Intro 2 CLARIN 4 Translation Studies

Darja Fišer
#ACDH
Vienna, 8 February 2018

CC-BY 4.0
Teaser

Why do DH and SS researchers need CLARIN?
Find out here.
Overview

• Intro to CLARIN
• CLARIN resources
• CLARIN tools
• Deposit to CLARIN
Intro to CLARIN
CLARIN in five bullets

• **CLARIN** is the **Common Language Resources and Technology Infrastructure**
• that provides easy and sustainable access for scholars in the **humanities and social sciences**
• 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** online environment
CLARIN data types and user communities

- Newspaper archives
- Literary texts
- Parliamentary records
- Literary texts
- Historical letters
- Broadcast archives
- Oral History data
- Social Media data
- ...
CLARIN services

- **CLARIN portal**: Get an example-based impression of what's currently available.
- **Depositing services**: Store language resources in a sustainable repository at a CLARIN centre.
- **Virtual Language Observatory**: Discover language resources using a faceted browser or a map.
- **Easy access to protected resources**: Get easy access to protected resources, with your institutional username and password.
- **Web services and applications**: Explore and analyze language data with a wide variety of tools.
- **Virtual Collections**: Create your own digital bookmarks, ideal for citing data sets.
- **Language Resource Inventory**: Submit and access information about language resources relevant to your
- **Content Search (prototype)**: Search different corpora with a single search engine.
- **Consulting Services**: Searching for a specific data set or application? Wondering how CLARIN can...
CLARIN members and centres

A consortium of:

- 19 members: AT, BG, CZ, DE, DK, DLU, EE, FI, GR, HU, IT, LT, LV, NL, NO, PL, PT, SE, SI
- 2 observers: FR, UK;
- >40 centres
What CLARIN Centres offer

- **Repository**
  - library of linguistic data and tools
  - search for data and tools and easily use them online or download them
  - deposit your data and be sure it is safely stored, everyone can find it, and correctly cite it

- **Federated single sign-on**
  - log in once with your existing institutional credentials
  - get access to protected resources

- **Metadata**
  - describe content, provenance and formats of linguistic data and tools
  - facilitate preservation and dissemination of linguistic data and tools

- **Persistent Identifier (PID or handle)**
  - a special permanent URL that provides a permanent link to linguistic data and tools
  - will resolve correctly even if in some distant future the data is moved
  - should be used as URL in citations

- **Licensing**
  - Public
  - Academic
  - Restricted

- **Preservation (Data Seal of Approval)**
  - committed to long-term care of items in the repository
  - ensure the archived data can be found, understood and used in the future
CLARIN K-Centres

- [https://www.clarin.eu/content/knowledge-centres](https://www.clarin.eu/content/knowledge-centres)

  - K-Centres for languages
    - Polish, Spanish, Danish, Sweden
  
  - K-Centres for resources
    - Treebanking, Phonogrammarchiv

  - K-Centres for research areas
    - Speech Analysis, Language Learning Analysis, Multimodal Analysis
CLARIN resources: corpora

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

Key families of resources
• newspaper corpora
• parliamentary corpora
• social media corpora
• learner corpora
• parallel corpora
Parallel corpora

• great potential for reuse and re-purposing within many fields of study in the humanities and social sciences (and beyond):
  - suited for both *close reading* and *distance reading*
  - *Humanities*: translation studies, contrastive and comparative linguistics, ...
  - *Social sciences*: social and cultural dynamics, political sciences, economics ...
  - *NLP*: machine translation, multilingual information extraction, ...

• considered a rich data type
  - apart from linguistic content, rich in metadata and extralinguistic clues

• but also
  - often presenting itself as messy or noisy data (not 100% parallel)
  - calling for links with data in other modalities than text and speech
  - created under specific circumstances that need to be well understood before strong conclusions can be drawn (the status of source and target languages)
Accessing resources through national repositories

Tourism English-Croatian Parallel Corpus 2.0
(Abu-MaTran project / 2016-01-28)

Author(s):
Toral, Antonio; Esplà-Gomis, Miquel; Klubička, Filip; Ljubešić, Nikola; Papavassiliou, Vassilis; Prokopidis, Prokopis; Rubino, Raphael; Way, Andy

Post-edited and error annotated machine translation corpus PErr 1.0
(Insight Centre for Data Analytics, National University of Ireland, Galway / 2016-05-24)

Author(s):
Popović, Maja; Arčan, Mihael
Accessing resources through the Virtual Language Observatory

**English-Czech Corpus** from Wikipedia


<table>
<thead>
<tr>
<th>Name</th>
<th>English-Czech Corpus from Wikipedia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collection</td>
<td>LINDAT / CLARIN Data &amp; Tools</td>
</tr>
<tr>
<td>Language</td>
<td>English</td>
</tr>
<tr>
<td></td>
<td>Czech</td>
</tr>
<tr>
<td>Organisation</td>
<td>NLP Centre, Masaryk university</td>
</tr>
<tr>
<td>National project</td>
<td>LINDAT (CLARIN-CZ)</td>
</tr>
</tbody>
</table>

**Availability**
- Public (28)
- Academic (3)
- Restricted for Individual (91)
- Unspecified (59)

**Resources**
- [1362553](#)
- [cs-en.sentences.txt](#)
Overview

• Coverage
  - 47 corpora
  - most European but also other languages
  - EU, wikipedia, web, newspapers, fiction
  - 100,000 sentences – 1 billion

• Size (in tokens)
  - largest: UK (1.6 billion)
  - smallest: Portuguese (1 million)

• Availability
  - most available for download
  - must under CC-BY licence

• Full overview available here
Beyond VLO

• META SHARE
  - http://www.meta-share.org
  - network of repositories for language data and tools
  - for resources yet to be integrated with CLARIN
  - from researchers yet to join CLARIN

• LRE map
  - http://lremap.elra.info
  - research-oriented
  - specialized domains

• CorporaList
  - https://mailman.uib.no/public/corpora/
  - most recent releases
CLARIN tools:
- searching for resources
- searching in resources
- manual annotation
- automatic annotation
- combining tools
Accessing resources through the Virtual Language Observatory

Showing 21 to 28 of 28 results within selection for parallel corpus, Public, English

English-Czech Corpus from Wikipedia
(Part of LINDAT / CLARIN Data & Tools)


Name: English-Czech Corpus from Wikipedia
Collection: LINDAT / CLARIN Data & Tools
Language: English, Czech
Organisation: NLP Centre, Masaryk university
National project: LINDAT (CLARIN-CZ)

Resources:
- cs-en.sentences.txt
- 136253

Availability:
- Public (28)
- Academic (3)
- Restricted for Individual (91)
- Unspecified (59)
Today eight out of every 10 elephant deaths is the result of poaching, compared with four out of 10 deaths six years ago, the prosecutor's office said.

Tiger parts, elephant ivory, rhino horn and exotic birds and reptiles are among the most trafficked items.

Mayor Naheed Nenshi appeared unimpressed, calling it "the elephant not in the room."

Even before the spike in recent years, Africa's elephant population is estimated to have shrunk from 1.3 million in 1979 to 450,000 in 2007.

"The essence of the whole human-elephant conflict is basically that people have encroached on elephant habitat," says Ravi Corea, president of the Sri Lanka Wildlife Conservation Society.

"It's like the parable about the elephant," said Forrest.

"I think it's really brave to talk about the elephant in the room," Maya Walters, a teenager with high blood pressure who appeared in one of the ads, told the Atlanta Journal-Constitution.

He's a one-man gay pride parade in a series of Ms. Lauer's quirky costumes: a red-and-white-striped bodysuit topped by a yellow frizz wig; a demure Alice in Wonderland frock accessorized by a miniature tricycle; a bright pink, furry elephant outfit.
WebAnno

https://webanno.github.io/

• developed by Chris Biemann et al. within CLARIN-D
• general purpose web-based manual annotation tool for a wide range of linguistic and non-linguistic annotations including various layers of morphological, syntactic, semantic and other annotations
• multi-user tool supporting different roles such as annotator, curator, and project manager
• progress and quality of annotation projects can be monitored and measured
Why annotate?

1. **linguistic analysis**
   - analysis in context
   - quantitative analysis
   - theory validation
   - feedback on problematic cases
2. **language technologies**
   - testing set for annotation tools
   - training set for statistical annotation tools
   - automatic generation of lexicons
3. **practical benefits**
   - searching by categories
   - illustration of examples
   - long-term data usability
WebLicht
https://weblicht.sfs.uni-tuebingen.de/

• developed by the Tübingen CLARIN-D team
• execution environment for automatic annotation of text corpora
• combining tokenizers, part of speech taggers, parsers etc. into custom-made processing chains
• many text processing tools must be invoked from the command line or from programming code
  - sometimes difficult to install
  - not always available for all operating systems
  - confusing for those not accustomed to running command line tools or who don’t like to program
  - not possible to create some annotations with one tool and other annotations with a different tool
  - Input/Output formats differ
Workflow

Germany’s parliament has voted in favour of legalising same-sex marriage, joining many other western democracies in granting gay and lesbian couples full rights, including adoption.

Norbert Lammert, president of the parliament, said 393 lawmakers voted in favour, 226 voted against and four abstained.

The chancellor, Angela Merkel, said she voted against the move because she believed marriage was for a man and a woman. She said the decision for her was a personal one.
Other languages

https://www.clarin.eu/content/web-services
Language Resource Switchboard

[Image of an old book]

[Table]

<table>
<thead>
<tr>
<th>tokens</th>
<th>POS tags</th>
<th>lemmas</th>
<th>named Entities</th>
</tr>
</thead>
<tbody>
<tr>
<td>den</td>
<td>ART</td>
<td>d</td>
<td></td>
</tr>
<tr>
<td>Schwerpunkt</td>
<td>NN</td>
<td>Schwerpunkt</td>
<td></td>
</tr>
<tr>
<td>des</td>
<td>ART</td>
<td>d</td>
<td></td>
</tr>
<tr>
<td>Reiches</td>
<td>NN</td>
<td>Reich</td>
<td></td>
</tr>
<tr>
<td>nach</td>
<td>APPR</td>
<td>nach</td>
<td></td>
</tr>
<tr>
<td>Ungarn</td>
<td>NE</td>
<td>Ungarn</td>
<td>B-GPE</td>
</tr>
<tr>
<td>zu</td>
<td>PTKZU</td>
<td>zu</td>
<td></td>
</tr>
<tr>
<td>verlegen</td>
<td>VVINF</td>
<td>verlegen</td>
<td></td>
</tr>
</tbody>
</table>

[Diagram of a virtual language observatory]

- Named Entity Recognition
- N-Gramming
- Part-Of-Speech Tagging
- Text Analytics
CLARIN community: how to get involved
UI outreach

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

• Videolectures
  • http://videolectures.net/clarin/

• Guest blog posts
  • https://www.clarin.eu/content/guest-blog-posts

• Showcases
  • https://www.clarin.eu/showcases
  • https://www.clarin.eu/content/showcases-i
  • https://www.clarin.eu/content/showcases-ii
CLARIN for researchers

• Entry page:
  - https://www.clarin.eu/content/clarin-for-researchers

• Basic info:
  - https://www.clarin.eu/blog/using-clarin-digital-research

• Get in touch:
  - https://www.clarin.eu/content/clarin-eric-user-involvement-coordinators
  - https://www.clarin.eu/content/national-user-involvement-coordinators

• Apply for funding:
  - Mobility grants https://www.clarin.eu/news/clarin-mobility-grants

• Course Registry:
  - https://registries.clarin-dariah.eu/courses/
UI events by CLARIN ERIC in 2017

https://www.clarin.eu/events

UI events organized by CLARIN ERIC
• 1-3 CLARIN-PLUS workshops
• 4 CLARIN tutorial

UI events (co)funded by CLARIN ERIC
• 5-7 Summer schools
• 8 Workshop
• 9-10 Tutorials

UI events endorsed by CLARIN ERIC
• 11 Workshop

Invited contributions at 3rd party events
• 12 Conference
• 13, 15 Summer schools
• 14, 16 Workshops
UI events by national consortia in 2017

• 160 UI activities recorded, 30 more than in 2016
  - attended by up to 200 participants, in total over 5300 participants
  - 92 humanities and digital humanities, 82 linguistics, 46 social sciences, 27 computer science, 26 literature
  - 30 English 21 Finnish, 18 German, 17 Polish, 15 Dutch
UI surveys

• Focus Group
  - What do researchers need?
  - What are they struggling with?

• Overview of key research networks for CLARIN outreach
  - Which are the active research communities that could benefit from our infrastructure?

• Overview of key families of resources and tools in CLARIN
  - What do we have to offer?
  - Are there gaps in our infrastructure?
  - How can we improve our infrastructure?

• Overview of requests and announcements on LRT on CorporaList
  - What researchers need?
  - Can we help?
  - What researchers are developing?
Deposit to CLARIN
Depositing services

https://www.clarin.eu/content/depositing-services

• advantages:
  - free and safe
  - respect your license
  - easy to find, giving you maximal visibility of your work
  - easy to cite, giving you maximal credit for your work
  - easy to enrich

• find a CLARIN centre near you:
  - help researchers to store their resources
  - assist with the technical and organisational details
Deposit your resource in 9 simple steps

1. Log in
2. Start a new submission
3. Select type of your submission
4. Describe your item
5. Upload files
6. Select licenses
7. Leave a note
8. Review your submission
9. Submit
Veni, vidi, CLARIN!

darja.fiser@ff.uni-lj.si