Introduction to the CLARIN Infrastructure

infrastructural support for the study of
language as linguistic, social and cultural data

Francesca Frontini

Istituto di Linguistica Computazionale – CNR & CLARIN ERIC

CLARIN

Research Infrastructures

The European Commission defines research infrastructures as “**facilities, resources and services used by the science community to conduct research and foster innovation**. They include major s**cientific equipment, resources such as collections, archives or scientific data**, e-infrastructures such as data and computing systems, and communication networks.

They can be used beyond research, e.g. for education or public services and they may be single-sited, **distributed**, or **virtual**”.

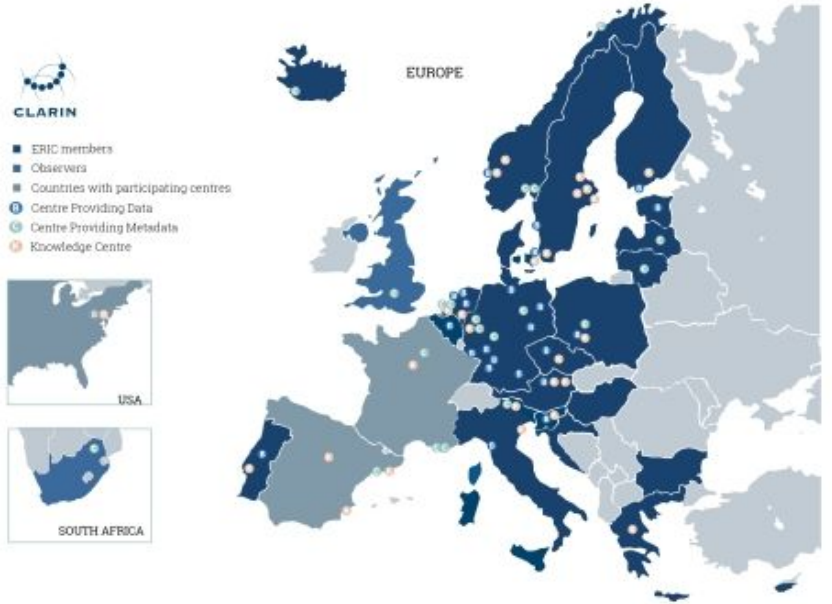
CLARIN …

- is the *Common Language Resources and Technology Infrastructure*
- has **ERIC** status since 2012, ESFRI **Landmark** since 2016
- provides easy and sustainable access for scholars in the **humanities and social sciences** (SSH) and beyond
  - to **digital language data** (in written, spoken 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 and training**
- is one of the European RIs in the SSH cluster (aka SCI)
- and actively involved in shaping the **European Open Science Cloud** (see **clarin.eu/eosc**)

CLARIN Value Proposition: https://www.clarin.eu/content/value-proposition (link to pdf)

CLARIN ERIC in members and centres

- a distributed network of **70 centres**
- **22 members**: AT, BE, BG, CY, CZ, DE, DK, EE, FI, GR, HR, HU,IS, IT, LT, LV, NL, NO, PL, PT, SE, SI
- **2 observers:** UK, ZA
- **1 third party**

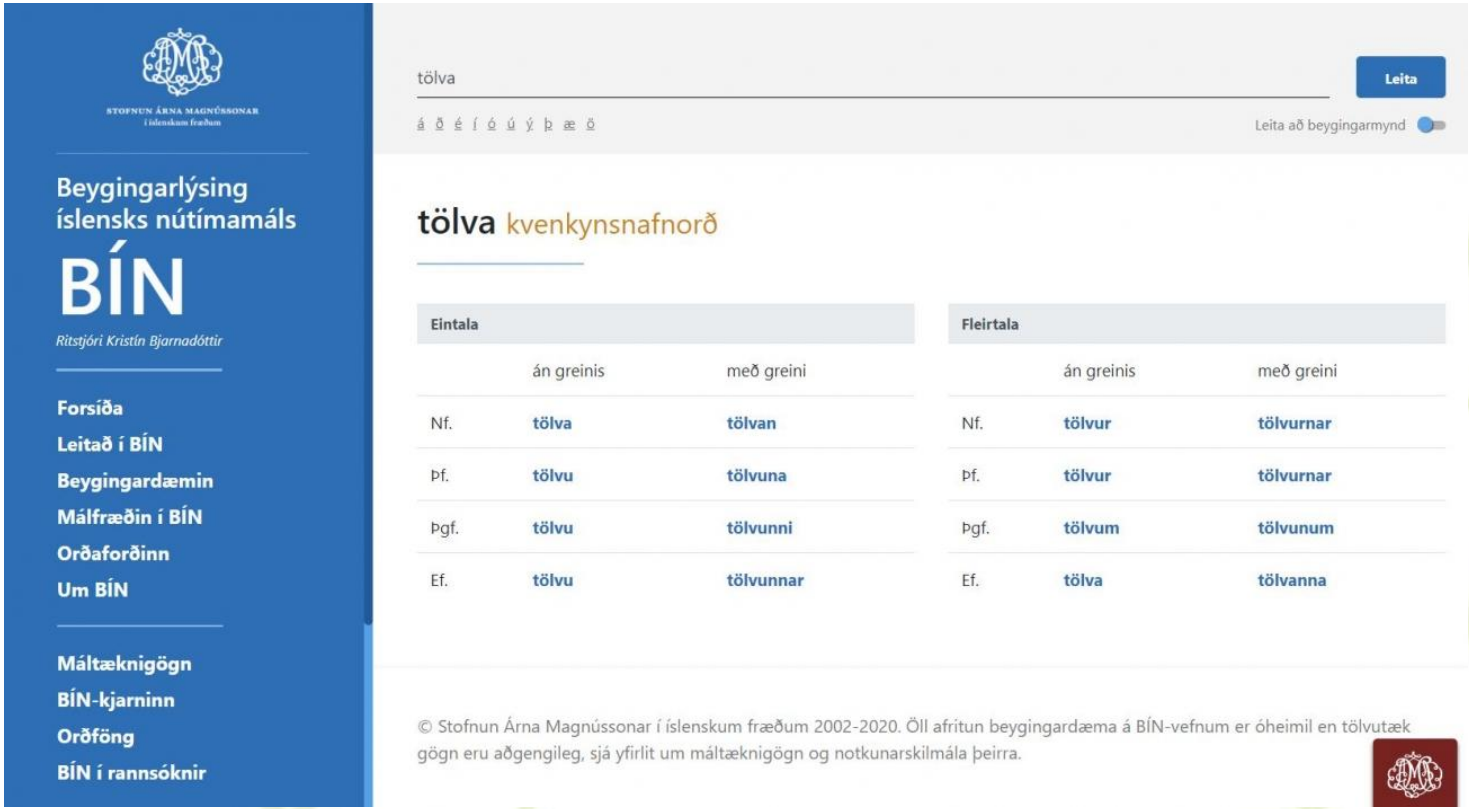


CLARIN in data types and communities of use

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

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

The Database of Modern Icelandic Inflection



https://bin.arnastofnun.is/

Universal Dependencies

This page pertains to UD version 2.

nmod: nominal modifier

The nmod relation is used for nominal dependents of another noun or noun phrase and functionally corresponds to an attribute, or genitive complement.

New from v2: The nmod relation was previously used also for nominal dependents of verbs, adjectives, and adverbs. The latter are now covered by the new obl relation.

In conjunction with the case relation, nmod provides a uniform analysis for the possessive alternation (with the option of a subtype like nmod:poss to distinguish non-adpositional case):

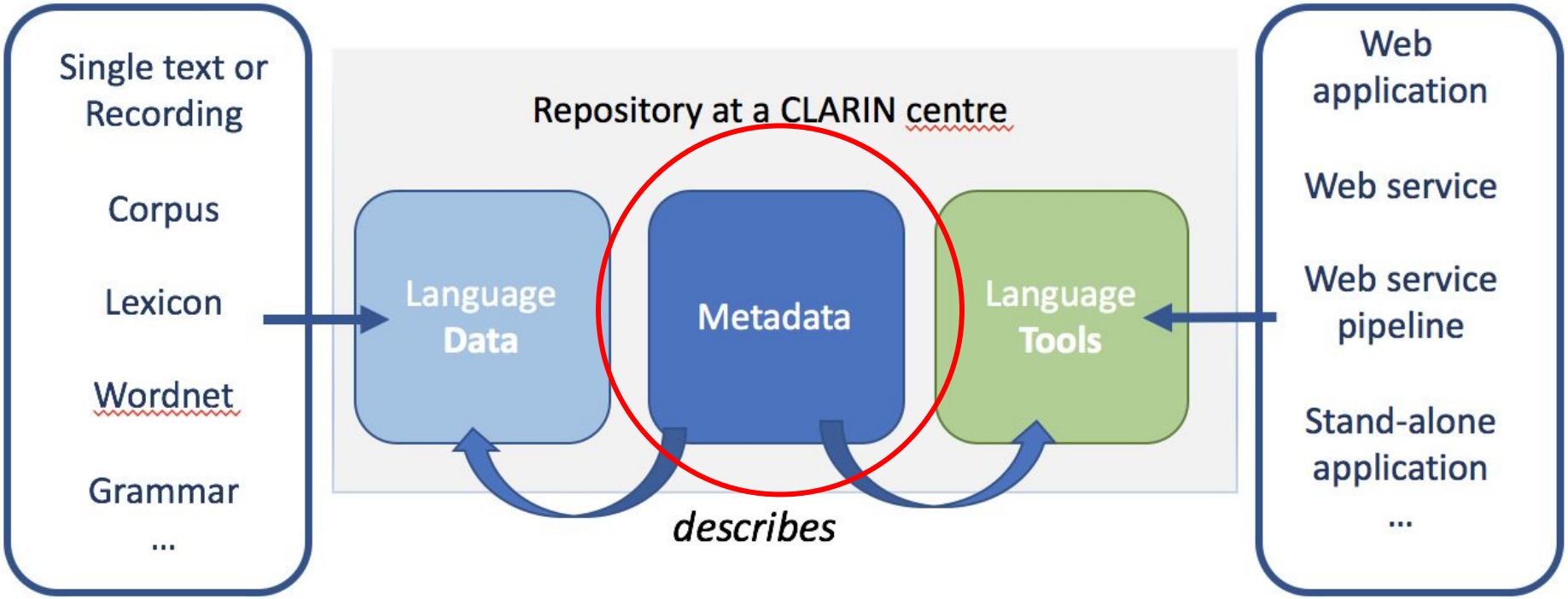
nmod in other languages: [am] [ar] [bg] [bxr] [ca] [ckb] [cop] [cs] [cu] [da] [de] [el] [en] [es] [et] [eu] [fa] [fi] [fo] [fr] [ga] [gl] [got] [grc] [he] [hi] [hr] [hu] [id] [it] [ja] [kk] [kmr] [ko] [la] [lv] [mr] [nl] [no] [pl] [pt] [ro] [ru] [sa] [sk] [sla] [sl] [so] [sr] [sv] [swl] [ta] [tr] [u] [ug] [uk] [ur] [urj] [vi] [yue] [zh]

http://universaldependencies.org/

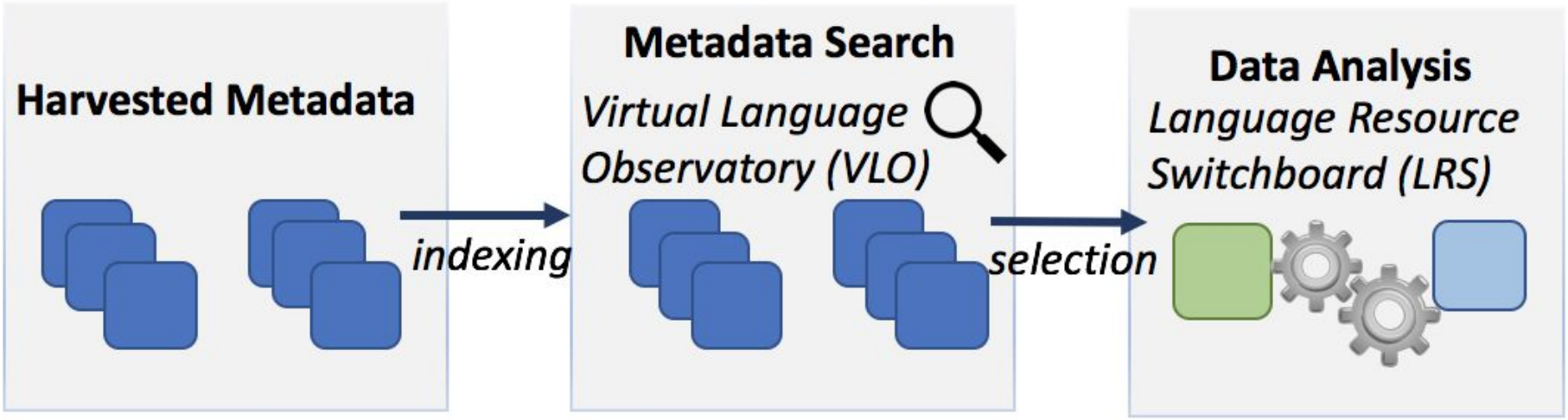
HuComTech corpus

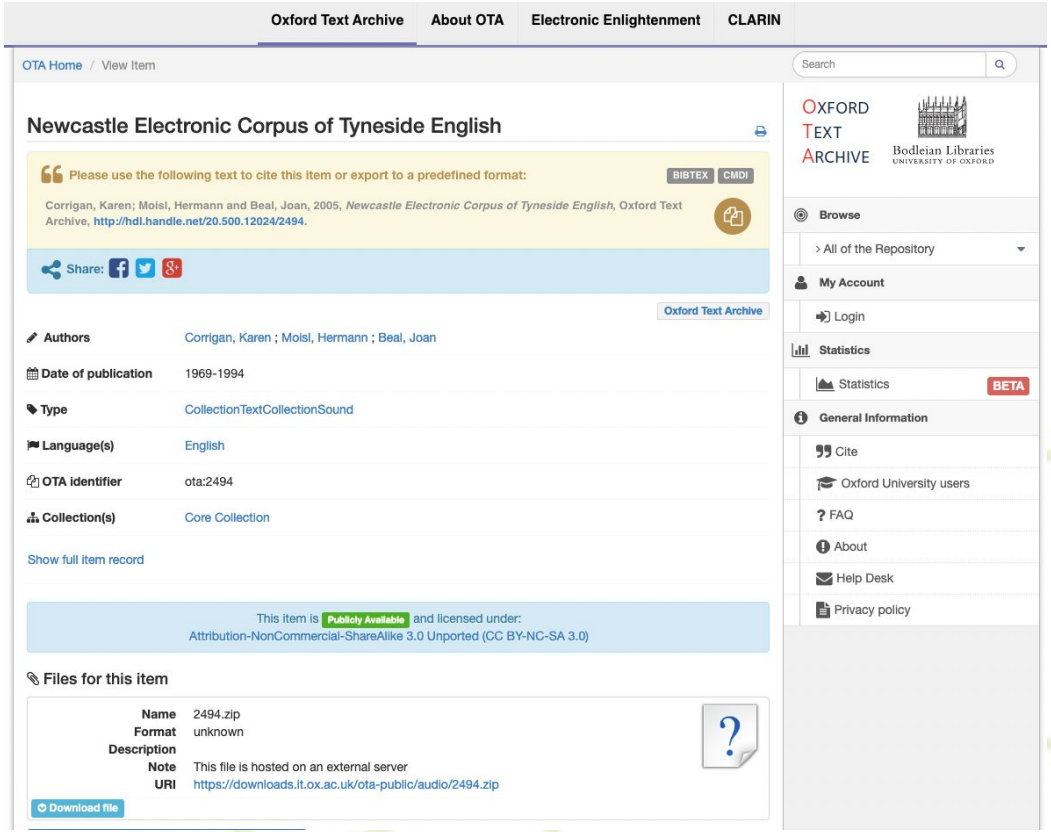
http://tla.nytud.hu

The CLARIN data architecture:
a network of distributed centre repositories



The CLARIN data architecture: *central processing of metadata*



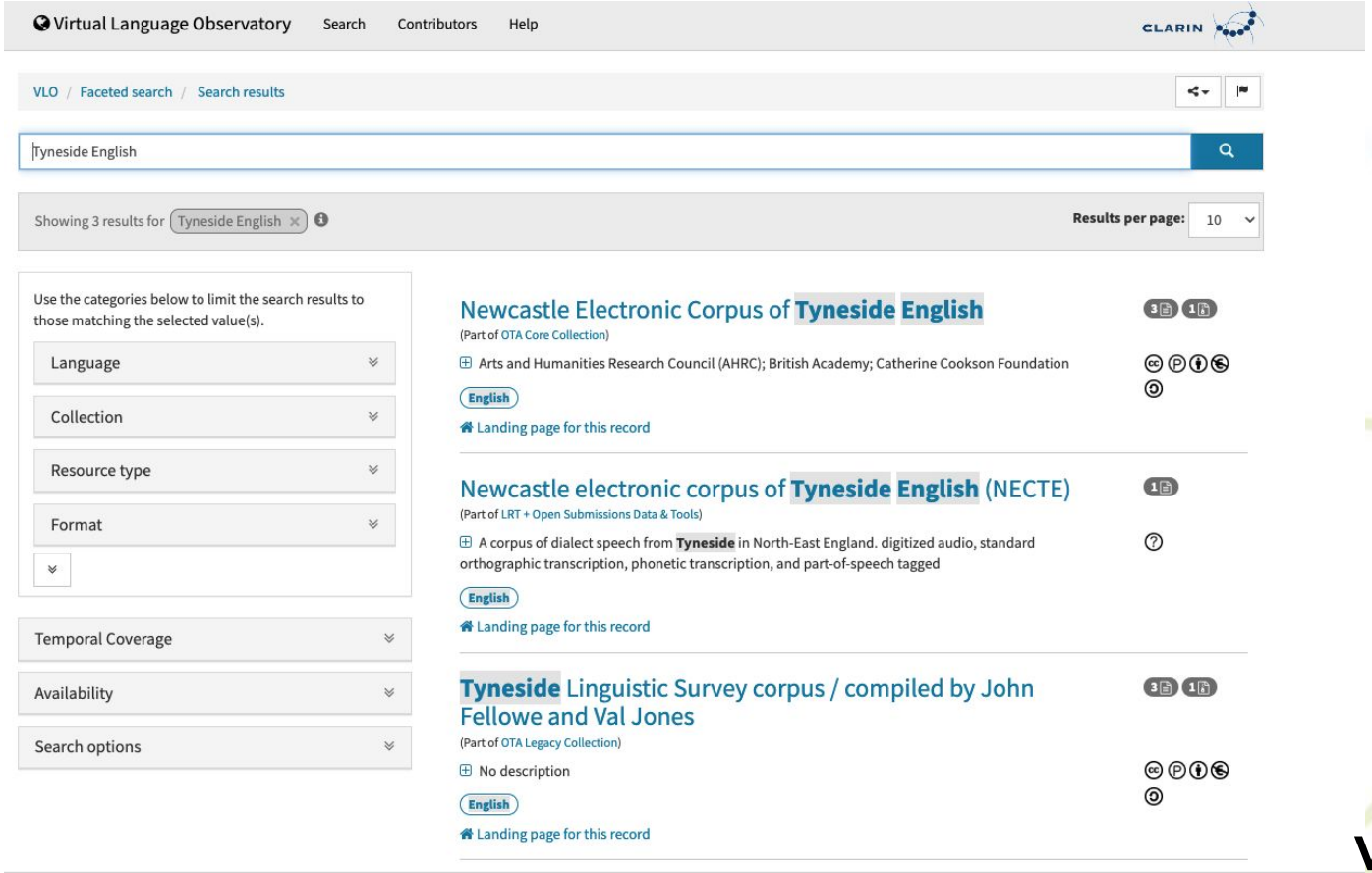


An example of a corpus deposited on the Oxford Text Archive, one of CLARIN's repositories

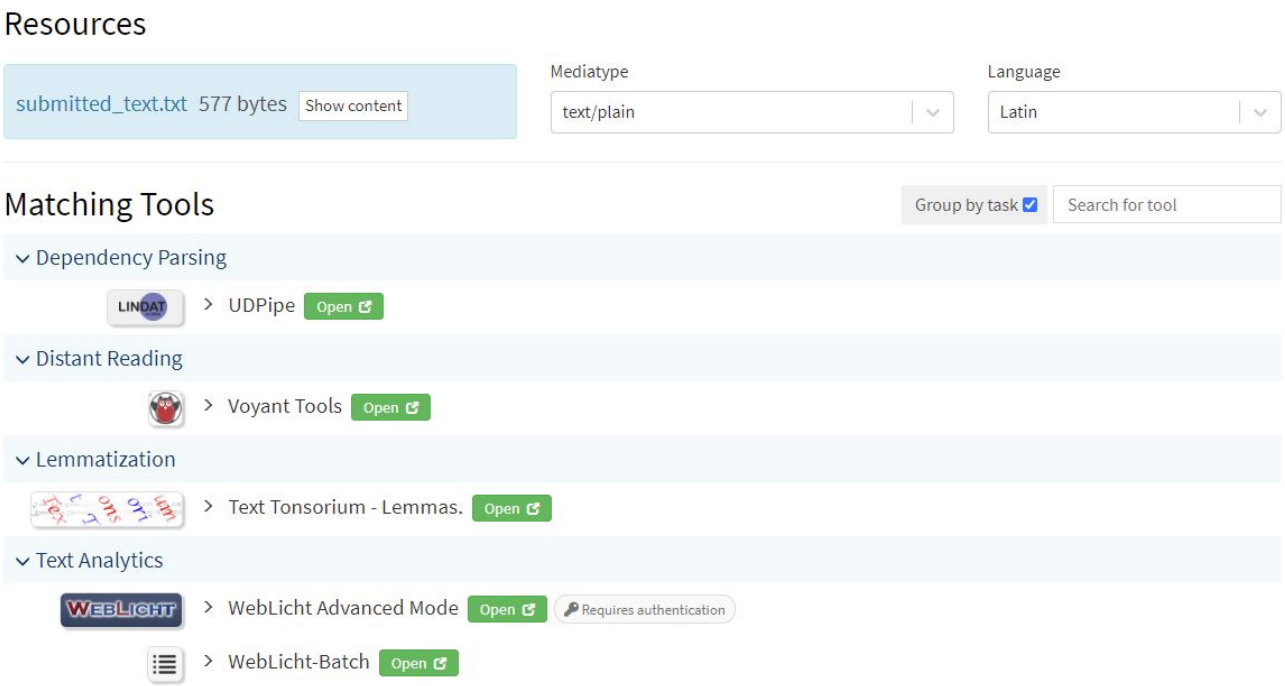
Depositing resources in CLARIN ensures

- long term preservation
- clear licensing
- findability

http://hdl.handle.net/20.500.12024/2494



vlo.clarin.eu



switchboard.clarin.eu

Other central services

Virtual Collection Registry: where scholars can create and publish their lists of datasets and resources, providing them with a persistent identifier. Examples of use use:

- data as mentioned in an article's footnotes gathered in a single virtual collection
- a virtual collection with links to data illustrating a book (video and sound recordings)

Federated Content Search: enable researchers to search for specific patterns across collections of data, CLARIN offers a search engine that connects to the local data collections that are available in the centres.

Resource Families

Introduction

The aim of the CLARIN Resource Families initiative is to provide a user-friendly overview per data type of the available language resources in the CLARIN infrastructure for researchers from the digital humanities, social sciences and human language technologies. The overviews are meant to facilitate comparative research and the listings are sorted by language.

The listings for each family include the most important metadata and brief descriptions, such as resource size, text sources, time periods, annotations and licences as well as links to download pages and concordancers. In addition to the resources found in the CLARIN infrastructure, an overview is provided of other existing valuable language resources which have not yet been integrated into the infrastructure.

The listings also provide hyperlinks to other relevant materials, such as the thematic CLARIN workshops and tutorials and their accompanying videolectures, as well as a list of key publications on the resources surveyed.

Currently, overviews are available for 13 corpora families, 5 families of lexical resources, and 4 tool families. See below. For information about applying for funding for small projects that can help to extend the scope of the initiative, see https://www.clarin.eu/content/clarin-resource-families-project-funding.

Corpora

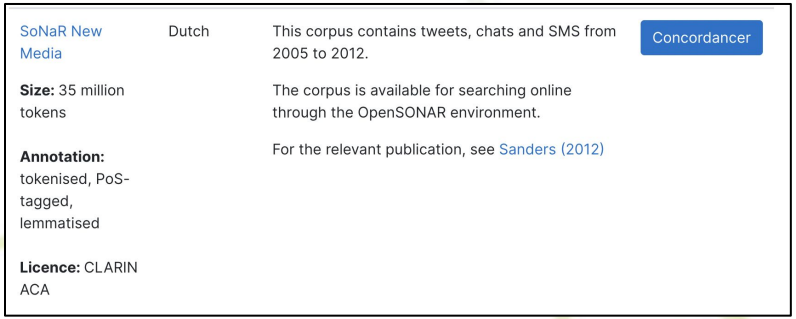
- Computer-mediated communication corpora
- Corpora of academic texts
- Historical corpora
- L2 learner corpora
- Legal corpora
- Literary corpora
- Manually annotated corpora
- Multimodal corpora
- Newspaper corpora
- Parallel corpora
- Parliamentary corpora
- Reference corpora
- Spoken corpora

Lexical Resources

- Lexica
- Dictionaries
- Conceptual Resources
- Glossaries
- Wordlists

Tools

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



https://www.clarin.eu/resource-families

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 500.000 professional SSH researchers

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.

Knowledge Infrastructure

Tour de CLARIN

Funding

Video Channels

Best-Practice Papers

K-centres

CLARIN knowledge centres (K-centres) bring together expertise for a specific **domain**, **topic**, data **modality**, **language**, etc.

Currently, the K-centres, which can be operated by a single institute/group or arranged as a distributed structure, already cover a large number of research topics, languages and resource types.

CLARIN's strategy aims and **broadening** the range of topics covered by K-centres, incentivising closer **cooperation** between them, and promoting their geographic distribution across CLARIN member countries.

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

DiPText-KC

CLARIN Knowledge Centre for Digital and Public Textual Scholarship

DiPText-KC offers expertise on methods, data, instruments and technologies relevant in the field of Philological and Literary Studies, History, Art History and Cultural Heritage.

Its actions aim at:

- sharing information with scholars and students about the state of the art in digital scholarly editing and text annotation through domain-specific languages;
- supporting scholars and students in the creation and publication of digital scholarly editions and resources;
- organizing training activities (for instance webinars, workshops and summer schools).

DiPText-KC is one of the Centres of CLARIN-IT, the Italian node of CLARIN (Common Language Resources and Technology Infrastructure), a digital infrastructure of pan-European interest identified by ESFRI (European Strategy Forum on Research Infrastructures) and classified as a Landmark Research Infrastucture for the Social Sciences and Humanities (ESFRI Landmarks SSH RI).

Being officially recognized as a CLARIN Knowledge Centre (K-Centre), DiPText-KC is allowed to use the CLARIN K-Centre logo in its official communications.

CLARIN K CENTRE

DiPText-KC	CLARIN Knowledge Centre for Digital and Public Textual Scholarship
Areas of competence	The CLARIN Knowledge Centre for Digital and Public Textual Scholarship offers expertise on;methods, data, instruments and technologies relevant in the field of Philological and;Literary Studies, History, Art History and Cultural Heritage.;Its actions aim at 1) sharing information with scholars and students about the state of the art in digital;scholarly editing and text annotation through Domain-Specific Languages; 2) supporting scholars and students in the creation and publication of digital scholarly editions and resources; 3) organizing training activities (for example webinars, workshops, and summer schools).
Audiences served	- Digital philologists; - Computational philologists; - Scholars in literary studies; - Art historians; - Cultural heritage scholars; - Computational Linguists
Types of services	; - FAQ; - Helpdesk; - Training; - User assistance
Is portal for language(s)	-
Other languages covered	- Italian; - Italian; - Ancient Greek; - Latin; - Arabic; - English; - French; - German
Modalities covered	; - Text; - Images
Linguistic topics	- Text encoding; - Phonology; - Morphology; - Syntax; - Semantics; - Lexicography
Language processing	- Automated text encoding; - Information extraction; - Domain-Specific Language parsing
Data types	- XML-TEI encoded documents; - Treebanks; - Wordnets; - Dictionaries; - Ontologies
Resource families	; - Historical corpora; - Literary corpora; - Manually annotated corpora; - Parallel corpora; - Lexica; - Dictionaries; - Wordnets
Generic topics	- Text encoding; - Linked Open Data; - Domain-Specific Languages
Other keywords	- Data modeling in the domain of philological and literary studies; - Software engineering in the domain of philological and literary studies
Tour de CLARIN	- -

TdC showcases the richness of the CLARIN landscape and displays the full range of **activities throughout the network**.

Tour de CLARIN highlights prominent user involvement activities of different CLARIN national consortiums, shines the spotlight on individual Knowledge Centres (K-centres) and Service Providing Centres (B-centres)

https://www.clarin.eu/Tour-de-CLARIN

CONSORTIA

K-CENTRES

B-CENTRES

follows the tag set and principles defined in Branco et al. (2015), while the universal dependencies option converts the CINTIL dependencies into Universal Dependencies (de Marneffe et al. 2014).

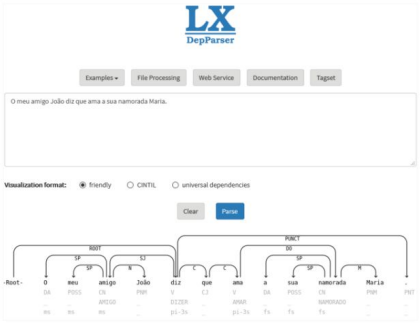


Figure 2: Output of the dependency parser, in-browser, in the "friendly" format.

Figure 2 shows the user-friendly output format of the tool for the example sentence *O meu amigo João diz que ama a sua namorada Maria* ("My friend João says he loves his girlfriend Maria"). The ROOT relation pointing to *diz* ("says") identifies it as the main predicate of the entire sentence, whereas the SJ (subject) relation pointing to *amigo* ("friend") defines the latter – or rather the entire phrase that it heads, *O meu amigo João* ("My friend João") – as the syntactic subject of the main clause.

LX-DepParser is run as the last step of a language processing pipeline. The preceding steps handle tokenization, part-of-speech tagging and morphological analysis. The

#id	form	lemma	cpos	pos	feat	head	deprel	phead	pdeprel
1	O	-	DA	DA	ms	3	SP	3	SP
2	meu	-	POSS	POSS	ms	3	SP-ARG1	3	SP-ARG1
3	amigo	AMIGO	CN	CN	ms	5	SJ-ARG1	5	SJ-ARG1
4	João	-	PNM	PNM	-	3	N	3	N
5	diz	DIZER	V	V	pi-3s	0	ROOT	0	ROOT
6	que	-	CJ	CJ	-	5	C-ARG2	5	C-ARG2
7	ama	AMAR	V	V	pi-3s	6	C	6	C
8	a	-	DA	DA	fs	10	SP	10	SP
9	sua	-	POSS	POSS	fs	10	SP-ARG1	10	SP-ARG1
10	namorada	NAMORADO	CN	CN	fs	7	DO-ARG2	7	DO-ARG2
11	Maria	-	PNM	PNM	-	10	M-PRED	10	M-PRED
12	.	-	PNT	PNT	-	5	PUNCT	5	PUNCT

Figure 3: Output of the dependency parser, in-browser, in the CINTIL format.

Figure 3 shows the output, for the same sentence, in the CINTIL tabular format. This format is akin to the commonly used CoNLL format and is amenable to being read by a computer. This is also the format produced by the other modes of accessing the tool: the file processing service and the web service.

For natural language processing, LX-DepParser has mostly been used as a component in larger processing pipelines, such as LX-SRLabeler (for semantic role labelling) or LX-Suite (a suite of shallow processing tools). It has also been used, for instance, to provide features for a machine-learning classifier in a work where the dependencies produced by the parser were used as features for a classifier that assigns deep lexical types for handling out-of-vocabulary words in a deep processing grammar (Silva 2014). These grammars make use of lexica with extremely fine-grained syntactic categorization, but cannot proceed when a word is not found in their lexicon, and relying only on the coarse annotation of a normal part-of-speech tagger leaves too much ambiguity unresolved for a useful and efficient analysis. The classifier was able to use the features provided by the parser to assign fine-grained tags.

For the study of language, the parser has been used to quickly provide a tentatively annotated corpus that was then manually corrected, leading to the creation of CINTIL-DependencyBank PREMIUM. Moreover, from personal communications we are aware that the dependency

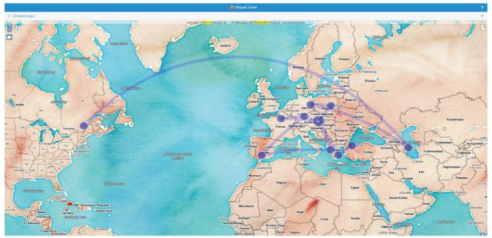


Figure 18: Geomapping applied to the Axioti corpus. The blue arc shows the occurrences of named entities as they appear in the text.

Although Dreamscape is currently in its beta-release stage, which means that the identification of locations is not really fully reliable, it still gives us an idea of how easily such tools could be used in the near future, as well as of their impact on macro-analyses. Of course, accurate geospatial representation of texts is still possible today through a combined use of existing tools – e.g., in the case of Greek texts, I think that CLARIN:EL Annotator of Named Entities can be of great help.

>

In preparing such resources, has NLP:EL helped you with any kind of linguistic or structural annotation?

<

Linguistic and structural annotation is very important in literary and translation studies, but what is crucial for qualitative research is to combine them with extra-structural markup. In the Melpo Axioti corpus, we have first used annotation tools provided by NLP:EL through CLARIN:EL – specifically, the CLARIN:EL Annotator of Named Entities and the GrNE-Tagger – in order to tag the recurrence of persons and places in Melpo Axioti's works, and the ILSP Feature-based multi-tiered POS Tagger and ILSP Lemmatizer to study the morphological changes in her language, which are connected with her ideological choices in different periods of her work. Afterwards thereby created a double tagset, in which the house is tagged as either a positive or a negative space, with subsets defined by positive (e.g., "home", "friendly") and negative ("disgust", "loneliness") connotations of the word, as seen in Figure 19.

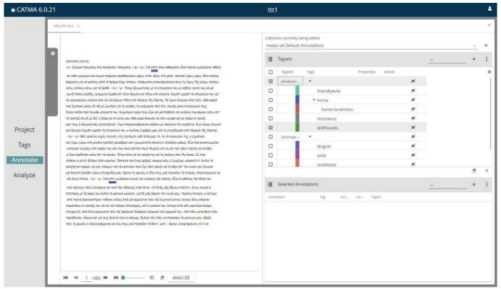


Figure 19: CATMA applied to the Melpo Axioti corpus.

On the basis of this, we then linked the presence or absence of the occurrences and particular meanings and connotations of the term *house* to broader thematic and stylistic choices, as well as to different periods, in Melpo Axioti's work. This sort of extra-linguistic annotation is a good example of how a researcher is able to work with qualitative research questions using a digital approach, especially in humanities, as such questions can only be formulated accurately if the user knows the corpus well and tries to shed new light on it by applying both linguistic and extra-linguistic tools offered by different developers.

>

Your research has quite a large and interdisciplinary scope, combining corpus linguistics with approaches like post-structuralist discourse theory, as well as literary studies. Could you present a result or paper in which your research has benefitted from NLP:EL's involvement? How concretely did your research benefit from NLP:EL?

<

Impact Stories

Showcases high-quality and innovative research that uses CLARIN tools and resources.

IceTaboo: Offensive Word Database with Commercial Application

The IceTaboo database can be used to flag contextually inappropriate words in texts, and is already being used as part of an automatic proofreading software by an Icelandic online news website.

Read more

https://www.clarin.eu/content/clarin-impact-stories

Teaching with CLARIN

A new CLARIN platform, which encourages the integration of CLARIN resources, tools and services into the curricula of SSH-related disciplines.

Description of the Training Materials

(Sub)discipline, topic, language(s)	Phonetics, Linguistics Speech Analysis Finnish and English
Keywords	phonetics, acoustic phonetics, speech analysis, research methods
Course URL	http://urn.fi/urn:nbn:fi:lb-2021063021
CLARIN resources	Praat, ELAN, VLO, Kielipankki – The Language Bank of Finland, Route to A wing Corpus

https://www.clarin.eu/content/teaching-clarin

DH Course Registry

DIGITAL HUMANITIES
COURSE REGISTRY

FILTER SORTING

NAME	UNIVERSITY	PLACE	DATE	TYPE
Skills in Transnational History MO5161	University of St Andrews	St Andrews, United Kingdom	10 Sept ⟲ 1 semester	Module
MA in Digital Humanities	Lancaster University	Lancaster, United Kingdom	4 Oct ⟲ 12 months	Master Programme
MSc Digital Heritage	University of York	York, United Kingdom	1 Oct ⟲ 1 years	Master Programme
MSc Digital Archaeology	University of York	York, United Kingdom	1 Oct ⟲ 1 years	Master Programme
MA in Digital Humanities	King's College London	London, United Kingdom	24 Sept ⟲ 1 years	Master Programme
Introduction to Computational Linguistics (DHS)	University of Latvia	Riga, Latvia	7 Feb 2023 1 semester	Course
Introduction to Computational Linguistics	University of Latvia	Riga, Latvia	9 Feb ⟲ 4 months	Master Programme
Media and Information	Universiteit van Amsterdam	Amsterdam, Netherlands	1 Sept ⟲ 3 years	Bachelor Programme
Laboratorio di Informatica per Umanisti (Computing for the ...	Unviersità degli Studi di Udine	Udine, Italy	11 Feb ⟲ 10 weeks	Bachelor Programme
Digital Humanities Laboratory - Laboratorio di	University of Palermo	Palermo,	2 months	Course

INFO

https://dhcr.clarin-dariah.eu/

Funding and support

https://www.clarin.eu/funding

More knowledge sharing

- CLARIN Legal and Ethical Issues Committee (CLIC)
  - **Legal Information Platform**
    https://www.clarin.eu/content/legal-information-platform
- CLARIN Standards Committee (CSC)
  - **Standards Information System**
    https://standards.clarin.eu/sis/
- **CLARIN cafés**
  - informal and interactive space for discussion
- **Social media presence**
  - Newsflash
  - Twitter, LinkedIn

Getting involved in CLARIN

- Subscribe to the NewsFlash
  - https://www.clarin.eu/content/newsflash
- Check out our events
  - https://www.clarin.eu/events
- Open calls
  - https://www.clarin.eu/content/funding-opportunities
- Follow us on Twitter
  - @CLARINERIC

Thank you!

CLARIN

www.clarin.eu