CLARIN for Linguists
Concluding Overview

Jan Odijk
LOT Summerschool
Nijmegen, 2014-06-27
Overview

• This Course
• There is much more in CLARIN!
• There is even more NOT yet in CLARIN
This Course

- There is much more in CLARIN!
- There is even more NOT yet in CLARIN
GOALS OF THE COURSE

• Introduction to the CLARIN infrastructure
• Introduction to selected functionality within the CLARIN infrastructure
  – Focus on linguistics
  – Focus on results of CLARIN-NL
• Hopefully make you enthusiastic to use the CLARIN infrastructure and functionality in it
  – Because it improves your research
• If you use it, provide feedback
  – Helpdesk: helpdesk@clarin.nl
  – Service-specific user group lists
Course programme

• Introduction

• General functionality
  – General background and set-up of CLARIN
  – Searching for data
  – Searching in data
  – Storing Data in CLARIN

• 3 specific services
  – GrETEL for searching for syntactic structures
  – TTNWW for automatically enriching data with linguistic annotations
  – Cornetto for searching in a lexico-semantic resource

• With focus on syntax and semantics
This Course

- There is much more in CLARIN!
- There is even more NOT yet in CLARIN
<table>
<thead>
<tr>
<th>CLARIN-NL – Linguistics other subfields</th>
</tr>
</thead>
<tbody>
<tr>
<td>Historical linguistics</td>
</tr>
<tr>
<td>Language Acquisition</td>
</tr>
<tr>
<td>Morphology</td>
</tr>
<tr>
<td>Phonology</td>
</tr>
<tr>
<td>Typology</td>
</tr>
</tbody>
</table>
Overview

• CLARIN-NL – Linguistics
  – Lexical Data
  – Linguistically Annotated Corpora
  – Search in Lexical data
  – Search in linguistically Annotated Corpora
  – Annotation & Related Tools
  – Processing data

• CLARIN-NL other than linguistics

• Data and services from other countries
• This Course
• There is much more in CLARIN!
  ➢ There is even more NOT yet in CLARIN
Not all data (even some crucial data) are visible via the VLO or via Metadata Search

Very few tools and web services are currently visible via the VLO

Many tools are still prototypes or first versions

There are good search facilities for some individual resources but not for all

The search facilities so far are aimed at a single resource, or a small group of closely related resources.

Federated content search, which enables one to search with one query in multiple, quite diverse, resources, is still being worked on but difficult
Still Desired

- **TTNWW combined with analysis and search facilities**
- **Analysis and visualisation services for large search results**
- **Single query to search**
  - in all Dutch lexical resources
  - in all Dutch PoS-tagged corpora
  - in all Dutch treebanks
- **Chaining Search**
- **Parameterized queries**
- **Facilities to improve replicability**
- Fully CLARIN-compatible viewers and editors for annotated linguistic data
- Automatically retrainable enrichment services
- ...
A successor project is needed!

CLARIAH [www.clariah.nl](http://www.clariah.nl)

Proposal submitted in 2013

Decision announced on July 1\textsuperscript{st}, 2014.

If awarded project will start Jan 1\textsuperscript{st}, 2015

When CLARIN-NL ends
Thanks for Joining us in this Course
Lexical data

- **COAVA project** Curated Dutch Dialect Dictionaries for **Brabant** and **Limburg**
- **DuELME project** pre-CLARIN data and *interface* new *metadata* (data via the HLT-Agency)
Linguistically Annotated Data

- Database of the Longitudinal Utrecht Collection of English Accents (D-LUCEA) curated data ... expected in September
- **DISCAN** text corpus enriched with discourse Annotation and its metadata
- **EXILSEA project** enhancements of the Corpus NGT, the world’s first open access sign language corpus, by updating the existing IMDI metadata to CLARIN-standard CMDI descriptions using bilingual ISOcat categories ... expected in 2014
Linguistically Annotated Data

- **FESLI** curated specific language impairment data ...
- **INPOLDER** curated data
- **IPROSLA** project [website](#) and [metadata](#) via the VLO (license needed for access to the data)
- **LAISEANG** language documentation data ... expected in 2014
Linguistically Annotated Data

• **MIMORE project** metadata for **DiDDD**, **Dynasand**, and **GTRP** via Metadata Search (Use the MIMORE Search Engine to search in these data)

• **NEHOL project** Negerhollands data (via the Virtual Language Observatory)

• WIVU Hebrew Text Database curated by the **SHEBANQ project** ... expected in 2014
• **VALID project** curated five existing, digital data sets of language pathology data collected in the Netherlands, primarily on Dutch ... expected in 2014

• **VU-DNC** project  **Data and its metadata and Documentation**
• **COAVA** application Dialect Lexicon Browser

• **Cornetto-LMF-RFD project Interface** to Cornetto

• **DuELME project interface**

• **GTB** (Integrated Language Bank) including the **WFT-GTB** Frisian dictionary in the GTB) (Dutch interface)
Search in Lexical Resources

- GrNe project search interface for searching in a Greek-Dutch dictionary (letter Π only), Dutch interface
- SignLinc subproject enhancements to LEXUS (version 3.00 and higher) and ELAN tool (version 4.00 and higher) (SignLinC website)
COAVA application CHILDES browser

Search interface (beta) to Corpus Gysseling provided by INL

FESLI Search application for search in language selective impairment acquisition data... expected in October 2013
• **Mimore search** engine through 3 Dutch dialect databases and a **presentation of a demonstration scenario**

• **OpenSoNaR** tool for exploring the SoNaR-500 reference corpus ... expected in 2014

• **SHEBANQ** web application demonstrator that enables researchers to perform linguistic queries on the curated WIVU web resource and preserve significant results as annotations to this resource ... expected in 2014
SignLinc subproject enhancements to LEXUS (version 3.00 and higher) tool
– TDS-Curator project Access to the Typological Database System (TDS)
• **AAM-LR CLAM Webservice** supporting annotation of audio-files

• Extensions of the ELAN and ANNEX applications for the annotation and display of time-based resources by the **ColTime project** ... expected in 2014
• **EXILSEA project** enhancements of ELAN and ANNEX with the multilingual features of ISOCAT ... expected in 2014

• **Multicon** enhancements for multimodal collocations in new versions of the **ELAN** and **ANNEX** tools together with a **screencast** explaining the new functionality
• **SignLinc subproject** enhancements to ELAN tool (version 4.00 and higher) ([SignLinC website](http://signlinc.signlang.org/))

• **Transcription Quality Evaluation (TQE) Tool** and its CMDI metadata made by the **TQE subproject**
• @PhilosTEI open source, web-based, user-friendly workflow from textual digital images to TEI...expected in 2014

• Adelheid project website, web service for PoS-tagging, tokenizer, lexicon and editor/visualiser

• Gabmap website for analysis of dialect variation and introduction video (by the ADEPT subproject)
• INPOLDER project parsing application for Historical Dutch (also includes a workflow in which it is combined with the Adelheid Tagger)

• Namescape project Named Entity Tagger

• TICCLops project application and demonstrator for orthographic normalisation
• TTNWW enables automatic enrichment of text corpora
  – But that is just a first step. No researcher is interested in that in itself
  – It must be followed by e.g.
    • Search in the enriched data, or
    • Analysis of the enriched data (statistics, etc)
  – But using the TTNWW output in Search services is currently not possible yet
  – Analysis is possible but only in limited ways
• facilities for this are desired
• Search queries applied to large data often yields large results
  – Cannot be analyzed by hand
  – \(\Rightarrow\) flexible Workflows for search – analysis services – visualisation services
  • Each search tool should yield output formats suitable for existing analysis software (e.g. CSV format for input to Excel, Calc, R, SPSS, ...)
– (and/or) Search can apply to its own output
  • Incremental refinement
• Full-fledged federated content search is not possible yet
• But much simpler cases are not possible either
  – Search with one query in multiple Dutch lexical resources:
    • CGN-lexicon, CELEX, GTB, Cornetto, DuELME-LMF, ...
  – Search with one query in multiple Dutch pos-tagged text corpora
    • CGN, D-COI, SONAR-500, VU-DNC, Childes corpora, ...
  – Search with one query in multiple Dutch treebanks
    • CGN treebank, LASSY-Small, LASSY-Large
• This might be an incremental way to get to full-fledged federated content search
• [MPI’s TROVA offers some of the functionality described here]
Still desired

• Chaining Search, e.g.
  – GrETEL followed by semantic filtering (Cornetto)
    • Bare noun phrases where the head noun is count
    • N N constructions where first N indicates a quantity
  – GrETEL followed by morphological potential filtering (CGN/SONAR/CELEX lexicon)
    • Het adj-ø N where adj has no e-form potential
  – GrETEL followed by phonological filtering
    • Het adj-ø N where adj ends in /C+$C+$C+/
Parameterized queries (batch queries)

- give me all example sentences containing any word from a given set of `synonyms' of the adverb *zeer* (itself derived from Cornetto) and, for each word, statistics on the categories it modifies

• **Replicability**
  – Student tried to replicate similarity measure calculations on Wordnet of Patwardhan and Pedersen (2006) and Pedersen (2010)
  – in an excellent team: Piek Vossen and his research group
  – With help of one the original authors: Ted Pedersen
  – Using the exact same software and data

• They failed to reproduce the original results!

• Reason: ‘properties which are not addressed in the literature may influence the output of similarity measures’

• Many experiments and Pedersen’s unpublished intermediate results to find out
  – the original settings of all parameters (e.g. treatment of ties in Spearman $\rho$ )
  – Which aspects of the data had been used and how
One step towards a solution for this

- All tools must allow input of metadata associated with data
- All tools must provide provenance data
- All tools must provide a list with settings of all parameters (also usable as an input parameter, ‘configuration file’) as part of the provenance data
- All tools must generate new metadata for its results based on the input metadata, the generated provenance data, and possibly some manual input of a user

<table>
<thead>
<tr>
<th>Day</th>
<th>Who</th>
<th>What</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mon 23rd</td>
<td>Jan Odijk</td>
<td>Introduction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Illustrative Usage Case</td>
</tr>
<tr>
<td>Tue 24th</td>
<td>Liesbeth Augustinus</td>
<td>Treebank Mining By Example</td>
</tr>
<tr>
<td></td>
<td>Frank Van Eynde</td>
<td>Number agreement in copular constructions</td>
</tr>
<tr>
<td>Wed 25th</td>
<td>Marc Kemps-Snijders</td>
<td>Enriching your data with grammatical information using TTNWW</td>
</tr>
<tr>
<td>Thu 26th</td>
<td>Piek Vossen &amp; Isa Maks</td>
<td>Cornetto</td>
</tr>
<tr>
<td>Fri 26th</td>
<td>Jan Odijk</td>
<td>Storing data in CLARIN</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Concluding Overview</td>
</tr>
</tbody>
</table>