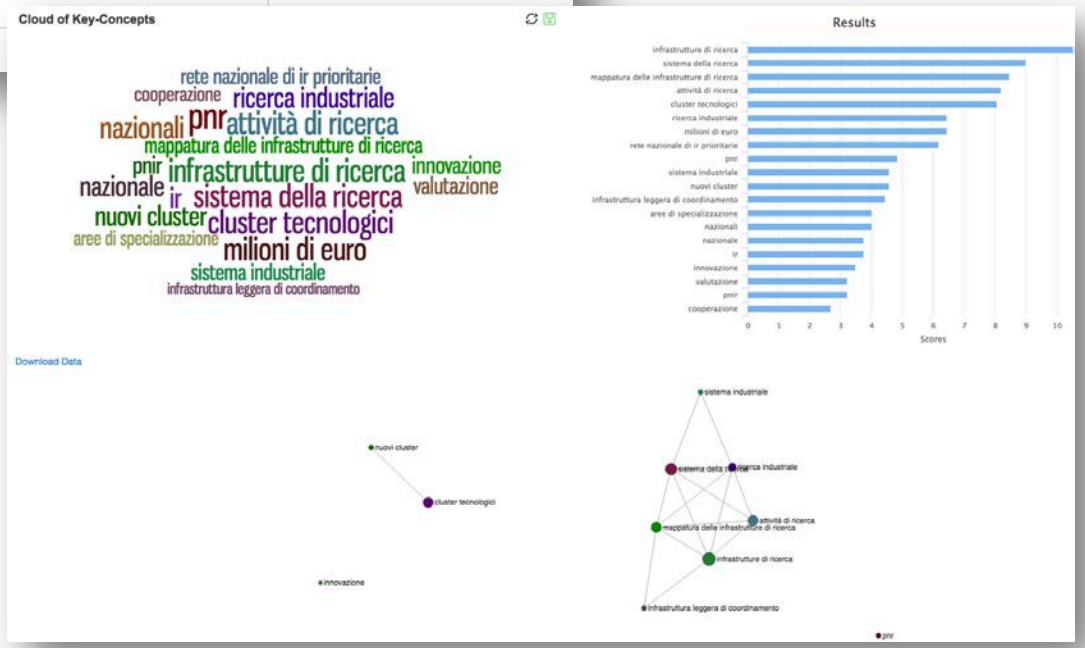
Part-of-speech distribution



A platform for readibilty and simplification

FBK – Fondazione Bruno Kessler
Digital Humanities Unit





Terninology extration and visualization

CLARIN-IT User Involvement

- A. Nicolosi (pers. member) Dept. of Human and Social Sciences – UniParma
- Interest for the use of digital resources and tools in the Humanities
- Survey to ascertain the current interest in digital methods by classical philologists
- The survey focused on the digital resources and tools needed to support an excellent digital edition of an ancient text
- MasterDegree thesis

- CLARIN-IT contribution to the wider impact of CLARIN on Italian users
- adoption of Digital Humanities methods in the field of Classics and ancient world studies

KPI – CLARIN-IT Events

- Disseminate results, involve users
- Introduce CLARIN to next generations, (Keynotes, Lectures)
- Participation in events (AIUCD, LREC, GARR, Workshops)
- Organization of events

Thematic Workshops	10
Workshops for National Coordinators	5
Annual Conferences (Scientific Committee)	4
General Assembly	3
Annual meetings Technical Coordination Committees	3
Technical Committees (CLARIN-IT members in committeees: standard, metadata, legal)	9
Organization of International Workshops	3
Organization of CLARIN-IT events (user involvement, raising awareness in Academia & research inst.)	19
Organization of the Annual Conference 2018 in Italy (CLARIN-IT Local Organizer)	1

KPI – CLARIN-IT Publications

Pubblicazioni CLARIN-IT	Rivista	2
	Atti	16
Pubblicazioni correlate a CLARIN-IT	Rivista	3
	Atti	12
	Volume	2

CLARIN-IT: the message of our National Delegate

- Dear Colleagues attending the CLARIN Annual Conference 2018, I regret not attending your annual conference, but I am overwhelmed with commitments and am negotiating several important matters at the CNR for which my presence in Rome is essential.
- …
- The qualitative leap will be one of the topics to which **the CNR - and the DSU in particular - intends to apply by investing in digital infrastructures for SSH, and guaranteeing support to CLARIN, DARIAH and E-RIHS,** with a view to optimizing resources first, integration of skills and interoperability of data, tools and services available to the research community in humanities, social sciences and heritage science.
- …
- **CLARIN-IT plays a central role in the national system of Research Infrastructures being one of the priority identified by the Italian RoadMap (PNIR) and ensuring the participation of the Italian sector research system in several European projects including E-LEXIS, PARTHENOS and SSHOC …**
- ..
- **The strategic aim of the CNR-DSU is to create the conditions at the national level to ensure the Italian SSH research system a central role** in the process of establishing the European Open Science Cloud, **enhancing the nodes of the European infrastructures present on our territory and enhancing the skills of researchers**, promoting methodological innovation in SSH research and the development of tools and services for cultural and creative industries.
- …
- The sustainability of this system - in the medium to long term - is **ensured by financial hedges from various national and European funding sources (H2020, ESIF, EOSC, PON, POR, etc.), such as the SSHOC Cluster** - a tool to consolidate space of the human and social disciplines in the EOSC - which are flanked by the National Operational Programs and other forms of investment at national and local level, aimed at strengthening the Italian infrastructure network.

http://www.clarin-it.it/it

Monica Monachini – *CLARIN Italian National Coordinator*

ILC-CNR – National Executor

https://ilc4clarin.ilc.cnr.it/

Alessandro Enea – *Data Center*

Paola Baroni – *Web, Dissemination, Membership Procedures*

Riccardo Del Gratta – *Repository*

Sebastiana Cucurullo – *Metadata*

Valeria Quochi – *User Involvement*

coordination@clarin-it.it

communication@clarin-it.it