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*
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 scientific 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

<table>
<thead>
<tr>
<th>Data Types</th>
<th>Communities of Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Newspaper archives</td>
<td>Digital humanities</td>
</tr>
<tr>
<td>Parliamentary records</td>
<td>Linguistics and Philology</td>
</tr>
<tr>
<td>Literary texts</td>
<td>Data Science /AI</td>
</tr>
<tr>
<td>Historical letters</td>
<td>Translation and Lexicography</td>
</tr>
<tr>
<td>Broadcast archives</td>
<td>Literary Studies</td>
</tr>
<tr>
<td>Oral History data</td>
<td>History</td>
</tr>
<tr>
<td>Social Media data</td>
<td>Political and Social Sciences</td>
</tr>
<tr>
<td>L-2 Learner Resources</td>
<td>Media Studies</td>
</tr>
<tr>
<td>Survey data</td>
<td>Culture, Folklore, Anthropology</td>
</tr>
<tr>
<td>Patient recordings</td>
<td>Speech therapy</td>
</tr>
<tr>
<td>Excavation reports</td>
<td>General Public</td>
</tr>
</tbody>
</table>
# The Database of Modern Icelandic Inflection

The Database of Modern Icelandic Inflection (BÍN) provides access to detailed information about the Icelandic language, including its inflectional patterns. The website, [https://bin.arnastofnun.is/](https://bin.arnastofnun.is/), offers interactive features to help users understand and learn about the language.

## Inflectional Table

<table>
<thead>
<tr>
<th>Case</th>
<th>Nominative</th>
<th>Accusative</th>
<th>Genitive</th>
<th>Dative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nf.</td>
<td>tölvur</td>
<td>tölvur</td>
<td>tölvunnar</td>
<td>tölvunnar</td>
</tr>
<tr>
<td>Pf.</td>
<td>tölvu</td>
<td>tölvuna</td>
<td>tölvunni</td>
<td>tölvunni</td>
</tr>
<tr>
<td>Plf.</td>
<td>tölvum</td>
<td>tölvunum</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ef.</td>
<td>tölva</td>
<td>tölvanna</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This table provides the singular and plural forms for various cases in the nominative, accusative, genitive, and dative cases. The website also includes additional resources and tools for learning and research.

For more information and to explore the database in detail, visit the [BÍN website](https://bin.arnastofnun.is/).
**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):

```
1: the office of the Chair
2: the Chair’s office
```

*nmod* in other languages: [am] [ar] [bg] [bfr] [ca] [ckb] [cop] [cs] [cu] [da] [de] [el] [en] [es] [et] [eu] [fa] [fi] [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] [sl] [so] [sr] [sv] [sw] [ta] [tr] [ug] [uk] [ur] [ur] [vi] [yue] [zh]

[http://universaldependencies.org/](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
| Use the categories below to limit the search results to those matching the selected value(s). |
| Language | Collection | Resource type | Format |
| Temporal Coverage | Availability | Search options |

### Newcastle Electronic Corpus of Tyneside English

Newcastle Electronic Corpus of Tyneside English
(whole of the full-text collection)
- Arts and Humanities Research Council (AHRC); British Academy; Catherine Cookson Foundation
- English
- Landing page for this record

### Newcastle electronic corpus of Tyneside English (NECTE)

Newcastle electronic corpus of Tyneside English (NECTE)
(whole of the full-text collection)
- A corpus of dialect speech from Tyneside in North-East England. digitized audio, standard orthographic transcription, phonetic transcription, and part-of-speech tagged
- English
- Landing page for this record

### Tyneside Linguistic Survey corpus / compiled by John Fellowe and Val Jones

Tyneside Linguistic Survey corpus / compiled by John Fellowe and Val Jones
(whole of the full-text collection)
- No description
- English
- Landing page for this record
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 include:

- 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 video lectures, 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

SoNaR New Media
- Dutch
- This corpus contains tweets, chats and SMS from 2005 to 2012.
- Size: 35 million tokens
- The corpus is available for searching online through the OpenSONAR environment.
- Annotation: tokenised, PoS-tagged, lemmatised
- Licence: CLARIN ACA

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

Knowledge Infrastructure

Knowledge Centres

Digital Humanities Course Registry

Tour de CLARIN

Teaching

Annual Conference

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
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 Infrastructure 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.
<table>
<thead>
<tr>
<th><strong>DiPText-KC</strong></th>
<th><strong>CLARIN Knowledge Centre for Digital and Public Textual Scholarship</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Areas of competence</strong></td>
<td>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).</td>
</tr>
<tr>
<td><strong>Audiences served</strong></td>
<td>- Digital philologists; - Computational philologists; - Scholars in literary studies; - Art historians; - Cultural heritage scholars; - Computational Linguists</td>
</tr>
<tr>
<td><strong>Types of services</strong></td>
<td>; - FAQ; - Helpdesk; - Training; - User assistance</td>
</tr>
<tr>
<td><strong>Is portal for language(s)</strong></td>
<td>-</td>
</tr>
<tr>
<td><strong>Other languages covered</strong></td>
<td>- Italian; - Italian; - Ancient Greek; - Latin; - Arabic; - English; - French; - German</td>
</tr>
<tr>
<td><strong>Modalities covered</strong></td>
<td>; - Text; - Images</td>
</tr>
<tr>
<td><strong>Linguistic topics</strong></td>
<td>- Text encoding; - Phonology; - Morphology; - Syntax; - Semantics; - Lexicography</td>
</tr>
<tr>
<td><strong>Language processing</strong></td>
<td>- Automated text encoding; - Information extraction; - Domain-Specific Language parsing</td>
</tr>
<tr>
<td><strong>Data types</strong></td>
<td>- XML-TEI encoded documents; - Treebanks; - Wordnets; - Dictionaries; - Ontologies</td>
</tr>
<tr>
<td><strong>Resource families</strong></td>
<td>; - Historical corpora; - Literary corpora; - Manually annotated corpora; - Parallel corpora; - Lexica; - Dictionaries; - Wordnets</td>
</tr>
<tr>
<td><strong>Generic topics</strong></td>
<td>- Text encoding; - Linked Open Data; - Domain-Specific Languages</td>
</tr>
<tr>
<td><strong>Other keywords</strong></td>
<td>- Data modeling in the domain of philological and literary studies; - Software engineering in the domain of philological and literary studies</td>
</tr>
<tr>
<td><strong>Tour de CLARIN</strong></td>
<td>- -</td>
</tr>
</tbody>
</table>
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 consortia, shines the spotlight on individual Knowledge Centres (K-centres) and Service Providing Centres (B-centres).

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

<table>
<thead>
<tr>
<th>Id</th>
<th>form</th>
<th>lemma</th>
<th>spc</th>
<th>pos</th>
<th>gen</th>
<th>feat</th>
<th>head</th>
<th>depend</th>
<th>plural</th>
<th>person</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>SP</td>
<td>3</td>
<td>SP</td>
</tr>
<tr>
<td>2</td>
<td>am</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>SP.ARG1</td>
<td>3</td>
<td>SP.ARG1</td>
</tr>
<tr>
<td>3</td>
<td>am</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>SP.ARG1</td>
<td>3</td>
<td>SP.ARG1</td>
</tr>
<tr>
<td>4</td>
<td>João</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>N</td>
<td>3</td>
<td>in</td>
</tr>
<tr>
<td>5</td>
<td>diz</td>
<td>-</td>
<td>0.03</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>ROOT</td>
<td>3</td>
<td>ROOT</td>
</tr>
<tr>
<td>6</td>
<td>que</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>C-ARG2</td>
<td>3</td>
<td>C-ARG2</td>
</tr>
<tr>
<td>7</td>
<td>ama</td>
<td>-</td>
<td>0.07</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>A-ARG3</td>
<td>3</td>
<td>A-ARG3</td>
</tr>
<tr>
<td>8</td>
<td>a</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>A-ARG3</td>
<td>3</td>
<td>A-ARG3</td>
</tr>
<tr>
<td>9</td>
<td>suo</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>SP</td>
<td>3</td>
<td>SP</td>
</tr>
<tr>
<td>10</td>
<td>nam</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>SP</td>
<td>3</td>
<td>SP</td>
</tr>
<tr>
<td>11</td>
<td>Maria</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>M-PRED</td>
<td>3</td>
<td>M-PRED</td>
</tr>
<tr>
<td>12</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>PUNCT</td>
<td>3</td>
<td>PUNCT</td>
</tr>
</tbody>
</table>

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-SRL abler (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 parser has been used in a classroom setting to demonstrate students of Portuguese...
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 extrastructural 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 GReNE-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 13.

Figure 13: 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 benefited 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.

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

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

https://www.clarin.eu/content/teaching-clarin
DH Course Registry

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