# What CLARIN has to offer to Linguists

Jan Odijk

TIN-dag

Utrecht, 2015-02-07

# Overview

- What is CLARIN?
- What CLARIN has to offer to linguists
- How you can learn to use the functionality offered
- Current Status and Near Future

# Overview

- **What is CLARIN?**
- What CLARIN has to offer to linguists
- How you can learn to use the functionality offered
- Current Status and Near Future

# CLARIN Infrastructure

- A **research infrastructure** for **humanities researchers** who work with digital **language resources**
  - **Research infrastructure:** facilities, resources and related services used by the scientific community to conduct top-level research
  - **Humanities researchers:** here focus on **linguists**
  - **Language resources:** lexicons, corpora, databases, …; text, audio, video, ….
  - **Language** in various functions: object of inquiry, component of identity, means of communication, carrier of cultural content

# CLARIN Infrastructure

- The CLARIN infrastructure
  - Is distributed: implemented in a network of CLARIN centres
  - Is virtual: it provides services electronically (via the internet)
- The CLARIN infrastructure
  - Is still under construction
    - Highly incomplete
    - Fragile in some respects
  - But you can use many parts already

# CLARIN in Europe

- Prepared by CLARIN preparatory project (2008-2011)
  - coordinated by Utrecht University
- From Feb 2012 coordinated by CLARIN ERIC, hosted by the Netherlands
  - ERIC: a legal entity at the European level specifically for research infrastructures
  - 12 ERIC members (=countries), 1 observer, and growing

# Overview

- What is CLARIN?
- **What CLARIN has to offer to linguists**
- How you can learn to use the functionality offered
- Current Status and Near Future

# CLARIN Infrastructure

- CLARIN-NL
  - Has set up a network of centres that offer access to data, tools and services
  - Has adapted existing data and tools to improve interoperability
  - Has created new easy and user-friendly tools for exploring, searching, analysing, visualising and enriching data
  - Has not created any new data

# CLARIN Infrastructure

- CLARIN offers
  - Applications to search for data and tools
  - Facilities to store data and tools in a sustainable manner and make them available to the whole research community
  - Access to data and tools, and facilities to apply tools to data

# Search For Data and Tools

- CLARIN-NL Portal: http://portal.clarin.nl
  - Faceted search in data and services created by CLARIN-NL
    - Search by research domain, language, tool task, linguistic annotation and several other facets
  - Links to search facilities in the whole CLARIN infrastructure
  - And much more….

# Storing Data and Tools

- Facilities for YOUR research data and tools
  - To make them CLARIN-compatible
  - to create `metadata' for them
  - to make them visible and accessible in CLARIN
  - to securely store them

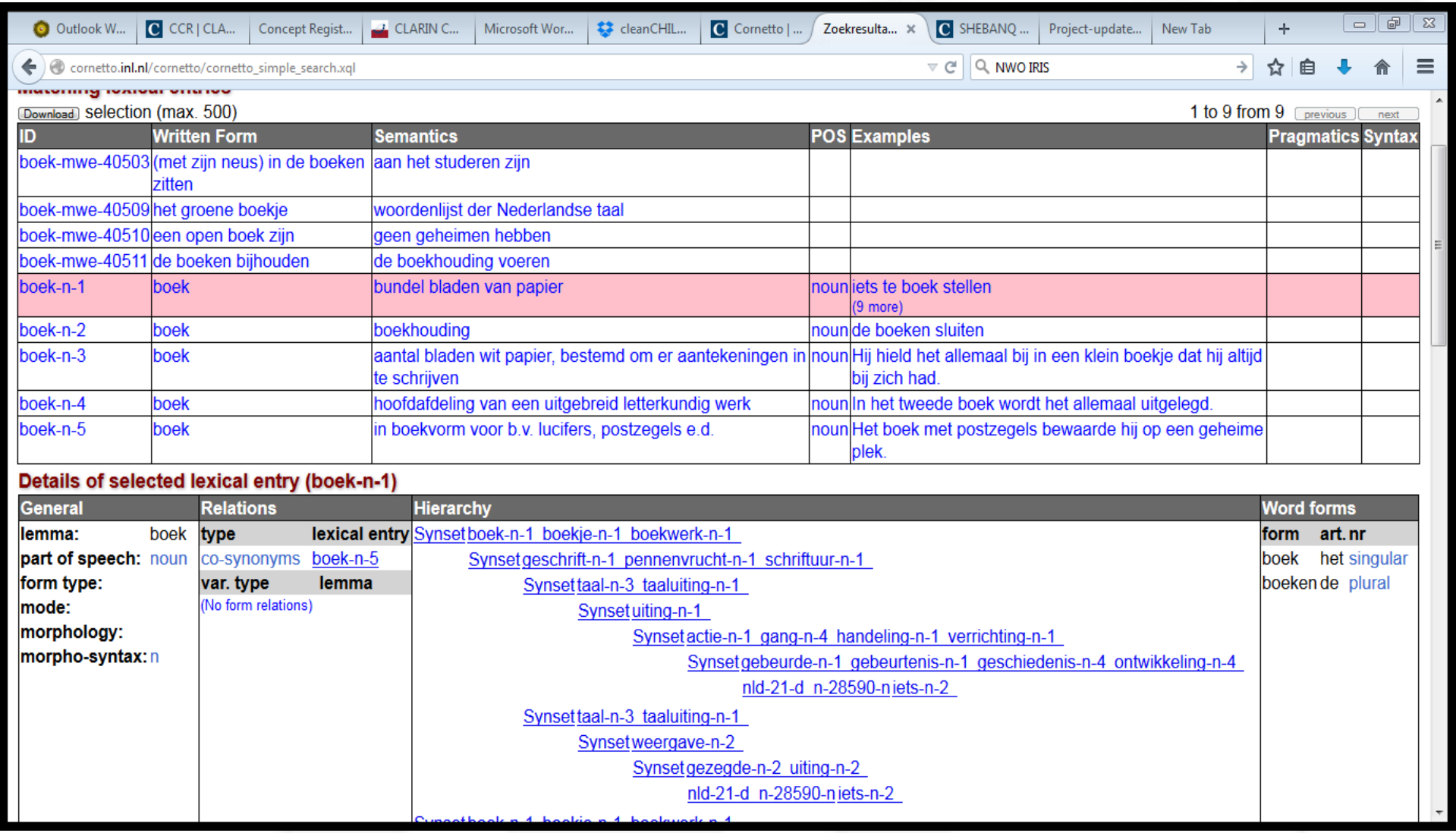
# Data and Tool Types

- CLARIN-NL –Linguistics
  - Lexical Data
  - Linguistically Annotated Corpora
  - Search in Lexical data
  - Search in Linguistically Annotated Corpora
  - Tools for Enriching Your Own data
  - Tools for searching in your enriched data

# Selected Examples

- Interface to the Cornetto lexical semantic database
  - Combination of WordNet and ReferentieBestand Nederlands
  - Search for
    - words,
    - their grammatical and semantic properties
    - Relations between meanings of words (synonymy, hyponomy, antonymy, etc..

# Selected Examples



Outlook W... | CCR | CLA... | Concept Regist... | CLARIN C... | Microsoft Wor... | cleanCHIL... | Cornetto | ... | Zoekresulta... | SHEBANQ ... | Project-update... | New Tab

cornetto.inl.nl/cornetto/cornetto_simple_search.xql

NWO IRIS

Download selection (max. 500) 1 to 9 from 9 previous next

| ID | Written Form | Semantics | POS | Examples | Pragmatics | Syntax |
|---|---|---|---|---|---|---|
| boek-mwe-40503 | (met zijn neus) in de boeken zitten | aan het studeren zijn | | | | |
| boek-mwe-40509 | het groene boekje | woordenlijst der Nederlandse taal | | | | |
| boek-mwe-40510 | een open boek zijn | geen geheimen hebben | | | | |
| boek-mwe-40511 | de boeken bijhouden | de boekhouding voeren | | | | |
| boek-n-1 | boek | bundel bladen van papier | noun | iets te boek stellen (9 more) | | |
| boek-n-2 | boek | boekhouding | noun | de boeken sluiten | | |
| boek-n-3 | boek | aantal bladen wit papier, bestemd om er aantekeningen in te schrijven | noun | Hij hield het allemaal bij in een klein boekje dat hij altijd bij zich had. | | |
| boek-n-4 | boek | hoofdafdeling van een uitgebreid letterkundig werk | noun | In het tweede boek wordt het allemaal uitgelegd. | | |
| boek-n-5 | boek | in boekvorm voor b.v. lucifers, postzegels e.d. | noun | Het boek met postzegels bewaarde hij op een geheime plek. | | |

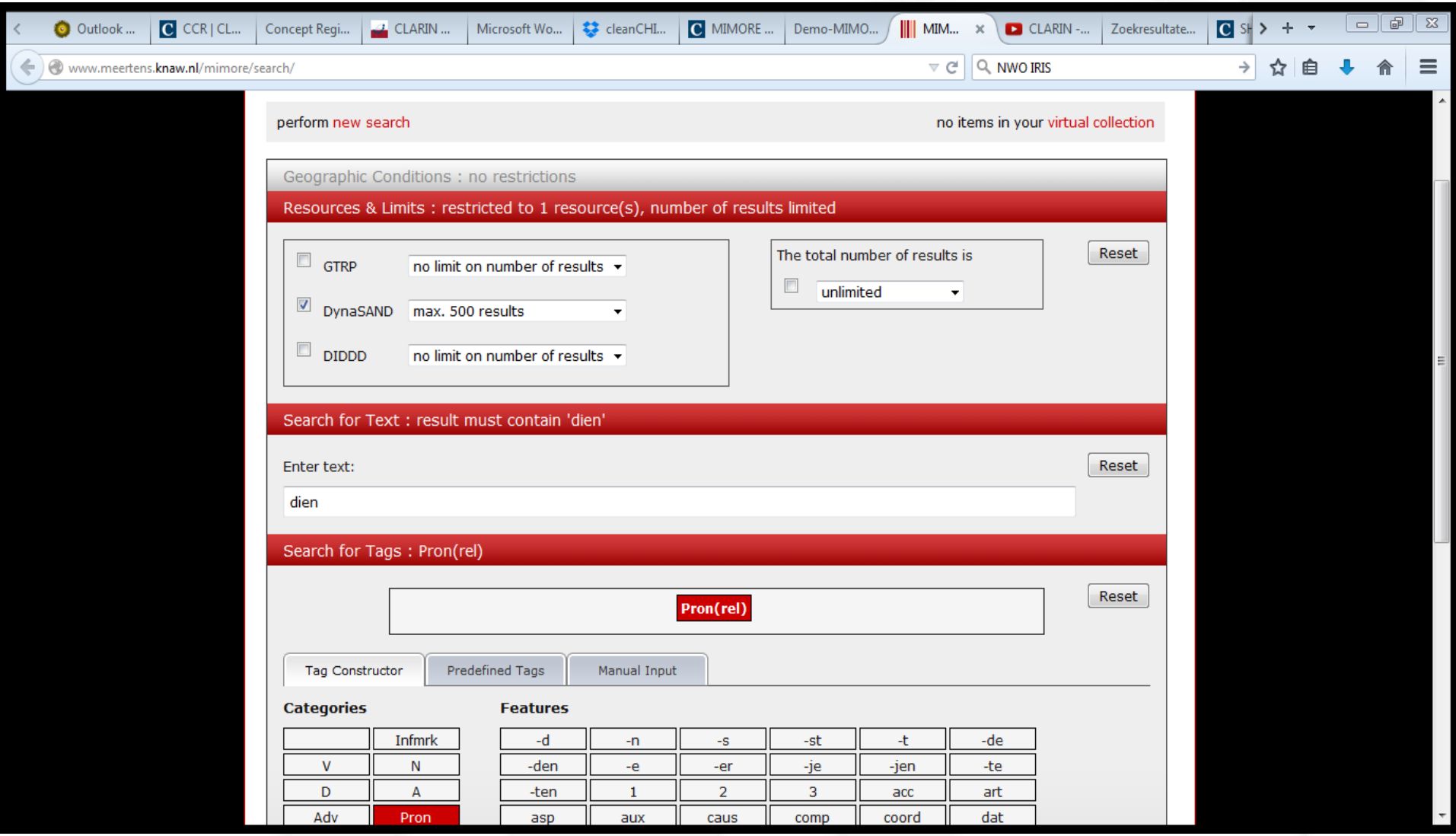
**Details of selected lexical entry (boek-n-1)**

| General | | Relations | | Hierarchy | Word forms | | |
|---|---|---|---|---|---|---|---|
| lemma: | boek | type | lexical entry | Synset boek-n-1 boekje-n-1 boekwerk-n-1 | form | art. nr | |
| part of speech: | noun | co-synonyms | boek-n-5 | Synset geschrift-n-1 pennenvrucht-n-1 schriftuur-n-1 | boek | het | singular |
| form type: | | var. type | lemma | Synset taal-n-3 taaluiting-n-1 | boeken | de | plural |
| mode: | | (No form relations) | | Synset uiting-n-1 | | | |
| morphology: | | | | Synset actie-n-1 gang-n-4 handeling-n-1 verrichting-n-1 | | | |
| morpho-syntax: | n | | | Synset gebeurde-n-1 gebeurtenis-n-1 geschiedenis-n-4 ontwikkeling-n-4 | | | |
| | | | | nld-21-d n-28590-n iets-n-2 | | | |
| | | | | Synset taal-n-3 taaluiting-n-1 | | | |
| | | | | Synset weergave-n-2 | | | |
| | | | | Synset gezegde-n-2 uiting-n-2 | | | |
| | | | | nld-21-d n-28590-n iets-n-2 | | | |

# Selected Examples

- Mimore search engine
  - Search through 3 Dutch dialect databases (DiDDD, DynaSand, GTRP )
  - Analyze the results
  - presentation of a demonstration scenario

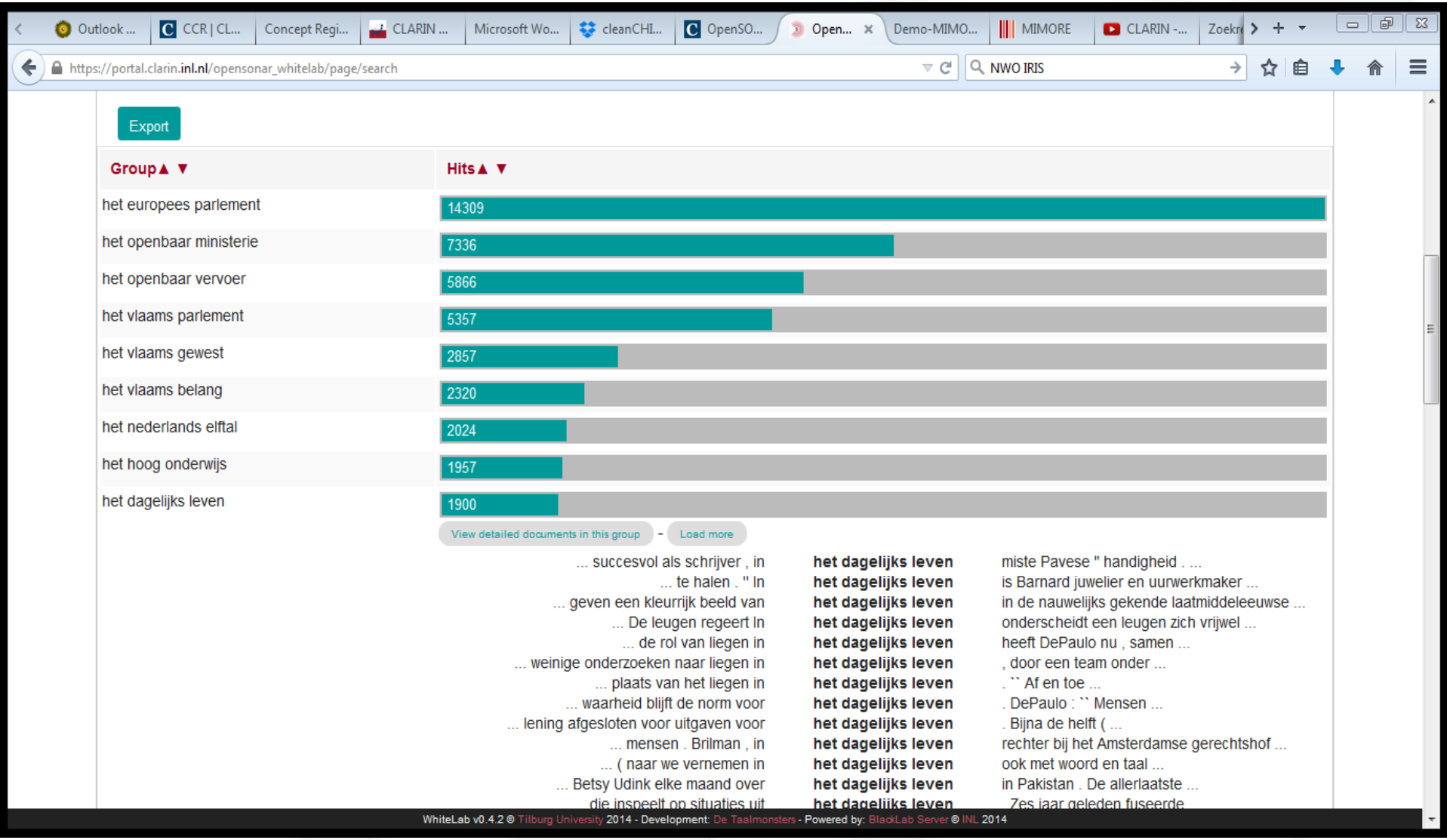
# Selected Examples



# Selected Examples

- OpenSoNaR Interface
  - To the SONAR and the SONAR New Media corpus
  - Written Dutch (> 500 m tokens)
  - Exploration and Search
  - For extended pos-tags, word forms, lemmas
  - Combined with metadata for each document
  - 4 different interfaces from simple to expert
  - Login with the account of your own organisation

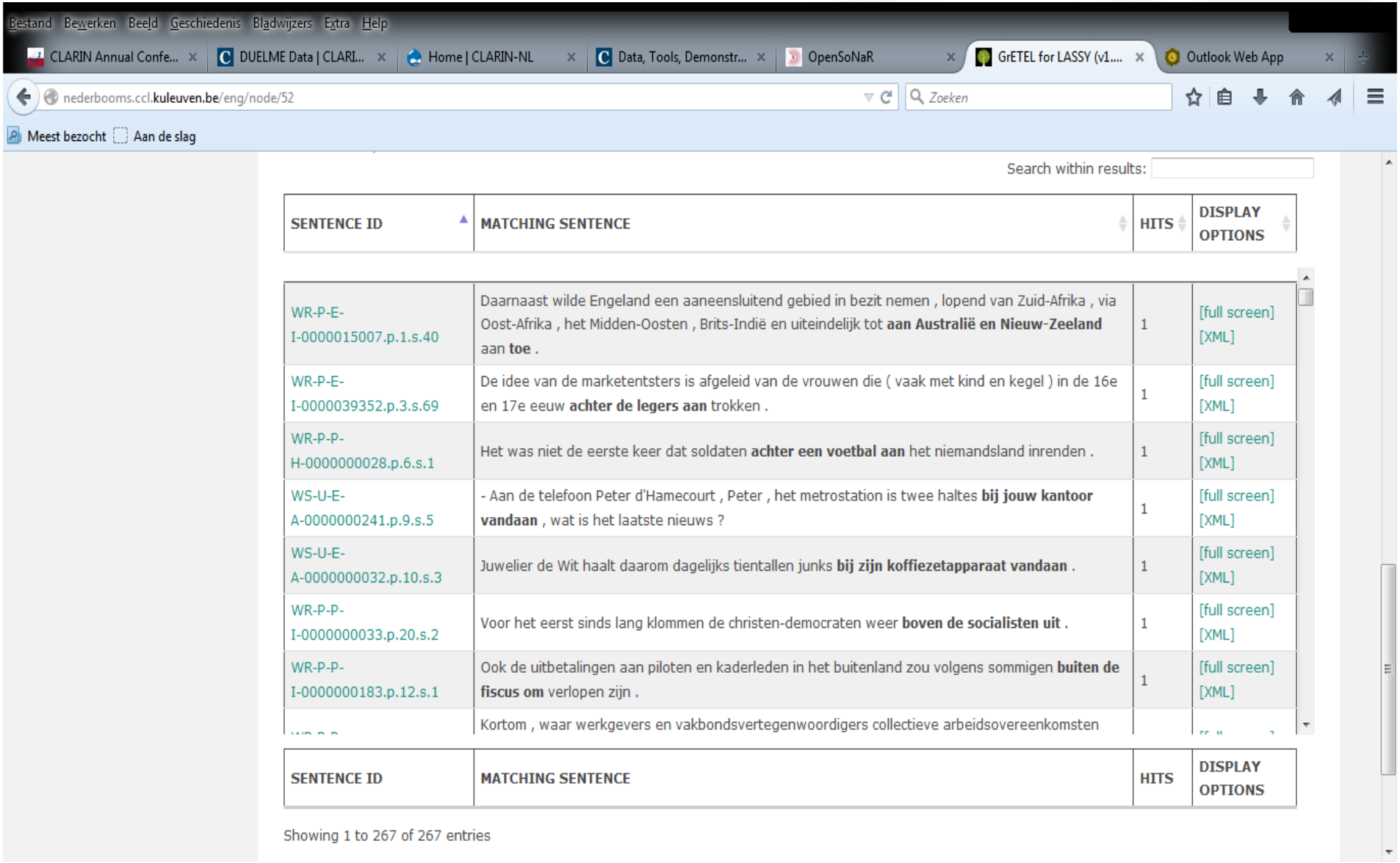
# Selected Examples



# Selected Examples

- GrETEL
  - Example-based search in treebanks
  - Lassy-Small (1 m tokens, manually verified)
  - Spoken Dutch Corpus (CGN) – 1 m tokens, manually verified
  - SONAR Corpus (500 m tokens, automatically parsed)
  - Result of cooperation with Flanders

# Selected Examples




| SENTENCE ID | MATCHING SENTENCE | HITS | DISPLAY OPTIONS |
|---|---|---|---|
| WR-P-E-I-0000015007.p.1.s.40 | Daarnaast wilde Engeland een aaneensluitend gebied in bezit nemen , lopend van Zuid-Afrika , via Oost-Afrika , het Midden-Oosten , Brits-Indië en uiteindelijk tot **aan Australië en Nieuw-Zeeland** aan **toe** . | 1 | [full screen] [XML] |
| WR-P-E-I-0000039352.p.3.s.69 | De idee van de marketentsters is afgeleid van de vrouwen die ( vaak met kind en kegel ) in de 16e en 17e eeuw **achter de legers aan** trokken . | 1 | [full screen] [XML] |
| WR-P-P-H-0000000028.p.6.s.1 | Het was niet de eerste keer dat soldaten **achter een voetbal aan** het niemandsland inrenden . | 1 | [full screen] [XML] |
| WS-U-E-A-0000000241.p.9.s.5 | - Aan de telefoon Peter d'Hamecourt , Peter , het metrostation is twee haltes **bij jouw kantoor vandaan** , wat is het laatste nieuws ? | 1 | [full screen] [XML] |
| WS-U-E-A-0000000032.p.10.s.3 | Juwelier de Wit haalt daarom dagelijks tientallen junks **bij zijn koffiezetapparaat vandaan** . | 1 | [full screen] [XML] |
| WR-P-P-I-0000000033.p.20.s.2 | Voor het eerst sinds lang klommen de christen-democraten weer **boven de socialisten uit** . | 1 | [full screen] [XML] |
| WR-P-P-I-0000000183.p.12.s.1 | Ook de uitbetalingen aan piloten en kaderleden in het buitenland zou volgens sommigen **buiten de fiscus om** verlopen zijn . | 1 | [full screen] [XML] |
| | Kortom , waar werkgevers en vakbondsvertegenwoordigers collectieve arbeidsovereenkomsten | | |


| SENTENCE ID | MATCHING SENTENCE | HITS | DISPLAY OPTIONS |
|---|---|---|---|


Showing 1 to 267 of 267 entries

# Selected Examples

- PaQu
  - Enrich your (Dutch) data with full parses via TTNWW or PaQu
  - Upload your data in PaQu
  - Search in your enriched data with PaQu
  - Intermediate version available
    - See also my presentation at 16:00hrs in 3.02

# Overview

- What is CLARIN?
- What CLARIN has to offer to linguists
- **How you can learn to use the functionality offered**
- Current Status and Near Future

# Education & Training

- How do you learn to use these tools?
  - Courses / tutorials regularly organized
  - LOT summer / winter school courses
  - Demonstration scenarios and/or screen casts
    - E.g. for Gabmap GrETEL OpenSONAR
  - Educational modules via the portal:
    - https://dev.clarin.nl/node/CLARIN%20Educational%20Packages
  - Helpdesk: helpdesk@clarin.nl

# Education & Training

- Do you want to know more?
  - Visit the CLARIN-NL portal
  - View the CLARIN-NL movies
  - Ask me (or others) during this TIN-dag
  - I can demonstrate some tools during the breaks
  - Visit my presentation 16:00hrs in 3.02 (on PaQu)
  - Visit the CLARIN-NL Final event:
    - March 13, 2015, Beeld & Geluid, Hilversum (in the morning)

# Overview

- What is CLARIN?
- What CLARIN has to offer to linguists
- How you can learn to use the functionality offered
- **Current Status and Near Future**

# Current Status & Future

- CLARIN-NL will finish by April 1, 2015
- Successor project CLARIAH www.clariah.nl
  - Jan 1st, 2015 – Jan 1st, 2019
  - 12 million euro budget
  - Linguistics is one of the 3 core disciplines (with social-economic history and media studies)
  - Kick-off March 13, 2015, Beeld & Geluid, Hilversum (in the afternoon)

# Thanks for Attention!

# Subdisciplines

| CLARIN-NL – Linguistics subdisciplines | | |
|---|---|---|
| Historical linguistics | Dialectology | Discourse Studies |
| Language Acquisition | Language Documentation | Lexicology / Lexicography |
| Morphology | Morpho-syntax | Phonetics |
| Phonology | Pidgin & Creole Studies | Semantics |
| Sign Language | Specific Language Impairment (SLI) | Syntax |
| Typology | … | |

# Languages Covered

| CLARIN-NL – Languages | | |
|---|---|---|
| Dutch | English | German |
| Frisian | Dutch Sign Language | Classical Greek |
| French | Hebrew | Aramaic |
| Syriac | > 50 languages from Insular South East Asia and West New Guinea | And more… |

- And many more (> 2,000) from all CLARIN countries

# What you can do

| CLARIN-NL – Functionality Offered | | |
|---|---|---|
| Searching | Browsing | Analysis |
| Corpus Exploration | Annotation | Tokenization |
| Pos-tagging | Lemmatization | Orthographic normalisation |
| Grammatical relation assignment | Named entity recognition | Chunking |
| Co-reference assignment | Multiword unit assignment | Parsing |
| Visualisation | Diarisation | Speech recognition |