

Other Platforms for Multilingual Technologies and Future Directions

UniDive WG3 Meeting (Istanbul/2023)

A. Seza Doğruöz

What are the other language resource platforms?

- In addition to ELG, there are other language resource platforms.

CLARIN

- CLARIN (Common Language Resources & Technology Infrastructure)
- ELRA (European Language Resources Association)

The research infrastructure for language as social and cultural data

CLARIN is a digital infrastructure offering data, tools and services to support research based on language resources.

<https://www.clarin.eu/>

Access to Data (CLARIN)

- Virtual Language Observatory (VLO) and/or through member centers.
- 20 certified centers.
- *“The Virtual Language Observatory (VLO) provides a means of exploring language resources and tools. Its aim is to provide an easy-to-use interface, allowing for a uniform search and discovery process for a large number of resources from a wide variety of domains. Facets make it easy to explore and access available resources. A powerful query syntax makes it possible to carry out more targeted searches as well. It also makes it easy to review processing options for discovered resources via the Language Resource Switchboard, and to create virtual collections based on search results via the Virtual Collection Registry.” (CLARIN website, 2023).*

CLARIN Virtual Language Observatory

Welcome to the VLO!

Use the **search bar** below to start searching through hundreds of thousands of language resources, or [continue](#) to browse everything and use **facets** to narrow down to your area of interest or discover new resources.

See all records

Take a quick tour

Search through 858,477 records



Showing all records (622,549 results)

Results per page: 10

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

Language

Collection

<< < 1 2 3 4 5 6 7 8 9 10 > >>

SmartKom Audio

(Part of [Bavarian Archive for Speech Signals \(BAS\)](#))

This corpus contains the audio recordings of all actors who use the SmartKom system; it covers the audio recordings (no video) and annotations of all three original SmartKom

447

1



Level of Granularity (Taxonomy for Language Resources)

- Language
- Collection
- Resource Type
- Modality
- Format
- Keyword
- Genre
- Subject
- Country
- Organization
- Data Provider
- National Project
- Temporal Coverage
- Availability

Language Resource Families (CLARIN)

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
- Oral History Corpora
- Parallel Corpora
- Parliamentary Corpora
- Reference Corpora
- Sign Language Resources
- Spoken Corpora

Lexical Resources

- Language Models
- Lexica
- Dictionaries
- Conceptual Resources
- Glossaries
- Wordlists

Tools

- Corpus Query Tools
- Normalisation
- Named Entity Recognition
- Part-of-Speech Tagging and Lemmatisation
- Tools for Sentiment Analysis

Language Resource Switchboard

“The Language Resource Switchboard can assist with finding the right language processing tool for your data. Upload a file, or enter a URL, and the Switchboard will provide step-by-step guidance to have your data processed with a CLARIN tool”
(CLARIN website, 2023).

How to upload and process your own data on CLARIN

Switchboard

Switchboard helps you find tools that can process your data.

The data will be shared with the tools via public links. For more details, see the [FAQ](#).



Upload files or text

Tool inventory


Tool Inventory

Group by task Search for tool






▼ Constituency Parsing

-  > WebLicht Const Parsing DE Requires authentication
-  > WebLicht Const Parsing EN Requires authentication

▼ Coreference Resolution

-  > Concraft -> Bartek Not secure

▼ Dependency Parsing

-  > Concraft -> DependencyParser Not secure
-  > MaltParser
-  > UDPipe
-  > WebLicht Dep Parsing DE Requires authentication
-  > WebLicht Dep Parsing EN Requires authentication

▼ Distant Reading

-  > Voyant Tools

Frequently Asked Questions

- *Q: What happens to the research data that is transferred to the Switchboard?*
A: The tools that process your resource need to have access to it. For this, your research data is uploaded to a temporary file storage on a CLARIN-based server for a limited time. The resource is shared on the basis of a unique URL that is not publicly listed and hard to guess. This URL is shared with the tool, which may or may not redistribute or store this location. Note that in some cases the URL cannot be passed through the tool over a secure (encrypted) channel. All uploaded research data is deleted at regular intervals. This procedure is carried out for uploaded files, resources provided by reference (pasted URLs and incoming via other services) as well as text inserted manually into the input box on the Switchboard web page.
- *Q: Can I use the Switchboard with resources containing sensitive or private information, or which are subject to restrictions regarding redistribution?*
A: No. CLARIN does not have full control over the shared resource and therefore cannot guarantee its secure transfer, storage or processing. Therefore it is not advised to upload files or enter content that contains sensitive or private information, or is subject to restrictions regarding redistribution.
- *Q: What information is collected when visiting the Switchboard site?*
A: The Switchboard makes use of cookies and uses Matomo (Piwik) to track user visits.
- *Q: Is there a way to perform a batch processing of files?*
A: The Switchboard is not able to batch processing many documents itself. It can, however, invoke a tool capable of batch processing. At the time of writing, there is a single tool connected to the Switchboard that can batch process documents: "WebSty". To invoke WebSty, upload a zip archive of Polish plain text files and follow the instructions.
- *Q: Is there a way for the Switchboard to run tools in a pipeline?*
A: The Switchboard is not a workflow engine. It can, however, invoke a tool capable of executing pipelines. The WebLicht workflow engine, for instance, has multiple Switchboard entries each advertising a complex analysis such as constituency parsing or named entity recognition. WebLicht then orchestrates the execution of the pipeline that divides the complex analysis in simpler tasks.
- *Q: Is there a way to give the Switchboard two or more resources at once?*
A: At the time of writing, only a single resource can be processed. But we are well aware of tools that need two or more inputs (e.g., for

Annual Report (CLARIN)

- [2022](#), [2021](#) reports (publicly available)
- Updates about lectures
- Available resources
- Projects & Awards
- Finances

ELRA

- European Language Resources Association
- *ELDA: Evaluations & Language Resources Distribution Agency*

"It is the association's operational body, and it is in charge of the development and the execution of ELRA's missions and tasks as defined by the Board of the association. ELDA is incorporated as a company in order to handle all the commercial and business-oriented tasks of the association."

New Membership Drive

Share this page!     

General Meeting @ LREC 2018 on May 9, 2018

The slides presented during the meeting are available as **PDF**.

The new membership at ELRA now encompasses two (2) types of members:

1. the Institutional Members (formerly called ELRA Members), including the ELRA Subscribers
2. the Individual Members.

For the **Institutional Members**, the main change lies in the discontinuation of the Fidelity Program, replaced by a new mechanism **rewarding the loyal members**. Their benefits remains those described here: <http://www.elra.info/en/join-elra/members-benefits/>.

The **Individual Member**, the new category of ELRA Members, includes:

1. the individual researchers and users of LRs who will join the association by paying individual membership fees
 2. the employees of institutional ELRA members who are ELRA members by default, without having to pay for the individual membership fees since the institutional membership covers for that.
- Individual members are represented by a Board member designated among the members and elected by them.

Individual members cannot buy resources from ELRA; this is legally restricted to institutional members. Individual members may, however, obtain the free resources for which ELRA has the right

- ➔ Join ELRA
- ➔ Membership Fees

Links

Tags

- association
- benefits
- ELRA member
- fees
- join
- membership fidelity

You are here » Catalogues

Catalogues

Share this page!     

This section gives you access to all the LR distribution and sharing features initiated and/or maintained by ELRA:

- [Catalogue of Language Resources](#)
- [Language Resources Announcements](#)
- [META-SHARE](#)
- [LRE Map](#)
- [R&D Catalogue](#)
- [Universal Catalogue](#)
- [Free Resources](#)

Links

→ [ELRA Catalogue](#)

Tags

- [catalogue](#)
- [distribution](#)
- [ELRA](#)
- [identification](#)
- [language resources](#)
- [r&d](#)
- [sharing](#)

You are here » [Catalogues](#) » ELRA Catalogue

Catalogue of Language Resources

Share this page!     



An increasing number of language resources in the various fields of HLT (namely spoken, written and terminological resources) are made available via the Evaluations & Language resources Distribution Agency (ELDA), in the catalogue which you can browse on-line @ catalog.elra.info

The latest Language Resources announced are also listed [here](#).

Search the ELRA catalogue below:



ELRA is a partner of the Open Language Archives Community (OLAC).
The ELRA Catalogue can be viewed as an [OLAC repository](#)

Links

- ➔ [ELRA Catalogue](#)
- ➔ [LR Announcements](#)
- ➔ [Purchase a LR](#)
- ➔ [Order Procedure](#)

Tags

- [catalogue](#)
- [evaluation package](#)
- [language resources](#)
- [monolingual](#)
- [multilingual](#)
- [speech corpus](#)
- [written corpus](#)

OLAC (Open Language Archives Community) Repository



Archive Details

ELRA Catalogue of Language Resources

Size	1573
Repository Name	ELRA Catalogue of Language Resources
Institution	ELRA (European Language Resources Association)
ArchiveURL	http://catalogue.elra.info/
Location	9 rue des Cordelières, 75013 Paris, France
Short Location	Paris, France
Synopsis	ELRA is the driving force to make available the language resources for language engineering and to evaluate language engineering technologies. In order to be active in identification, distribution, collection, validation, standardisation, improvement, in promoting the production of language resources, in support of evaluation campaigns and in developing a scientific field of language resources and evaluation.
Access	Language resources from the ELRA catalogue can be obtained upon licensing through ELRA.
Administrator	mapelli@elda.org
Participants	Valérie MAPELLI (Project manager and sales coordinator)
Base URL	http://www.elra.info/elrac/elra_catalogue.xml
Repository ID	catalogue.elra.info
OAI Version	2.0
OLAC Version	1.1
Records in Archive	http://www.language-archives.org/archive_records/catalogue.elra.info
Faceted search	http://search.language-archives.org/search.html?q=archive_facet%3A%22ELRA+Catalogue+of+Language+Resources%22

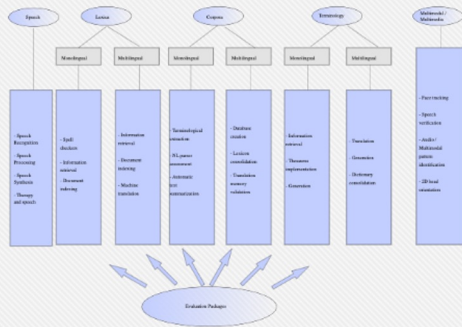


elra catalogue

1,573 language resources at your disposal

Type in your keywords, please...

Search



An increasing number of LRs in the various fields of Human Language Technology (see image on the left-hand side) are distributed on behalf of ELRA via its operational body ELDA, thanks to the contribution of various players of the HLT community.

Our aim is to provide Language Resources, by means of this repository, so as to prevent researchers and developers from investing efforts to rebuild resources which already exist as well as help them identify and access those resources.

ELRA Level of Granularity (Taxonomy)

- Language
- Resource Type
- Media Type
- Availability
- Licence
- Restrictions of Use
- Validated
- Subject
- Language Variety
- Foreseen Use
- Use is NLP Specific
- Linguality Type
- Multilinguality Type
- Modality Type
- Conformance to the Standards
- Domain
- Geographic Coverage
- Time Coverage



European languages and beyond?

- What are the options?
- How do we make the decisions to include new languages/regions?
- How can we work together?

Africa

- Masakhane NLP <https://www.masakhane.io/>
- Ghana NLP <https://ghananlp.org/>
- Digital Umuganda in Rwanda <https://digitalumuganda.com/>
- Galsen AI <https://galsen.ai/>



Insights from Dr. David Adelani (UK): *“Many of these efforts are quite new, started 3-5 years ago. We can learn a lot from the UniDive initiative on long term sustainability of these projects, and how to properly create and sustain language artifacts that are being created.”*

Americas NLP

<https://turing.iimas.unam.mx/americasnlp/>

- Indigenous language of the Americas
- Multiple Workshops
- Shared Tasks



NusaCrowd - Indonesian Languages

IndoNLP/**nusa-crowd**



A collaborative project to collect datasets in Indonesian languages.

 40

Contributors

 27

Issues

 240

Stars

 57

Forks



Time for Discussion for WG3

- Where do we go from here?
- What are the next steps?

Thank you!
Let me know if you have any questions & comments!