

CLARIAH WP3: Brief Status Update

Jan Odijk 2017-03-10

- INT: Search in corpora and lexica
- Meertens: Search in corpora and associated metadata, concept registry
- Radboud Universiteit Nijmegen: FoLIA and associated tools, various language processing tools, metadata curation
- RUG: search in treebanks
- UU: metadata curation, search in treebanks
- VU: (semantic) lexicons, tools for extracting information from text

/instituut voor de Nederlandse taal/

OpenSoNaR: search engine for SoNaR Corpus

Problematic queries faster

- negative clauses 10+ times
- Search for XML tags 5+ times
- Other performance improvements

Extended Pos independently searchable

- “find plural forms”
- “find imperatives of verbs”

More grouping-/sorting options

- “2nd word to the left of the hit”

Random sampling

- show 10% of the hits found

More to come!

The screenshot displays the OpenSoNaR search engine interface. At the top, there are navigation tabs for 'Home', 'Verken', and 'Zoek', with 'Zoek' being the active tab. Below the navigation, there are filters for 'Simpel', 'Uitgebreid', 'Geavanceerd', and 'Expert', with 'Resultaten' being the active filter. The 'openSoNaR' logo is visible in the top right corner.

The main content area shows a table of search tasks (Zoekopdrachten) and a detailed view of search results (Zoekresultaten).

#	Zoekopdracht	binnen	Metadata filters	Groepering	Status	Hits	Documenten	edit	x
1	[word="powerpoint"]	document			FINISHED	348	238	edit	x
2	[word="dia"]	document			FINISHED	697	428	edit	x
3	[word="dodo"]	document			FINISHED	498	256	edit	x

The 'Zoekresultaten' section shows a table with columns for 'Context links', 'Hit text', 'Context rechts', 'lemma', and 'woordsoort'. The 'Hit text' column contains the word 'PowerPoint' in various cases and forms. The 'Context links' column shows the surrounding text for each hit. The 'Context rechts' column shows the text to the right of the hit. The 'lemma' column shows the base form of the word, and the 'woordsoort' column shows the part of speech.

/instituut voor de Nederlandse taal/

INT hosts WebCelex (June 2015)

- Web application and service to the CELEX lexicon
- Manage WebCelex, incl. browser compatibility checks, GUI checks, basic software and data checks.
- put the application behind a CLARIN login

Taalportaal uses WebCelex

The screenshot displays the WebCelex interface, which is part of the Taalportaal project. The top navigation bar includes the logo 'WebCelex' and the text '/instituut voor de Nederlandse taal'. Below this, the page title 'Dutch Lemmas' is visible. A search bar is present with the text 'Query' and a search button. The main content area features a blue header for 'taalportaal' with the tagline 'the linguistics of Dutch, Frisian and Afrikaans online'. A search bar is also located here. The page is organized into a sidebar with tabs for 'Dutch', 'Frisian', and 'Afrikaans'. The 'Phonology' section is active, showing a list of sub-topics: 'Segment inventory', 'Phonotactics', 'Syllable level', and 'Onsets'. The 'Onsets' section is selected, displaying the title 'Onset: sequences of more than two consonants' and a brief description: 'Dutch SYLLABLE ONSETS consist of maximally three SEGMENTS. The first segment of such a three-consonant-sequence has to be an /s/'. Below this, a section titled 'The following Celex queries have been constructed for this topic:' lists two queries: 'This query finds 3-phoneme onsets, i.e. substrings preceded by a boundary marker and starting of /s/ and at least two consonants' with the query '(MorphStatus !~ /C/) && (PhonSylCLX =~ Abs[trbcdgfhjklmnpqrstvwxyz][trbcdgfhjklmnpqrstvwxyz])' and 'A-class (long) vowels in open syllables, i.e. A-class vowels followed by a boundary symbol' with the query '(MorphStatus !~ /C/) && (PhonSylCLX =~ /[aoiyue]\|b/)'. The bottom of the page shows a 'Morphology' section with the example 'spleet /spl-/ 'fissure''.

Meertens: MTAS

- MTAS = Multi Tier Annotation Search : Search engine for annotated text corpora
- <http://www.nederlab.nl/onderzoeksporaal/>
- CQL, KWIC, statistics, frequency lists, grouping, combined metadata/content search
- Various annotation formats (configurable indexing process) - FoLiA, ISO 24624:2016, and others
- Large data volumes
- Integrated with Nederlab framework: 14.5 m docs, 8.7 b tokens, 70 annotation tiers
- Source code: Github <https://github.com/meertensinstituut/mtas>
- Documentation: <https://meertensinstituut.github.io/mtas/>
- Fully automated build and integration environment: <http://www12.meertens.knaw.nl/jenkins/>

Meertens: OpenSKOS

- <http://145.100.58.150/ccr/public/> & <http://145.100.58.150/OpenSKOS-browser/>
- RDF triple store, SKOS Concept schemes, Collections, Concepts, Relations, SPARQL and REST access points for SKOS-based thesauri
- Cooperation:
 - CLARIN EU (co-financing, CCR = CLARIN Concept Registry, CLAVAS)
 - Netwerk Digitaal Erfgoed (Expert Group OpenSKOS)
 - DARIAH (Backbone thesaurus)

Radboud Universiteit Nijmegen

- PICCL Subservices

- SoNaR Spaces- 9 flavors of semantic spaces based on SoNaR-500: word2vec and GloVe. <http://ticclops.uvt.nl/vector/>
- AHA = Anagram Hashing Application- web application/service that derives character confusions and error type statistics from lists of word pairs . <http://ticclops.uvt.nl/AHA/>
- Transcriptor- web service for fuzzy search on NAMES databases, i.e. JRC-Names and Geonames enabling translingual search across scripts (Latin, Cyrillic, Chinese, Arabic,) <http://ticclops.uvt.nl/Transcriptor/>
- TICCL: word bigram (N-gram) correction (for solving split and run-on word problems; more precise short word correction; higher precision overall. Perl prototype being reimplemented in C++. Applied to Parliamentary data in Nederlab

Radboud Universiteit Nijmegen

- Frog (tokenization, lemmatization, morphological analysis, part-of-speech tagging, syntactic chunking, dependency parsing, and named-entity recognition), updates
 - Frog Generator, lemmatized and part-of-speech tagged corpus in any language -> lemmatizer/pos-tagger.
 - Successfully tested with Old Greek
 - more tests with historical and dialectal variants of Dutch are planned.
 - MBMA and MBLEM: morphological analysis and lemmatization: qualitative improvements
 - Refactoring and reprogramming of components + extensions + bug fixes
- Collection and curation of metadata records for collections
 - Proposed a complete attribute set along the way
 - 46 collections have been processed

FLAT FoLiA Linguistic Annotation Tool

<https://github.com/proycon/flat>

- A rich multi-user web-based tool for linguistic annotation of documents; fully open source
- Powered by the FoLiA format, unlocking a rich infrastructure of tools (<https://proycon.github.io/folia>)
- Supports a wide variety of linguistic annotation types
- Retains and visualises document structure; annotation on multiple levels

The screenshot shows the FLAT interface with several panels:

- Modes:** Perspective (Full Document), Text class (current, corrections).
- Legend - Part-of-Speech:** A list of linguistic classes with colored circles: N(soort, ev, basis, zijd, stan), VZ(init), LET(), LID(bep, stan, rest), BW(), WW(vd, vrij, zonder), ADJ(prenom, basis, me-t-e, stan).
- Tree Viewer:** A syntax tree for the sentence "Stemma is een ander woord voor stamboom". The root is "sentence", which branches into "subject verb" and "predicate". "subject verb" branches into "Stemma" and "is". "predicate" branches into "np" (which branches into "een" and "ander woord voor") and "PP" (which branches into "stamboom").
- Global annotations:** A view showing annotations above the text: "(sentence (subject) (verb) (predicate (np (pp)) (pp)) (pp))".
- Local annotations:** A detailed view for the word "Stemma", showing:
 - Text: Stemma
 - Part-of-Speech: N(eigen, ev, basis, zijd, stan)
 - Lemma: stemmas-nl
 - Syntax: sentence, subject
 - Dependency: su, hd: is, dep: Stemma

Different perspectives on the data

Coloured highlight of annotation focus

Visualisation of syntax and morphology trees

- Extensive support for correction of any annotation, including orthography, preserving originals
- Git-powered backend with undo functionality
- Annotator name and timestamps are stored with every annotation.

DEMO VIDEO!



<https://youtu.be/tYF6grtldVQ>

Click words to open the annotation editor

The Annotation Editor shows a text snippet: "In de historische wetenschap wordt zo'n stamboom, onder de naam stemma codicum (verwantschap tussen handschriften weer te geven)". The word "wordt" is highlighted. The editor displays the Part-of-Speech (WW(pv, tgw, met-t) - jij komt, hij speelt, zwijgt) and a list of features: wvorm (persoonsvorm), pvtijd (tegenwoordig), and pvagr (met t). A dropdown menu is open, showing options: enkelvoud, meervoud, and met t.

Add new annotations

Set annotations to span over multiple words

The Annotation Editor shows a text snippet: "Naam Maarten van Gompel, Universiteit Radboud University Nijmegen". The word "Radboud" is highlighted. A dropdown menu is open, showing a list of entities: Organisation, Location - Building, Location - City, Location - Country, Location - Line, Location - Metaphysical, Location - Nature elements, Location - Point, Location - Region, Location - Region - County, Location - Region - Estate, Location - Region - Province, Location - Region - Quarter, Location - Region - State, Location - Street, Location - Transport infrastructure, Location - Unspecified, Miscellaneous, Organisation, Person, Product.

Selection fields are populated by FoLiA Set Definitions referenced from the document, neither FLAT nor FoLiA predefine values, you can define your own!

RUG: PaQu

- Tool to search in syntactically annotated corpora
 - Simple queries: relation between two words: list all nouns that occur as direct object of the verb 'to drink'
 - Complex queries: full support for XPath2: are there sentences in which a fronted comparative adjective licenses an extraposed phrase?
 - *Eerder gaat een kameel door het oog van een naald dan dat een rijke ingaat in het koninkrijk der hemelen*
- Support for Lassy, CGN; Upload your own corpora
- Added
 - support for alternative input formats: Folia, TEI, tar and zip
 - support for meta-data in various formats for new corpora
 - counts of query hits per meta data attributes

RUG: PaQu - Example

- Compare the “green” verb order (*gekocht heeft*) and “red” verb order (*heeft gekocht*) in CGN between Dutch and Flemish speakers:

- **Green:**

items: 1 846
zinnen: 1 766

country — [toelichting bij tabel](#)

per item:		per zin:		
<i>aantal</i>	waarde	<i>aantal</i>	per 100 000	waarde
954	NL	918	1 060	NL
889	BE	845	2 000	BE
3	(leeg)	3	290	(leeg)

tijd: 5s

[download](#)

- **Red:**

items: 1 586
zinnen: 1 477

country — [toelichting bij tabel](#)

per item:		per zin:		
<i>aantal</i>	waarde	<i>aantal</i>	per 100 000	waarde
1 054	NL	995	1 148	NL
529	BE	479	1 134	BE
3	(leeg)	3	290	(leeg)

tijd: 4s

[download](#)

GrETEL-upload extension

- Search in syntactically annotated corpora
- Upload plain-text corpus, incl. metadata
- Parsed with Alpino, indexed in BaseX
- Choose which metadata to use in faceted search in GrETEL3
- Cleaner and Converter for CHAT data and metadata (to be integrated)
- Demo's, posters at Language Science Day (Utrecht), CLIN (Leuven), submitted to DH Benelux

Details for treebank *vklaura*

Components

Short name	Title	# sentences	# words
year1	year1	3,746	14,470
year2	year2	19,881	94,804
year3	year3	17,071	92,258
year4	year4	17,685	97,488
year5	year5	5,936	34,776
		64,319	333,796

Metadata

Field	Type	Minimum value	Maximum value	Facet	Shown?
age	text			Checkbox ▾	✗
birth_of	date	1986-05-06	1986-05-06	Date range ▾	✗
charencoding	text			Checkbox ▾	✗
comment	text			Checkbox ▾	✗
corpus	text			Checkbox ▾	✗
date	date	1988-02-10	1991-11-18	Date range ▾	✓
language	text			Checkbox ▾	✗
location	text			Checkbox ▾	✗
months	int	21	66	Slider ▾	✓
name	text			Checkbox ▾	✗
origutt	text			Checkbox ▾	✗

Example-based search

1 - Example 2 - Parse 3 - Matrix 4 - Treebanks 5 - Query 6 - Results

Step 6: Results

Quick navigation: Individual results Download results Query overview

Individual results

Click on a sentence ID to view the tree structure. The part of the sentence matching your input structure is set in bold.

of results: 500 / 2,886

Filter metadata Filter components

#	ID	Component	Sentence
1	year5-58385.xml	YEARS	jeeminee , kan je ook al le tt ertjes lezen ?
2	year5-58387.xml	YEARS	ik kan al lezen , hoor .
3	year5-58411.xml	YEARS	zeven letters moeten we hebben .
4	year5-58420.xml	YEARS	oh ja , dan mag ik die hebben .
5	year5-58451.xml	YEARS	je moet hier beginnen .
6	year5-58457.xml	YEARS	je moet bij het sterretje beginnen .
7	year5-58459.xml	YEARS	ja , of mama moet de eerste aanleggen en dan
8	year5-58477.xml	YEARS	hier staat geen le tt ertje , maar dan mag je zelf w als er geen lettertje staat .
9	year5-58501.xml	YEARS	oh , nou mogen we letters pakken .
10	year5-58507.xml	YEARS	ik mag er vijf pakken .
11	year5-58553.xml	YEARS	ja , ik moet even kijken , wat voor woord ik kan anleggen aanleggen .
12	year5-58598.xml	YEARS	dan kan ik geen echt woord aanleggen .
13	year5-58610.xml	YEARS	ik mag niet meehelpen .
14	year5-58617.xml	YEARS	ik mag weer vier pakken .
15	year5-58620.xml	YEARS	ik mag er weer vier pakken .

date

1988-02-10 1991-11-18

months

21 - 66

speaker

JAC (2147)

NIE (8)

LAU (728)

PET (1)

FRI (2)

Integration with GrETEL3

- Faceted search on metadata in uploaded corpora
- View metadata per hit

ik kan al lezen , hoor .

age	5:0:20
charencoding	UTF8
corpus	vankampen
date	1991-05-26
language	nld
months	60
name	Laura
origutt	ik kan al lezen , hoor .
role	Child
role	Child
situation	JAC and LAU are going to play scrabble
speaker	LAU
transcriber	Simone Boezewinkel
uttid	3

VU

- Shared Vocabularies

- VU developed scripts to convert diachronic lexicons to RDF
- Collaboration with WP2 and WP4
- Presented at DH Benelux 2016:
 - Isa Maks, Marieke van Erp, Piek Vossen, Rinke Hoekstra, Nicoline van der Sijs (2016) [Integrating Diachronous Conceptual Lexicons through Linked Open Data](#).
- Code and converted lexicons available at: <https://github.com/cltl/clariah-vocab-conversion>

- Formats

- Converters between NAF and FoLiA: NAFFoLiA
- <https://github.com/cltl/NAFFoLiAPy>
- Collaboration with Radboud University Nijmegen

VU

- Fine-grained entity typing for Dutch
 - Instead of 4 entity types, 59 or 269 such as person/politician and organisation/company/news
 - Paper currently under review, F-measure of .90 on 59 types and .57 on 269 types
 - <https://github.com/ctl/multilingual-finegrained-entity-typing>
- Occupations tagger
 - Can identify mentions in text and link them to concepts in a resource
 - Currently set up to detect mentions of occupations from HISCO
 - <https://github.com/ctl/SimpleTagger>
 - Collaboration with WP4