

CLARIN-IT

the Italian Common Language Resources and Technology Infrastructure

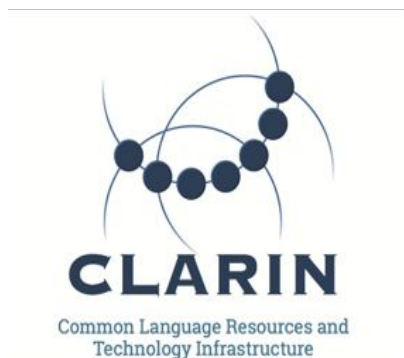


CLARIN ERIC and CLARIN-IT

Francesca Frontini - *CLARIN Board of Directors*

Monica Monachini - *National Coordinator CLARIN-IT*

Istituto di Linguistica Computazionale - ILC CNR

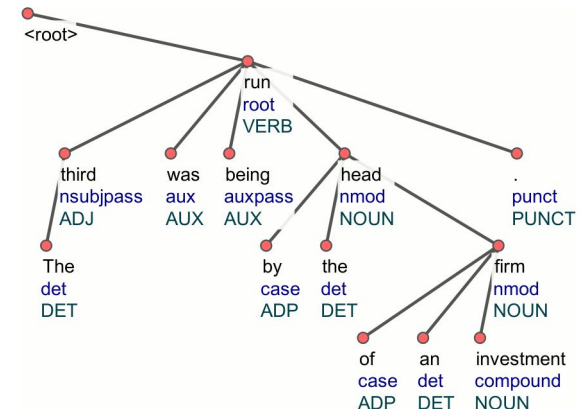


- What is CLARIN ERIC?
- What does CLARIN-IT offer?
- Some exercises to showcase CLARIN services



CLARIN

Common Language Resources and
Technology Infrastructure



CLARIN in a nutshell



- has the **ESFRI ERIC** status since 2012, **Landmark** since 2016
- provides easy and sustainable access for scholars in the **humanities and social sciences** and beyond
 - to digital language data (in written, spoken, video or multimodal form)
 - and **advanced tools** to discover, explore, exploit, annotate, analyse or combine them, wherever they are located
 - through a **single sign-on** environment
- serves as an ecosystem for **knowledge sharing**
- is an integral part of **the European Open Science Cloud**
 - See clarin.eu/eosc

CLARIN data and communities



- Newspaper archives
- Literary texts
- Parliamentary records
- Literary texts
- Historical letters
- Broadcast archives
- Oral History data
- Social Media data
- L-2 Learner Resources
- Survey data
- ...

For the CLARIN Resource Families initiative, see:
<https://www.clarin.eu/resource-families>

- Digital humanities
- Linguistics and Philology
- Translation and Lexicography
- Literary Studies
- History
- Political and Social Sciences
- Media Studies
- Culture, Folklore, Anthropology
- Speech therapy
- General Public
- ...

CLARIN today



- **21 members:** (AT, BG, CY, CZ, DE, DK, EE, FI, GR, HR, HU, IS, IT, LT, LV, NL, NO, PL, PT, SE, SI)
- **3 observers:** FR, UK, ZA
- **> 68 centres (25 CTS certified data centres)**



federated login



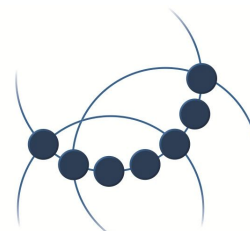
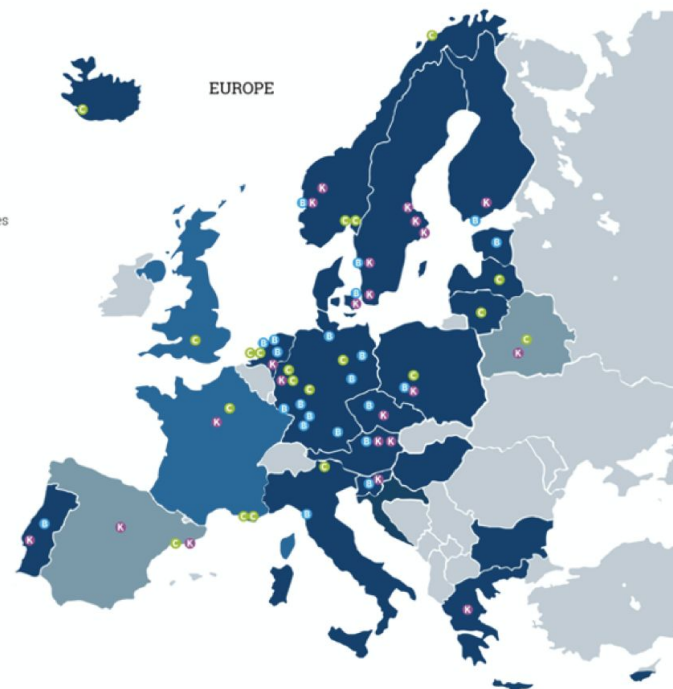
central harvesting of metadata



chained services



- ERIC members
- Observers
- Countries with participating centres
- Centre Providing Data
- Centre Providing Metadata
- Knowledge Centre



CLARIN

Common Language Resources and
Technology Infrastructure



CLARIN-UK

Infrastructure for Digital Language Resources and Tools

[Home](#)[Centres](#)[Resources](#)[People](#)[News](#)[Events](#)[Courses](#)[Blog](#)[About](#)

[Read more about CLARIN-UK](#)

Data & Tools

[More](#)

CorCenCC

Corpws
Cenedlaethol
Cymraeg
Cyfoes – the
National Corpus
of
Contemporary
Welsh

Latest News

[More](#)

New
members of
the
CLARIN_UK
consortium

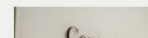
18 November
2020
Welcome
aboard!

Events

[More](#)

Lancaster
Symposium
on Innovation
in Corpus
Linguistics
2021

Wednesday 23
June, Online



Corpus

<https://www.clarin.ac.uk/>



CLARIN-UK

Infrastructure for Digital Language Resources and Tools

[Home](#)[Centres ▾](#)[Resources ▾](#)[People](#)[News](#)[Events](#)[Courses](#)[Blog](#)[About](#)[#LancsBox](#)[BNC](#)[CKLD](#)[CLAWS](#)[CLiC](#)[CorCenCC](#)[CQPweb](#)[ELAR](#)[GATE](#)[GATE Cloud](#)[Hansard at
Huddersfield](#)[Read more about CLARIN-UK](#)

Data & Tools

[More](#)

CorCenCC

Corpws
Cenedlaethol
Cymraeg
Cyfoes – the
National Corpus
of
Contemporary
Welsh

[More](#)

Events

[More](#)

Lancaster
Symposium
on Innovation
in Corpus
Linguistics
2021

Wednesday 23
June, Online



Corpus
Linguistics

<https://www.clarin.ac.uk/>

CLARIN Resource families



Corpora

- Computer-mediated communication corpora
- Corpora of academic texts
- Historical corpora
- L2 learner corpora
- Literary corpora
- Manually annotated corpora
- Multimodal corpora
- Newspaper corpora
- Parallel corpora
- Parliamentary corpora
- Reference corpora
- Spoken corpora

Lexical Resources

- Lexica
- Dictionaries
- Conceptual Resources
- Glossaries
- Wordlists

Tools

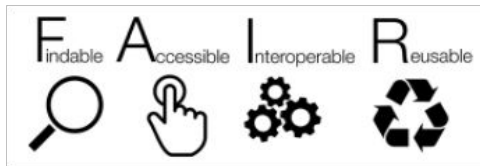
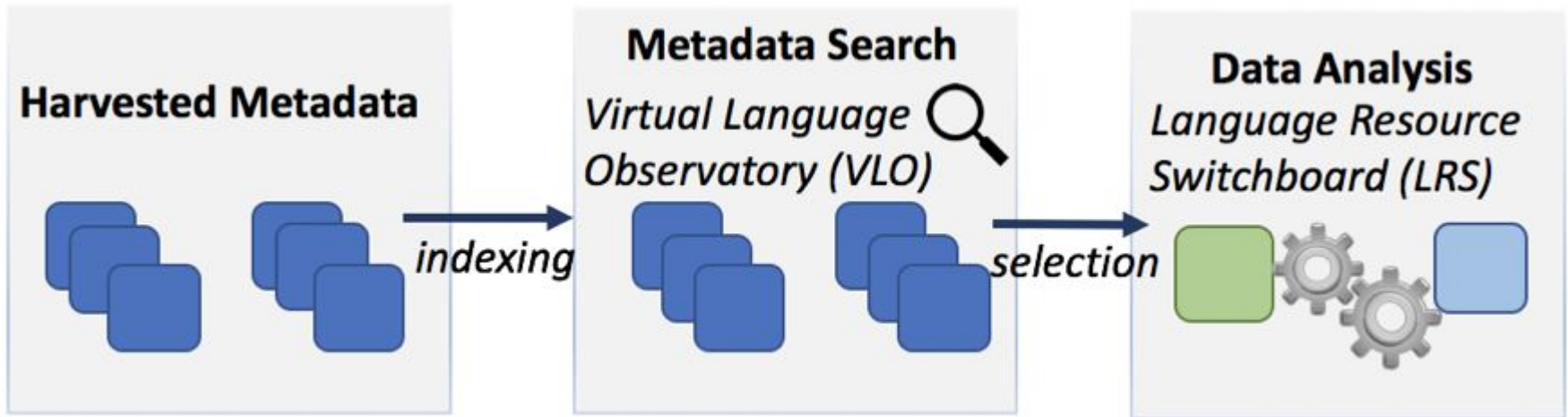
- Normalization
- Named entity recognition
- Part-of-speech tagging and lemmatization
- Tools for sentiment analysis

Spoken corpora in the CLARIN infrastructure

Corpora with transcriptions and audio recordings

Corpus	Language	Description	Availability
Arabic Speech Corpus Licence: CC BY 4.0	Arabic	The corpus is available for download from a dedicated webpage. For a relevant publication, see Halabi (2016) .	Download
DIALEKT v1: dialectal corpus with multi-tier transcription Size: 100,000 words Annotation: orthographically and phonetically (dialect features) transcribed, MSD-tagged, lemmatised Licence: Academic Licence Agreement for Czech National Corpus Data	Czech	This corpus contains traditional dialectological material, mostly unprepared monologue-type speech. The corpus is available download (upon request) and through the concordancer KonText. For a related publication, see Komrsková et al. (2018) .	Concordancer Download

The technical infrastructure



clarin.eu/fair



vlo.clarin.eu



switchboard.clarin.eu

Barack Obama's identity-building in the health care debate: A corpus-assisted discourse study

AUTHOR

[Katherina Riesner](#)

Summary, in English

In this study, I demonstrate that identity-building is an important discursive strategy for President Barack Obama in the seven-year long debate surrounding the Affordable Care Act (ACA). The data for the study comes from a 6-million word corpus of speeches that were held by Obama between January 2009 and January 2016, all published by the White House. The speeches are classified according to genre, audience, topic and date of delivery. Throughout the paper, I adopt the notion that identity is intentionally constructed by the speaker and strategically exploited for his communicative goals. With the help of two methodological approaches, I investigate what kind of identities Obama builds. The purely qualitative part of the study deals with three central corpus speeches from a discourse-analytic perspective. In the second, more quantitative part, I use a group of seven verbs with epistemic meaning to trace the usage of two predominant discursive identities in the ACA debate. The results suggest that President Obama repeatedly constructs the identities of father and teacher to persuade his audience. I argue that his use of these identities constitutes an attempt to reach the argumentative goals of effectiveness and reasonableness.

Department/s

Master's Programme: Language and Linguistics

Publishing year

2016

Language

English

Full text

[Available as PDF](#) - 2 MB



[Download statistics](#)


Document type


Student publication for Master's degree (two years)


Topic


Languages and Literatures



Lund University Humanities Lab




about
manual
register
user: anonymous
log in



METADATA SEARCH


CONTENT SEARCH


MANAGE ACCESS


BOOKMARK


REQUEST ACCESS


CITATION

IMDI Corpora

Lund Corpora

Eline Visser

ESST

Eye-Tracked Frog Stories

LACOLA

LANG-KEY

LUNDIC

REaChES

SpaceH

Strömqvist-Richthoff

Swedia2000

Tactile Reading

Test

ThaiSweVideo

The Barack Obama Corpus

the_barack_obama_corpus_information.txt

2009

2009

2010

2010

2011

2011

2012

2012

2013

2013

2014

2014

2015

2015

2016

2016

USE

VOKART

Corpus

Name

The Barack Obama Corpus

Title

The Barack Obama Corpus

Description

the_barack_obama_corpus_information.txt

Description

The Barack Obama Corpus (BOC) consists of 6,215,948 words (tokens), which are sourced from nearly 3,500 different texts, dating from January 2009 to January 2016. The texts, all taken from the White House Archives, comprise all speeches held by Barack Obama in his official capacity as 44th President of the United States of America. The earliest speech in the BOC is President Obama's inauguration speech and the last is his final State of the Union speech (January 2016). In total, the corpus includes 34,967 word types, which leads to a type/token-ratio of 0.56.

The files, which display the original titles given to them by the White House, have been tagged for genre, audience type, date and location of delivery, and principal topics. The genres include remarks, addresses, statements, press conferences, debates and question-

Description

How to cite this resource:

Riesner, Katherina (2017). The Barack Obama Corpus [Data set]. <http://hdl.handle.net/10050/00-0000-0000-0003-C53B-4@view>


Riesner, Katherina (2017). The Barack Obama Corpus [Data set].
<http://hdl.handle.net/10050/00-0000-0000-0003-C53B-4@view>

VLO / Faceted search / Search results

obama

Showing 5 results for obama  Results per page: 10 

Use the categories below to limit the search results to those matching the selected value(s).

Language Collection Modality 

The Barack **Obama** Corpus

(Part of Lund University Humanities Lab)



the_barack_obama_corpus_information.txt; The Barack **Obama** Corpus (BOC) consists of 6,215,948 words (tokens), which are sourced from nearly 3,500 different texts, dating from January 2009 to January 2016. The texts, all taken from the White House Archives, comprise all speeches held by Barack **Obama** in his official capac...



[Landing page for this record at corpora.humlab.lu.se](#)

**SWE-CLARIN**

metadata

**vlo.clarin.eu**

UDPipe

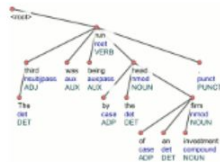
“ Please use the following text to cite this item or export to a predefined format:

[BIBTEX](#)
[CMDI](#)

Straka, Milan and Straková, Jana, 2016, *UDPipe*, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, <http://hdl.handle.net/11234/1-1702>.



Share:  



Authors:

Milan Straka, Jana Straková

Description:

UDPipe is an trainable pipeline for tokenization, tagging, lemmatization and dependency parsing of CoNLL-U files. UDPipe is language-agnostic and can be trained given only annotated data in CoNLL-U format. Trained models are provided for nearly all UD treebanks. UDPipe is available as a binary, as a library for C++, Python, Perl, Java, C#, and as a web service. UDPipe is a free software under [Mozilla Public License 2.0](#) and the linguistic models are free for non-commercial use and distributed under [CC BY-NC-SA license](#), although for some models the original data used to create the model may impose additional licensing conditions.

[Project home](#)
[Run](#)


Tool Inventory

▼ Constituency Parsing



> WebLicht Const Parsing DE



> WebLicht Const Parsing EN

▼ Coreference Resolution



> Concraft -> Bartek

▼ Dependency Parsing



> Concraft -> DependencyParser



> MaltParser



> Spacy (hosted by D4Science) - DE



> Spacy (hosted by D4Science) - EN



LINDAT

> UDPipe



> WebLicht Dep Parsing DE



> WebLicht Dep Parsing EN



Exercise 1

- Go to clarin.eu
- Find the VLO
- Search for texts by Robert Louis **Stevenson**
 - what can you find?
 - which format is the corpus in?
 - where are they hosted?
- On the Links tab, use the three dots (...) to activate the Switchboard.
 - Explore with **Voyant** or
 - Process with **UDpipe**

CLARIN Knowledge Infrastructure



Knowledge centres

CLARIN Resource Families

VideoLectures

Digital Humanities Course Registry

Workshops

Trainer Network Programme

Training Suite



<https://www.clarin.eu/content/clarin-for-researchers>

<https://www.clarin.eu/content/knowledge-sharing>

Knowledge Centres

List of all 22 CLARIN K-centres with expertise in specific linguistic topics

Click on the full name of the K-centre to go to its landing page, and click on the acronym to see its full organisation details

ACE	CLARIN Knowledge Centre for Atypical Communication Expertise
Areas of competence	Atypical communication encompasses language and speech as encountered during (second) language acquisition and development, and in language disorders, but also more broadly in bilingual language development and in sign language. ACE is specialised in this type of research and concomitant infrastructural issues related to data acquisition, processing and sharing, which is typically highly characterised by sensitivity issues. For data storage and access the centre collaborates with MPI's TLA (The Language Archive) which is a CLARIN B Centre and also based in Nijmegen.
Audiences served	- linguists; - psychologists; - neuroscientists; - computer scientists; - speech and language therapists; - education specialists
Types of services	- how-to documents; - access to document templates; - Access to data; - Depositing; - FAQ; - Helpdesk; - Technical support
Is portal for language(s)	-
Other languages covered	-
Modalities covered	- Audio: speech; - Text; - Video: sign language
Linguistic topics	- Language acquisition (L1 and L2); - language disorders; - Language learning
Language processing	-
Data types	-
Resource families	- Spoken corpora; - Manually annotated corpora; - Multimodal corpora
Generic topics	- Critical Data Management; - Legal and ethical issues
Other keywords	- Language acquisition; - sign language; - language pathologies
Tour de CLARIN	Introduction Interview

<https://www.clarin.eu/content/knowledge-centres>

CLARIN and Open Science



- Promoting the sharing and re-use of data through sustainable data registries
- All integrated datasets available in open access for research purposes
- Adherence to the FAIR data principles
 - Findable, Accessible, Interoperable, Re-usable
 - Interoperability through a common metadata framework
- Promotion of responsible data science
- Support for linguistic diversity
 - Data covering more than 1500 languages
 - Tools for many languages
 - Language resources in all modalities
- Strengthening the support for professional SSH researchers (> 500.000 in Europe)

[CLARIN: Towards FAIR and Responsible Data Science Using Language Resources.](#) In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, May 2018, 3259-3264.

CLARIN-IT - Consortium

CLARIN-IT (2015-onwards)



About

Governance

Consortium

Centres

Join

Access

Events

Initiatives

News

Home

the Italian Common Language Resources and Technology Infrastructure



English



ONLINE EVENTS

[EUPORIA 2021 Webinar - Encoding a Critical Apparatus](#)



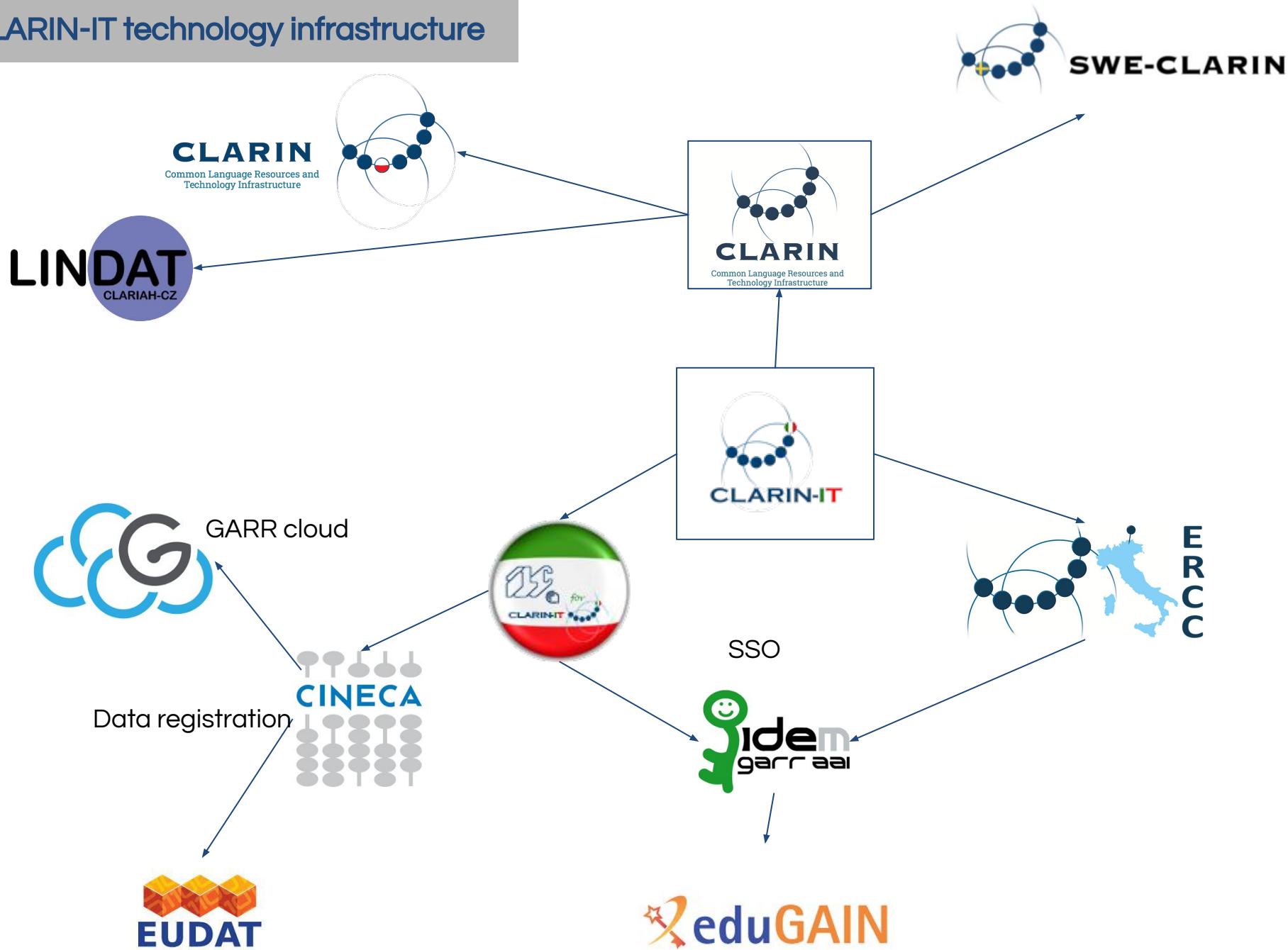
07/12/2020 - [Registration](#) to follow the Webinar via Zoom

CLARIN-IT Consortium



- The Department of Education, Human Sciences and Intercultural Communication of the University of Siena
 - The Department of Philology and Literary Criticism of the University of Siena
 - The Eurac Research Association (Bolzano)
 - The Bruno Kessler Foundation (Trento)
 - The Archival and Bibliographical Superintendence of Tuscany (Firenze)
 - The Department of Electrical Engineering and Information Technology and the Interdepartmental Research Center "URBAN/ECO" of the University of Naples Federico II
-
- The Catholic University of the Sacred Heart (Milano) has started the membership procedure.

CLARIN-IT technology infrastructure



CLARIN-IT Research Topics



- **Resources and Tools for the Italian Language**
 - create new resources by enriching existing corpora
 - lexical datasets with Linked Open Data
 - specialized corpora for computer-mediated communication.
 - natural language processing and analysis tools, offered as [web services](#) and integrated into [Weblicht](#).
- **Resources for Regional Languages and Multilingual corpora**
 - learner corpus for German, Italian and Czech,
- **Speech Archives**
 - Grafo, Caterina Bueno Archive
- **Digital Classics**
 - resources for Ancient Greek and Latin (LOD version of the TEI-dict Perseus Liddell-Scott Jones dictionary)
 - Italian Latinity of the Middle Ages
 - digital editions of ancient fragmentary texts

ILC4CLARIN



CLARIN
B CENTRE



[Chi siamo](#) [Organizzazione](#) [Repository](#) [Servizi](#) [Eventi CLARIN](#)



ILC4CLARIN

REPOSITORY

Easy to be found | Easy to be cited

A graphic illustrating a network of nodes (dark blue circles) connected by white lines, with one node highlighted by the Italian flag. To the right, a circular logo features twelve yellow stars surrounding the text "A European Research Infrastructure".

ItalWordNet v.2

Please use the following text to cite this item or export to a predefined format:

Roventini, Adriana; Marinelli,
hosted at Institute for Computational Linguistics
<http://hdl.handle.net/11362/44821>

BIBTEX CMDI

Share

Home

Authors

Item identifier

Project URL

Demo URL

Date issued

Type

Size

Language

Description

ALIM

ARCHIVIOVi.Vo.
Conservazione e diffusione degli archivi orali e audiovisivi

REGIONE TOSCANA

Suprintendenza Archivistica e Bibliografica della Toscana

UNIVERSITÀ DI SIENA

CASENTINO

ILC4CLARIN: deposit



CLARIN
B CENTRE



CLARIN-IT CLARIN

[Chi siamo](#) [Organizzazione](#) [Repository](#) [Servizi](#) [Eventi CLARIN](#)



[OPEN](#)

[Data, Tools and Services from other Italian institutions](#)

Exercise 2

- Find the ILC4CLARIN repository
 - Which is the most represented language in terms of records?
 - What kind of data can you find in Arabic?
 - How do you cite CophiWordNet?
- Try to log in with your institutional identifier, using the Login function (top right)

Lexical services



ItalWordNet

Parola: Mostra tutte le relazioni ☐

casa, Nome

- [1] - edificio o parte di esso in cui si abita;
([abitazione](#) [1], [casa](#) [1], [dimora](#) [2], [magione](#) [1], [ostello](#) [2], [tetto](#) [4])
- [2] - edificio con particolari funzioni.
([casa](#) [2])
- [3] - dinastia, casa regnante; l'insieme dei sovrani, appartenenti a una stessa famiglia, che si succedono sul trono
([casa](#) [3], [casa regnante](#) [1], [dinastia](#) [1])
"la casa di Savoia non ha regnato a lungo sull'Italia"
- [4] - la famiglia alla quale si appartiene
([casa](#) [4])
"sta pensando di mettere su casa"
- [5] - ditta, compagnia, in particolare nel settore dell'editoria e della moda
([casa](#) [5])
"casa editrice"
- [6] - casella nel gioco degli scacchi
([casa](#) [6])

NLP services



URL del servizio: [freeling_it](#) (WSDL)

i Come usare il servizio

Prova questo servizio attraverso il form sottostante.



Run service

Inputs

input

☐ as URL

☒ direct data or local file

Choose file No file chosen

language

it.cfg ▼

multiword

yes ☐ no ☒

ner

--Use Default-- ▼

output_format

--Use Default-- ▼

Reset fields

Report

mandatory
optional

CLARIN-ERIC Opportunities for Italian researchers



<https://www.clarin.eu/content/funding-opportunities>



17 european parliaments
including Senato della Repubblica italiana

ParlaMint-GB 2.0 (British parliament) : COVID_GB			ParlaMint-GB 2.0 (British parliament) : MPResReference			
(It	word	frequency	frequency/mill	frequency	frequency/mill	Score
	covid	9,667	412.9	1	0.0	413.9
	pandemic	10,341	441.7	0	0.1	402.4
	coronavirus	9,698	243.0	0	0.0	244.4
	Covid	5,676	242.4	0	0.0	243.0
	lockdown	5,521	237.9	2	0.2	199.0
	furlough	2,129	90.9	0	0.0	91.9
	PPE	1,564	66.8	0	0.0	67.8
	virus	6,009	256.7	40	4.0	51.5
	distancing	1,853	79.1	6	0.6	50.1
	CHS	860	36.7	0	0.0	37.7
	lockdowns	666	28.4	0	0.0	29.4
	Coronavirus	652	27.8	0	0.0	28.8
	SAGE	572	24.4	0	0.0	25.4
	tracing	1,048	44.8	8	0.8	25.6
	isolate	1,579	67.4	20	2.0	22.8
	quarantine	834	35.6	7	0.7	21.5
	masks	908	38.8	9	0.9	20.9
	vaccine	4,856	202.8	91	9.1	20.6
	trace	1,748	74.7	27	2.7	20.5
	shielding	492	21.0	1	0.1	20.0
	Virtual	397	17.0	0	0.0	18.0
	furloughed	396	16.9	0	0.0	17.9
	Trace	389	16.6	0	0.0	17.6
	Inaudible	370	15.8	0	0.0	16.8
	COVID	363	15.5	0	0.0	16.5
	infection	1,627	72.5	36	3.6	16.0
107		31.0	1	0.0	30.9	
210		263.9	208	7.6	30.9	
100		29.0	1	0.0	28.9	

<https://www.clarin.eu/content/parlamint-towards-comparable-parliamentary-corpora>

Exercise 3

- Explore the ParlaMint corpora using NoSketch Engine (public):
clarin.si/noske/index-en.html.
- Find the Italian corpus
 - Explore the **Corpus info** - how many subcorpora are there?
 - Create a wordlist for the COVID subcorpus
- Keyword extraction
 - extract **keywords** for the “COVID” subcorpus using the “REFERENCE” as **Reference (sub)corpus**

CLARIN for researchers



- Contact CLARIN for help with your research
- Visit this page
 - <https://www.clarin.eu/content/clarin-researchers>
- Participate in CLARIN (virtual) events
 - <https://www.clarin.eu/events>
- Tour de CLARIN
 - <https://www.clarin.eu/Tour-de-CLARIN>
- Use CLARIN Training and videolectures

Yankelevich, Tanya, Fiser, Darja, Lenardic, Jakob, Gorgaini, Elisa, & Braukmann, Ricarda. (2020, June). LIBER 2020 - Workshop: SSHOC Train-the-Trainer Bootcamp for Librarians. Zenodo. <http://doi.org/10.5281/zenodo.3970799>

Contacts



- CLARIN ERIC
 - Newsletter <https://www.clarin.eu/news>
 - @CLARINERIC
- CLARIN IT
 - www.clarin-it.it
 - @CLARIN_IT
 - coordination@clarin-it.it