Linguistics with CLARIN
Concluding Overview

Jan Odijk

LOT Winterschool

Amsterdam, 2014-01-16

Overview

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

Overview

- **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
- A few specific services
  - OpenSONAR for searching for words, word combinations and their morpho-syntactic properties
  - MIMORE for search in micro-comparative databases and analysis of the search results
  - TTNWW for automatically enriching data with linguistic annotations
  - PaQu for searching in and analysis of data thus enriched
- With focus on syntax and semantics

Overview

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

More in CLARIN

CLARIN-NL –Linguistics other subfields		
Historical linguistics	Dialectology	Discourse Studies
Language Acquisition	Language Documentation	Lexicology
Morphology	Morpho-syntax	Phonetics
Phonology	Semantics	Sign Language
Typology		

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

Overview

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

Still Desired

- 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
- …

Conclusion

- A successor project is needed!
- CLARIAH www.clariah.nl
- Proposal submitted in 2013
- Approval obtained on July 1st, 2014. (12 m euro)
- Project started Jan 1st, 2015 (until 1 Jan 2019)
- CLARIN-NL is finishing

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
- 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 end of 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

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

Linguistically Annotated Data

- VALID project curated five existing, digital data sets of language pathology data collected in the Netherlands, primarily on Dutch
- VU-DNC project Data and its metadata and Documentation

Search in Lexical Resources

- 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)

Search in Linguistically Annotated Corpora

- COAVA application CHILDES browser
- Search interface (beta) to Corpus Gysseling provided by INL
- FESLI Search application for search in language selective impairment acquisition data

Search in Linguistically Annotated Corpora

- Mimore search engine through 3 Dutch dialect databases and a presentation of a demonstration scenario
- OpenSoNaR tool for exploring the SoNaR-500 reference corpus
- 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

Search in Linguistically Annotated Corpora

- SignLinc subproject enhancements to LEXUS (version 3.00 and higher) tool
  - TDS-Curator project Access to the Typological Database System (TDS)

Annotation & Related Tools

- 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 end of 2014

Annotation & Related Tools

- EXILSEA project enhancements of ELAN and ANNEX with the multilingual features of ISOCAT … expected end of 2014
- Multicon enhancements for multimodal collocations in new versions of the ELAN and ANNEX tools together with a screencast explaining the new functionality

Annotation & Related Tools

- SignLinc subproject enhancements to ELAN tool (version 4.00 and higher) (SignLinC website)
- Transcription Quality Evaluation (TQE) Tool and its CMDI metadata made by the TQE subproject

Processing data

- @PhilosTEI open source, web-based, user-friendly workflow from textual digital images to TEI…expected in Feb 2015
- 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)

Processing data

- 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

Still desired

- 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
- Initial Work on it is ongoing: PaQu and AutoSearch projects

Still desired

- Search queries applied to large data often yields large results
  - Cannot be analyzed by hand
  - ➔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

Still desired

- 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+/

Still desired

- 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
  - allemachtig-adv-2 beestachtig-adv-2 bijzonder-a-4 bliksems-adv-2 bloedig-adv-2 bovenmate-adv-1 buitengewoon-adv-2 buitenmate-adv-1 buitensporig-adv-2 crimineel-a-4 deerlijk-adv-2 deksels-adv-2 donders-adv-2 drommels-adv-2 eindeloos-a-3 enorm-adv-2 erbarmelijk-adv-2 fantastisch-adv-6 formidabel-adv-2 geweldig-adv-4 goddeloos-adv-2 godsjammerlijk-adv-2 grenzeloos-adv-2 grotelijks-adv-1 heel-adv-5 ijselijk-adv-2 ijzig-a-4 intens-adv-2 krankzinnig-adv-3 machtig-adv-4 mirakels-adv-1 monsterachtig-adv-2 moorddadig-adv-4 oneindig-adv-2 onnoemelijk-adv-2 ontiegelijk-adv-2 ontstellend-adv-2 ontzaglijk-adv-2 ontzettend-adv-3 onuitsprekelijk-adv-2 onvoorstelbaar-adv-2 onwezenlijk-adv-2 onwijs-adv-4 overweldigend-adv-2 peilloos-adv-2 reusachtig-adv-3 reuze-adv-2 schrikkelijk-adv-2 sterk-adv-7 uiterst-adv-4 verdomd-adv-2 verdraaid-a-4 verduiveld-adv-2 verduveld-adv-2 verrekt-adv-3 verrot-adv-3 verschrikkelijk-adv-3 vervloekt-adv-2 vreselijk-adv-5 waanzinnig-adv-2 zeer-adv-3 zeldzaam-adv-2 zwaar-adv-10
  - (made possible in OpenSONAR)

Still desired

- 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 ρ)
  - Which aspects of the data had been used and how

Still desired

- 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

- Fokkens, A., M. van Erp, M. Postma, T. Pedersen, P. Vossen & N. Freire 'Offspring from Reproduction problems: What Replication Failure Teaches Us', *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics,* pages 1691–1701, Sofia, Bulgaria, 2013.

Course programme

Day	Who	What
Mon 12th	Jan Odijk	Introduction
		Illustrative Usage Case
Tue 13th	Jan Odijk	Search and Analysis with OpenSONAR
Wed 14th	Sjef Barbiers	Search and Analysis in micro-comparative databases with MIMORE
Thu 15th	Jan Odijk	Enrichment and Search (TTNWW, PaQu)
Fri 16th	Jan Odijk	Storing data in CLARIN
		Concluding Overview