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”

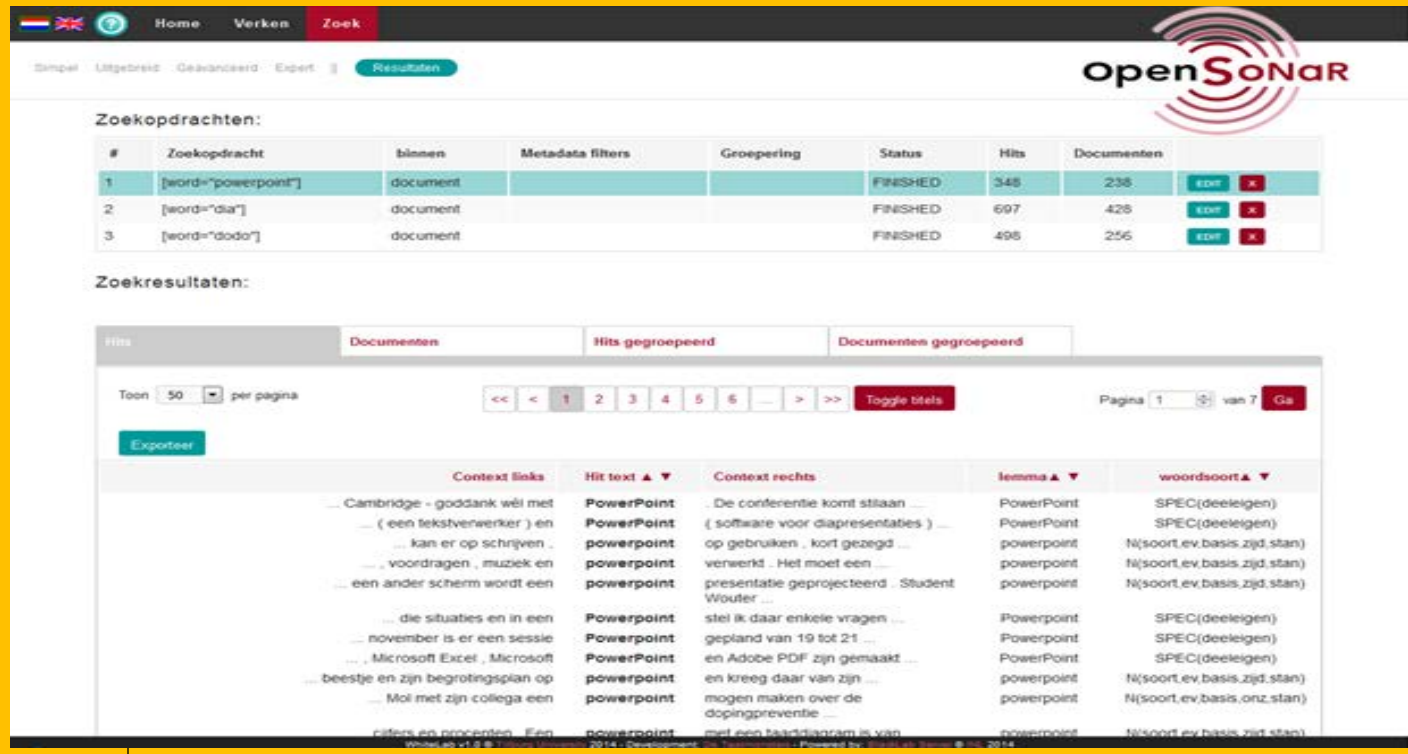
More grouping-/sorting options

- “2nd word to the left of the hit”

Random sampling

- show 10% of the hits found

More to come!

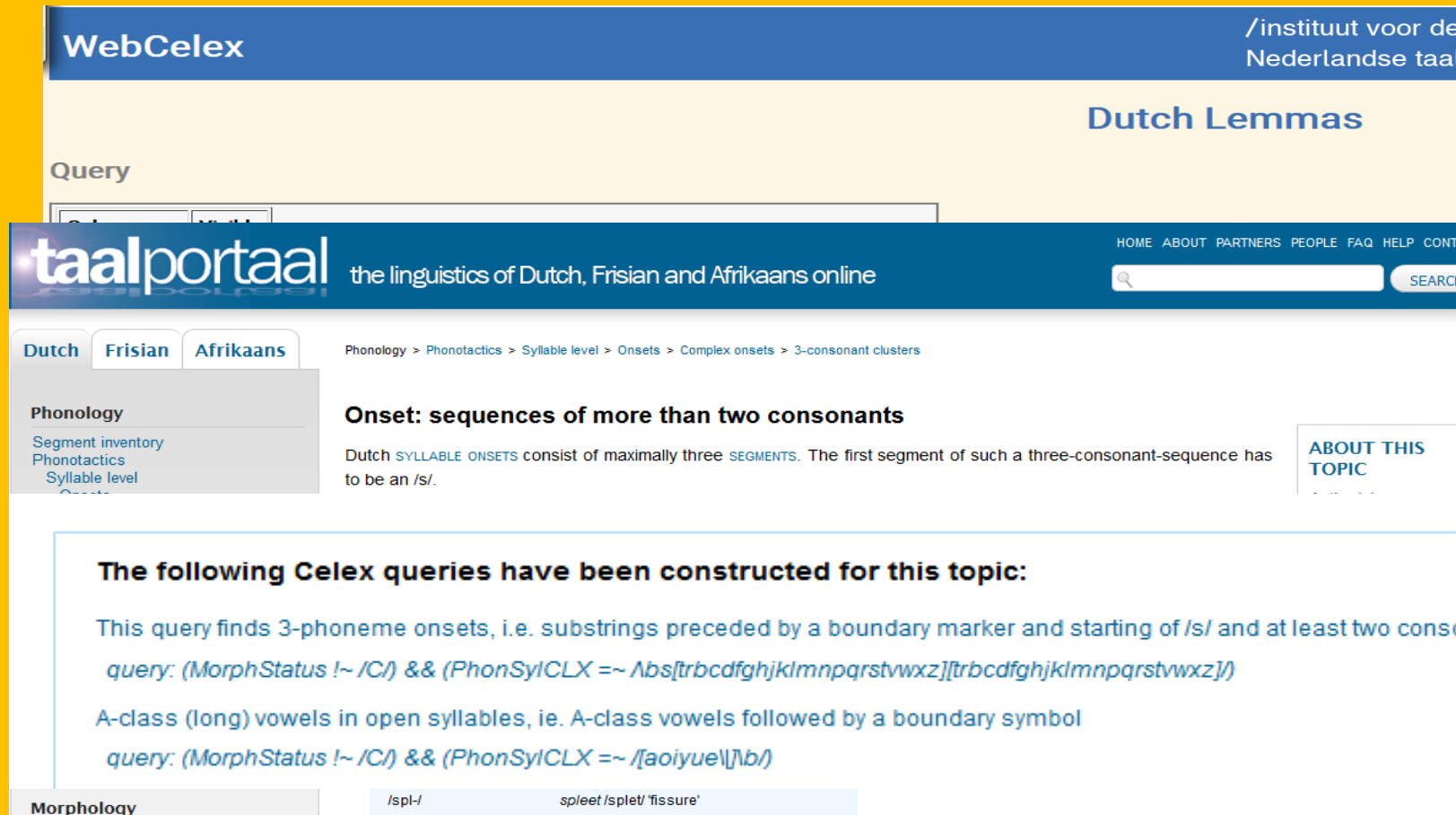


/instituut voor de Nederlandse taal/

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



Meertens: MTAS

- MTAS = Multi Tier Annotation Search : Search engine for annotated text corpora
- **http://www.nederlab.nl/onderzoeksportaal/**
- 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

2017-03

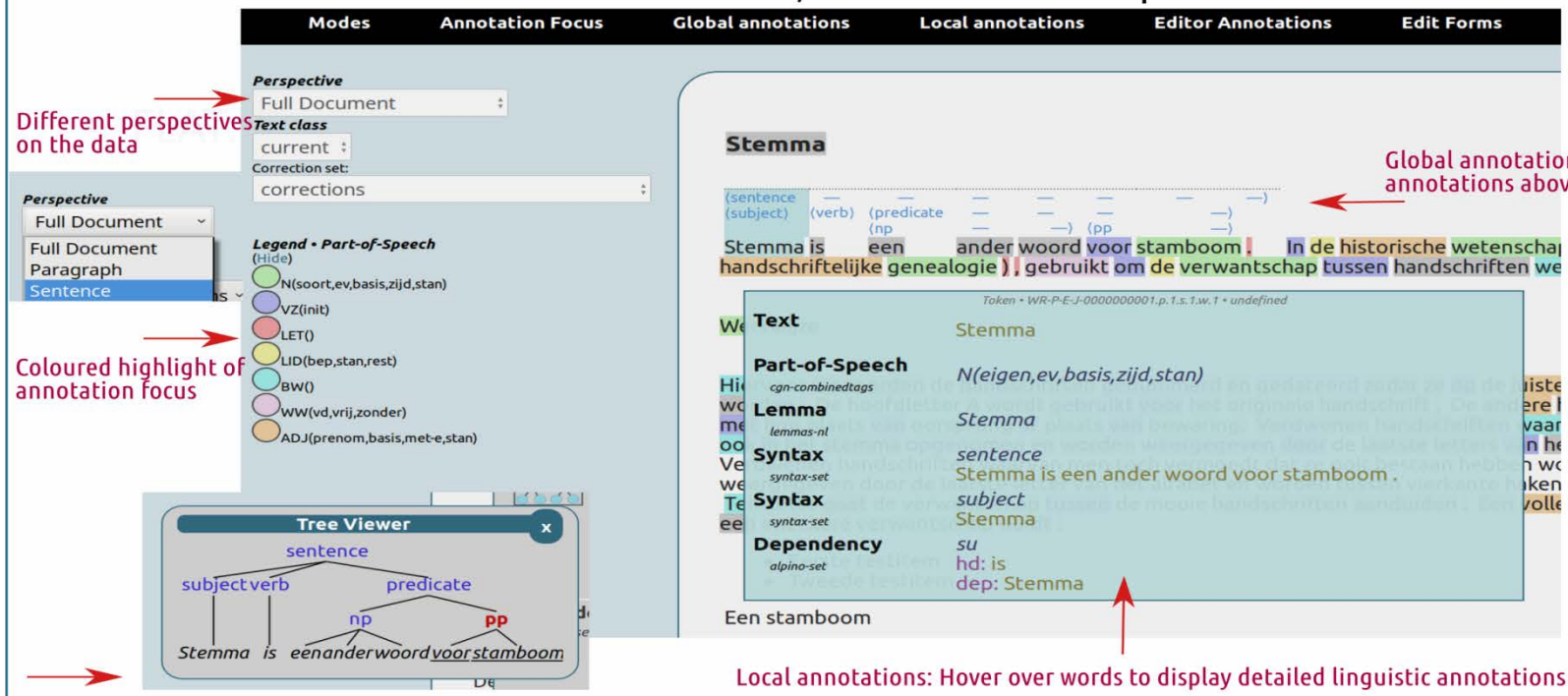
by Maarten van Gompel, funded by CLARIAH

Centre for Language and Speech Technology
Radboud University Nijmegen

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



Different perspectives on the data

Coloured highlight of annotation focus

Global annotations: view annotations above the text

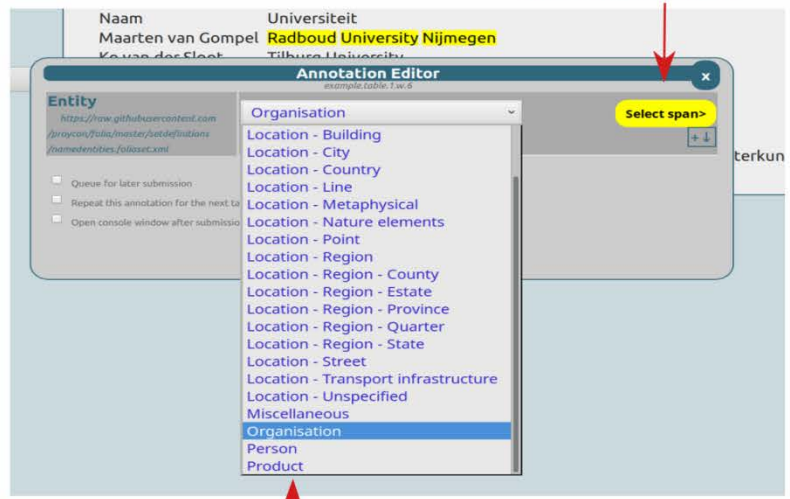
Local annotations: Hover over words to display detailed linguistic annotations

Visualisation of syntax and morphology trees

Click words to open the annotation editor

Add new annotations

Set annotations to span over multiple words



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

- 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

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:		
aantal	waarde	*aantal*	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

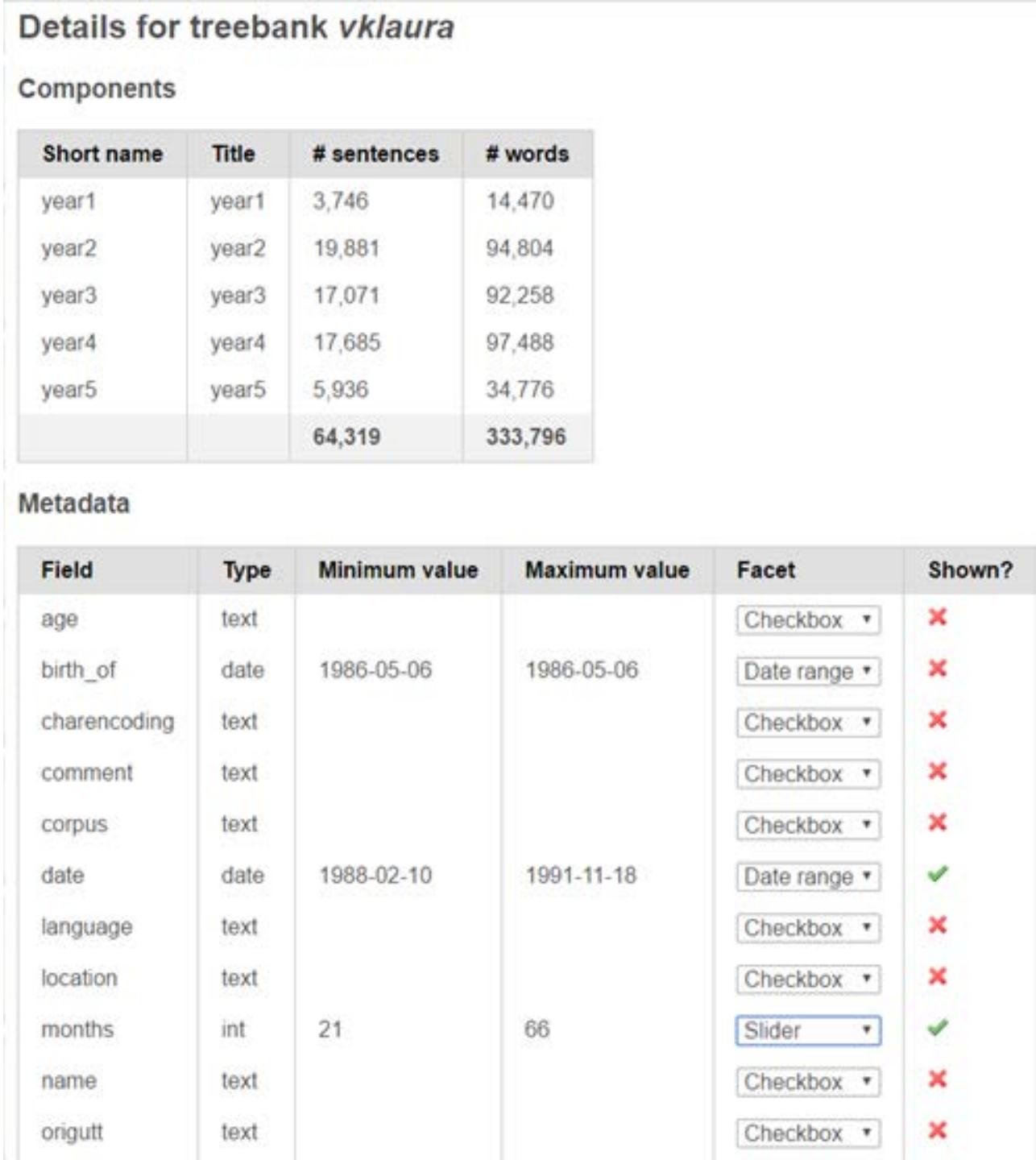
per item:		per zin:		
aantal	waarde	*aantal*	per 100 000	waarde
1 054	NL	995	1 148	NL
529	BE	479	1 134	BE
3	(leeg)	3	290	(leeg)

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 Convertor for CHAT data and metadata (to be integrated)
- Demo’s, posters at Language Science Day (Utrecht), CLIN (Leuven), submitted to DH Benelux



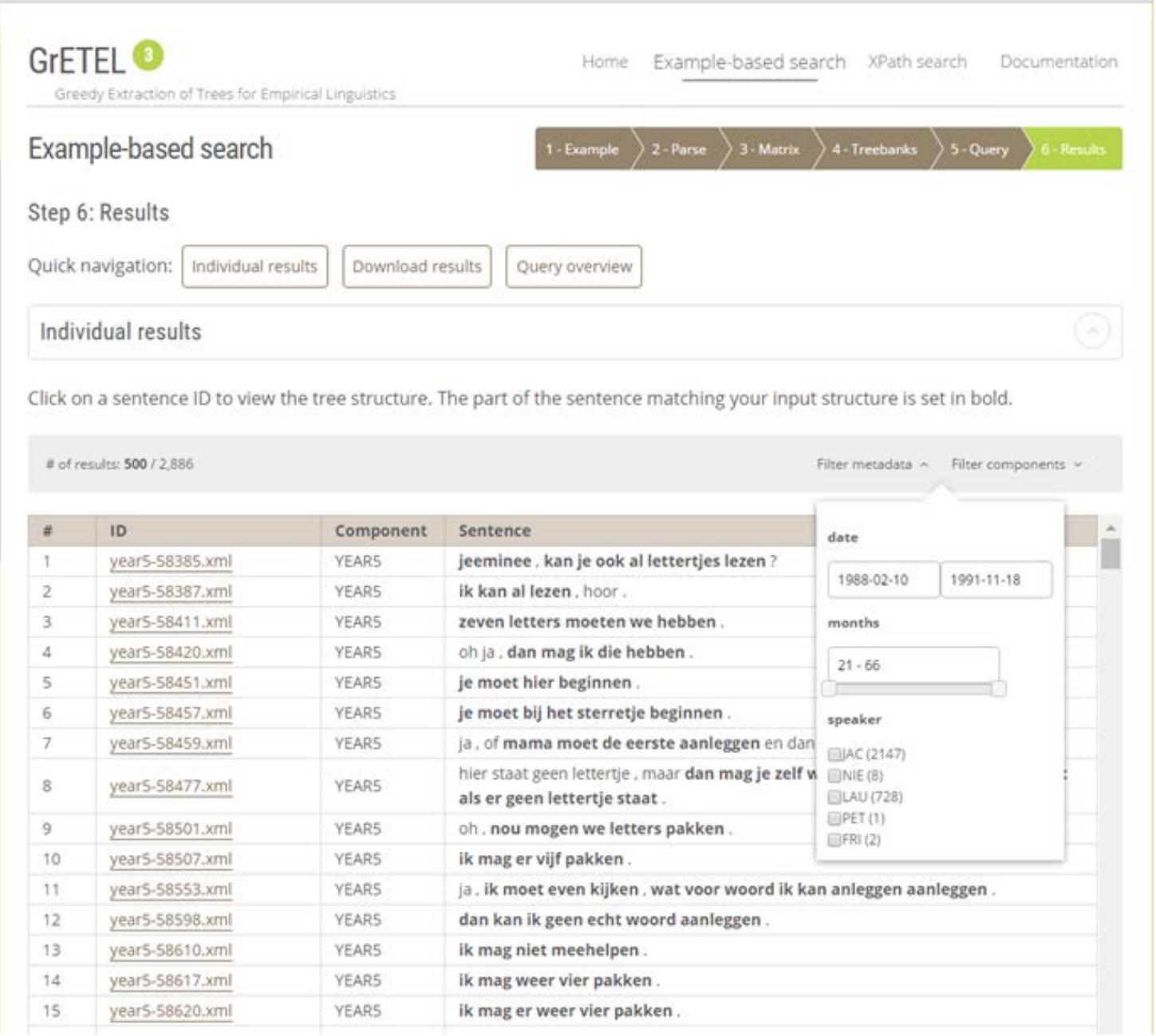
Details for treebank *vklaura*

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**

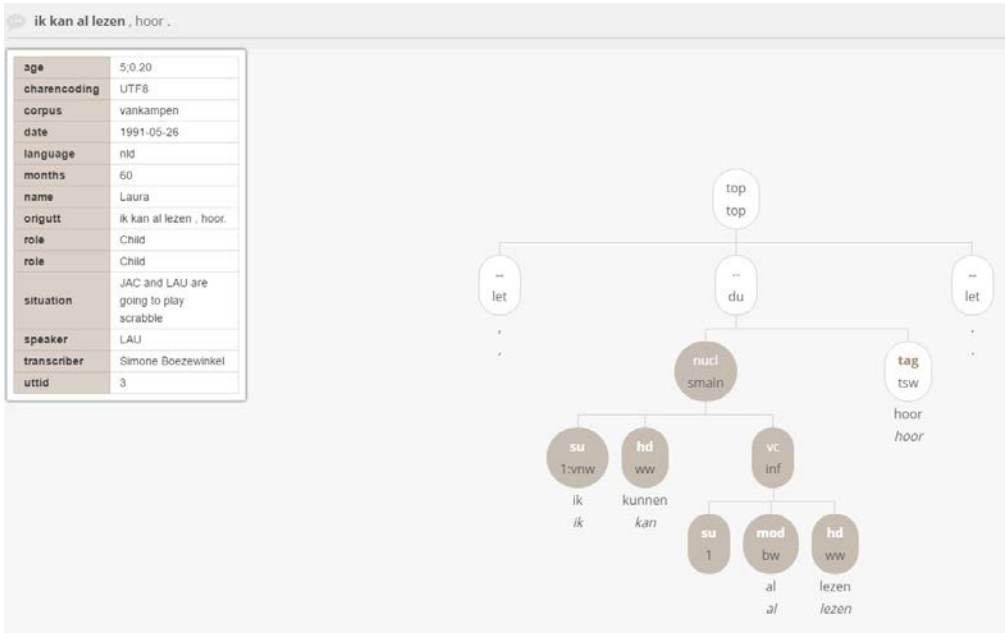
Metadata

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	✗



Integration with GrETEL3

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



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/cltl/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/cltl/SimpleTagger
  - Collaboration with WP4