Language resources and translation studies: “Search, find, archive and share”

Vesna Lušicky
University of Vienna
Austria

Tanja Wissik
Austrian Academy of Sciences
Austria
Overview

1. Introduction
2. Language resources
3. Language resource management lifecycle
4. Research infrastructures
5. Search, Find, Create, Use, Archive, Share
6. Benefits of CLARIN
7. Conclusions
Language Resources I.

“a set of speech or language data and descriptions in machine readable form, used for building, improving or evaluating natural language and speech algorithms or systems, or, as core resources for the software localisation and language services industries, for language studies, electronic publishing, international transactions, subject-area specialists and end users.” (ELRA)
Language Resources II.

Data

Tools

Standards,
Annotation frameworks…

(based on Mörh 2017)
Language Resources III.

- Examples of language resources can be:
  - corpora (written and spoken; mono-, bi- and multilingual; parallel and comparable),
  - translation memories,
  - termbases (terminology databases),
  - computational lexica,
  - digital dictionaries,
  - ontologies, etc.
Language Resources IV.

- **Creation and processing:**
  - Time consuming and costly,
  - Can involve legal challenges,
  - Can involve technical challenges, etc.
Language Resource Management Lifecycle

- **Search**
  - Lookup
  - Exploratory

- **Find**
  - RI
  - Catalogues
  - Repositories

- **Create**
  - Data collection
  - Data processing

- **Use**
  - For research
  - In teaching

- **Share**
  - Repositories
  - Catalogues
  - Citation
  - Legal Issues

- **Archive**
  - Institutional Repository
  - Domain specific Repository
Research Infrastructures

“Facilities, resources and related services that are used by the scientific community to conduct top-level research in their respective fields and covers major scientific equipment or sets of instruments; knowledge-based resources such as collections, archives or structures for scientific information […]”

SEARCH

- **Lookup**
  - Has precise search goals, such as finding facts to answer a specific question.

- **Exploratory search**
  - Includes a variety of qualitative definitions
  - Open-ended search goals
  - Faceted search
FIND

- USE CASE I.
  - Finding a language resource.

- USE CASE II.
  - Finding elements within a resource
FIND: Use Case I.

Repositories

- Preserve, manage, and provide access to language resources in a variety of formats;
- Curated to enable search, discovery, and reuse;
- Persistent Identifiers;
- Data can be cited;
- Long-term preservation;
- Examples: CLARIN Centres, Zenodo, institutional repositories.

Source: https://repo.clarino.uib.no
FIND: Use Case I.

Catalogues

- Contain **metadata**\(^*\) from repositories about language resources;
- Do not contain the data or tool itself but only refer to it;
- Often enable faceted search;
- Useful as a first step to which language resources are available.

\(^*\) Data that provides information about other data.

Source: https://vlo.clarin.eu

---

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.

Source: https://vlo.clarin.eu
FIND: Use Case II.

- **Aggregator: Content search in several resources and several repositories at the same time.**

- Useful as a first step to discover where interesting language resources are hosted.

Source: https://spraakbanken.gu.se/ws/fcs/2.0/aggregator
FIND: Use Case II.

- Download of the search results.
- Next step: Specialised search in the resource in the repository.

Source: https://spraakbanken.gu.se/ws/fcs/2.0/aggregator
FIND: Use Case II.

- Specialised search in the resource in the repository.
- Finding elements in authentic texts, e.g. monolingual corpus, as a support for the production of target text.
  - Example: Corpus of the Contemporary Lithuanian Language

Source: http://corpus.vdu.lt/en/?word=prieglobstis
FIND: Use Case II.

- Specialised search in the resource in the repository.
- Finding contrastive elements: e.g. in an aligned parallel corpus
  - Example: EU DGT 2015

Source: https://www.clarin.si/kontext/
CREATE

● Data collection
  - Data found in CLARIN
  - Data collection from scratch

● Data processing
  - CLARIN tools/services for data processing
  - Tool chains for data processing
CREATE: Example Weblicht

Source: https://weblicht.sfs.uni-tuebingen.de/weblicht/

CLARIN Workshop, Vienna, 8th February 2018

Vesna Lušicky & Tanja Wissik
CREATE: Example ContaWords

Source: http://contawords.iula.upf.edu
USE: Research – Example

Source: https://vlo.clarin.eu/record?0&docId=http://47.47.47.urn.fi_47.urn.fi_47_20160425028&q=the+finnish+broadcasting+corpus&index=1&count=5

Source: https://vlo.clarin.eu/record?3&docId=http://47.47.47.urn.fi_47.urn.fi_47_20140730134&q=finnish+broadcasting+corpus&index=0&count=6
USE: Research – Example

USE: Research – Example


Source: https://vlo.clarin.eu accessed 7th February 2018

Source: https://arche.acdh.oeaw.ac.at accessed 7th February 2018

CLARIN Workshop, Vienna, 8th February 2018
USE: Research – Example

USE: In the classroom

- Technology-related courses
- Information mining
- Language-specific courses

Source: Centre for Translation Studies, University of Vienna.

Source: Lušicky, Vesna (2017) Towards convergence of e-research in translation studies and blended learning in translator training through technology and language resources.
ARCHIVE

“A curation activity that ensures that data are properly selected, stored, and can be accessed, and for which logical and physical integrity are maintained over time, including security and authenticity”

(http://dictionary.casrai.org/Archiving)
Repositories

- Institutional repository
- Domain specific repository
- Infrastructural repository
CLARIN depositing services

- Decentralized
- CLARIN Centres
- List of CLARIN Centres: https://www.clarin.eu/content/depositing-services

Source: https://www.clarin.eu/content/services
SHARE

- Citability
- Persistent Identifier: Assigned permanent addresses
- Licences

Source: https://arche.acdh.oeaw.ac.at/browser/
BENEFITS OF CLARIN

- **Time and cost effectiveness**: Re-use of language resources from already existing repositories and catalogues.
- **Accessibility**: Distributed access to a larger number of resources.
- **Durability**: Archiving your own data, responding to technology changes.
- **Discoverability**: Make your data discoverable.
- **Citability**: Make your data citable and traceable for other researchers.
- **Interoperability**: Make resources reusable in various tools and scenarios.
Thank you!

Questions?

vesna.lusicky@univie.ac.at

tanja.wissik@oeaw.ac.at