KU LEUVEN

Zoeken in een Afrikaans corpus: baie maklik!

Liesbeth Augustinus
Ineke Schuurman
Vincent Vandeghinste
Peter Dirix
Frank Van Eynde

Colloquium Afrikaans - 23 oktober 2015

AFRIBOOMS PROJECT

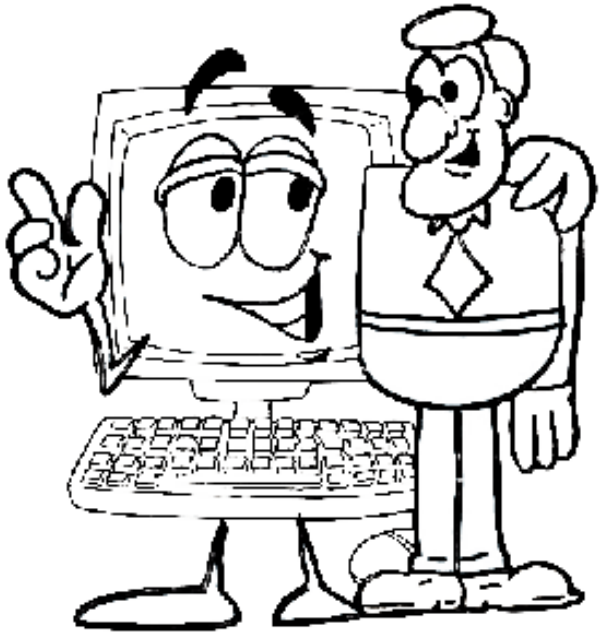
- Syntactisch geannoteerd corpus (treebank) voor Afrikaans
- Parser voor Afrikaans
- Doorzoekbaar maken van de treebank (voor taalkundig onderzoek)
- 2013 - 2014
- NWU (CTexT) & KU Leuven (CCL)

KU LEUVEN

GrETEL 4 AFRIKAANS

- **Gr**eedy **E**xtraction of **T**rees for **E**mpirical **L**inguistics

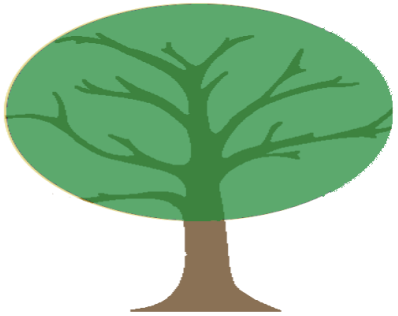
- **Zoekmachine** voor treebanks



KU LEUVEN

GrETEL

- **Gr**eedy **E**xtraction of **T**rees for **E**mpirical **L**inguistics
- Zoekmachine voor treebanks
- **Treebank** = syntactisch geannoteerd corpus
  - Penn Treebank (Engels)
  - TüBa (Duits)
  - LASSY, CGN, SoNaR (Nederlands)



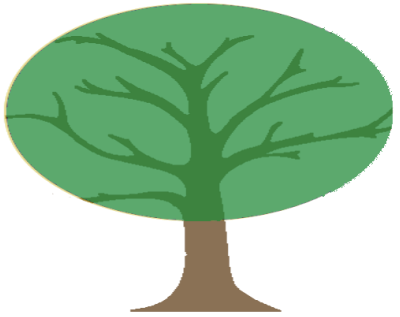
AFRIKAANSE TREEBANK

NCHLT treebank
Geschreven Afrikaans
Afrikaans deel van NCHLT Annotated Tekst Corpora (National Centre for Human Language Technologies)
44 715 woorden
1 934 zinnen
Manueel gecorrigeerd

KU LEUVEN

GrETEL

- **Gr**eedy **E**xtraction of **T**rees for **E**mpirical **L**inguistics
- Zoekmachine voor treebanks
- **Treebank** = syntactisch geannoteerd corpus
  - Penn Treebank (Engels)
  - TüBa (Duits)
  - LASSY, CGN, SoNaR (Nederlands)
- **Parser**
  - Bv. Alpino (Van Noord 2006)



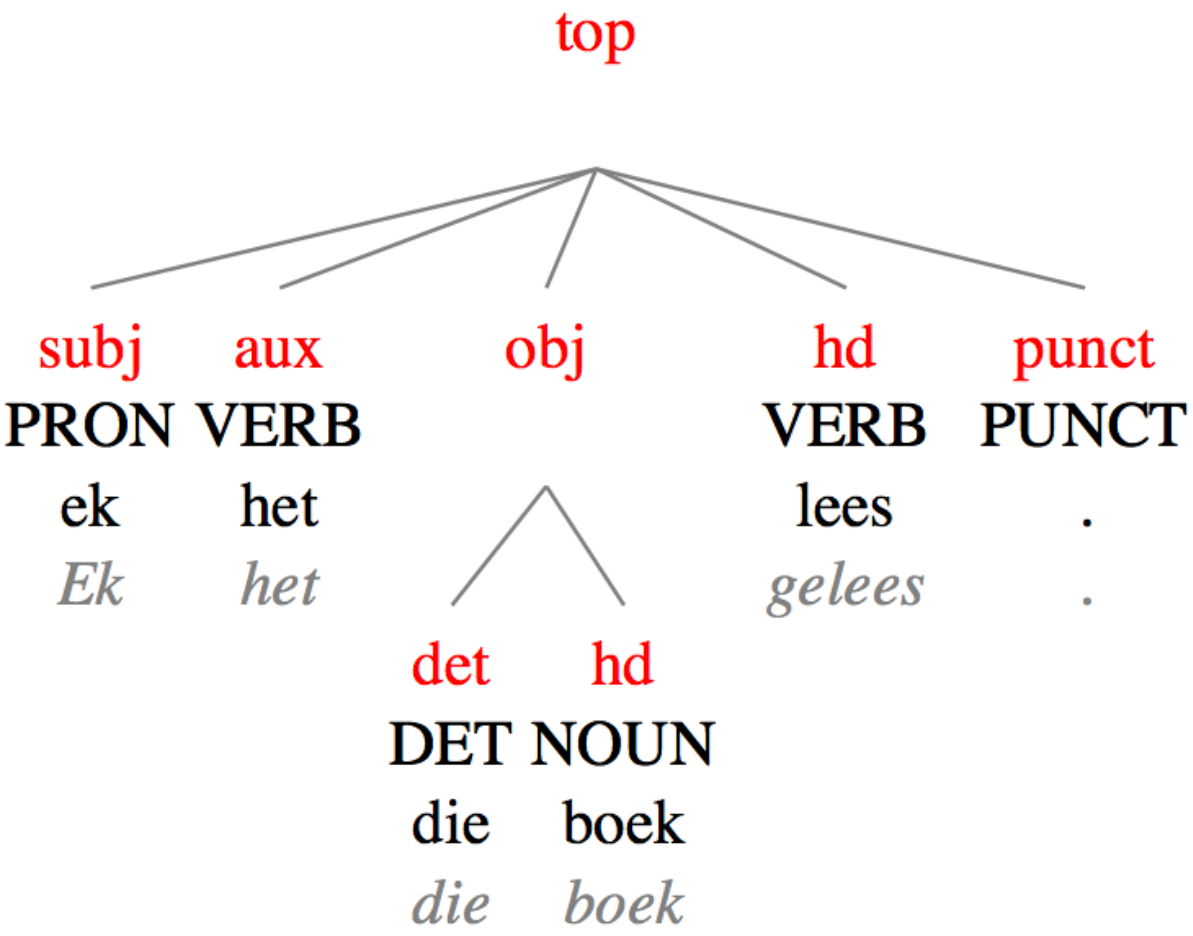
PARSER

Ek het die boek gelees. >> parser >>

KU LEUVEN

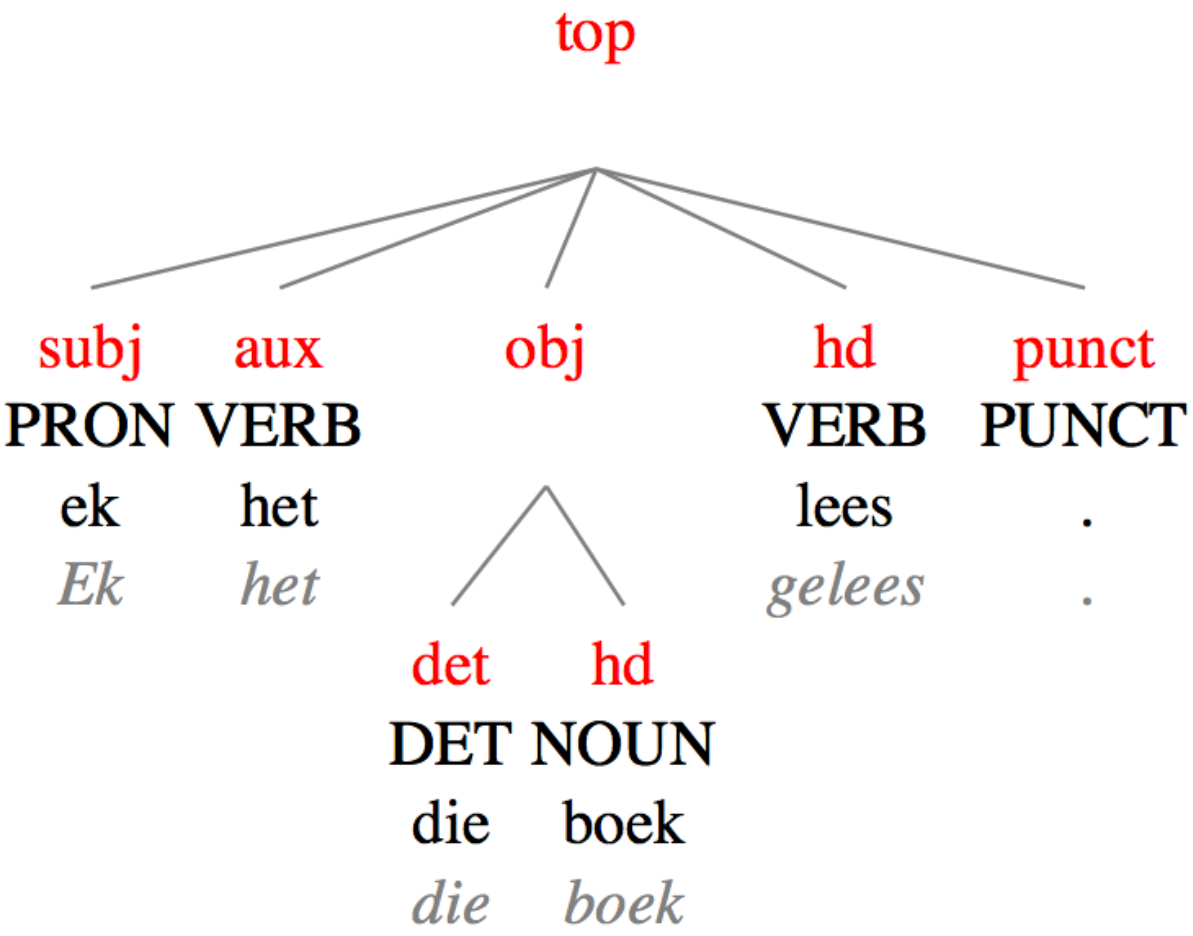
PARSER

>> parser >>



PARSER

>> parser >>

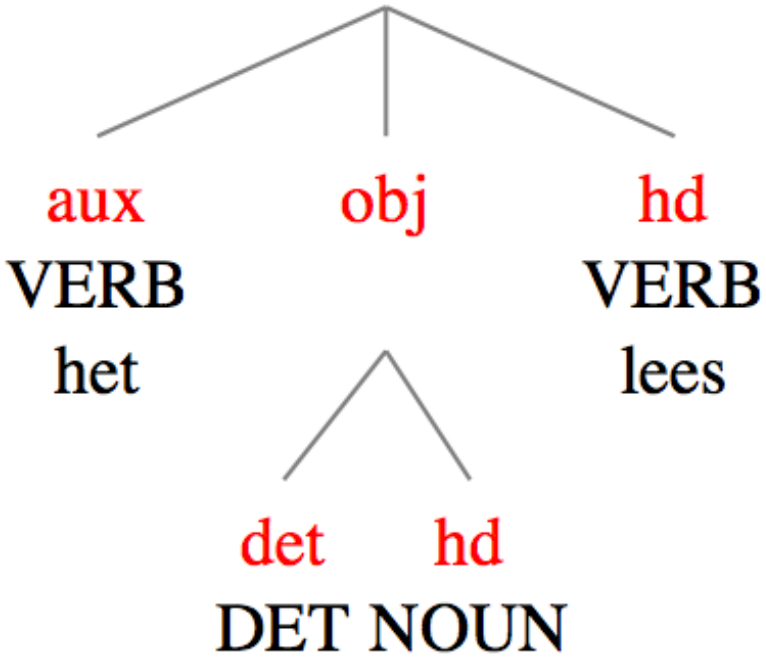


XML-boomstructuur

Zoektaal: **XPath**

KU LEUVEN

XPATH




```
//node[node[@rel="aux" and
@pt="VERB" and @lemma="het"]
and
node[@rel="obj" and
node[@rel="det" and
@pt="DET"] and
node[@rel="hd" and
@pt="NOUN"]]
and
node[@rel="hd" and
@pt="VERB" and
@lemma="lees"]]
```

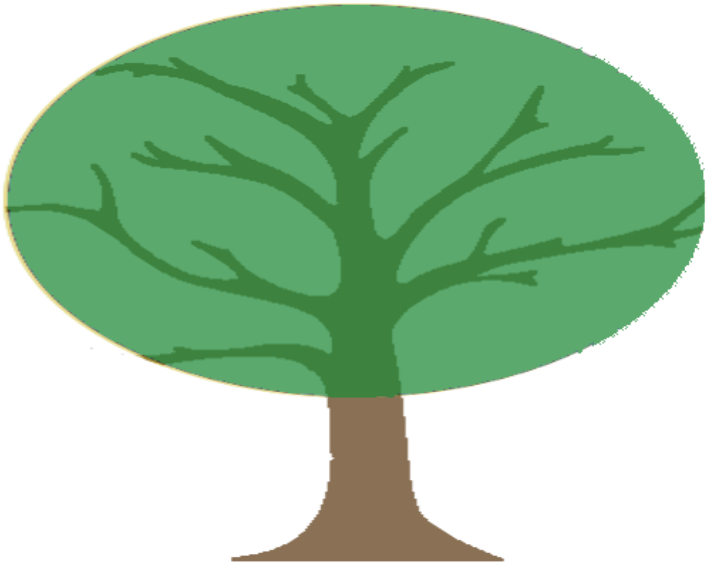
KU LEUVEN

XPATH

```
//node[node[@rel="aux" and
@pt="VERB" and @lemma="het"]
and
node[@rel="obj" and
node[@rel="det" and
@pt="DET"] and
node[@rel="hd" and
@pt="NOUN"]]
and
node[@rel="hd" and
@pt="VERB" and
@lemma="lees"]]
```

KU LEUVEN

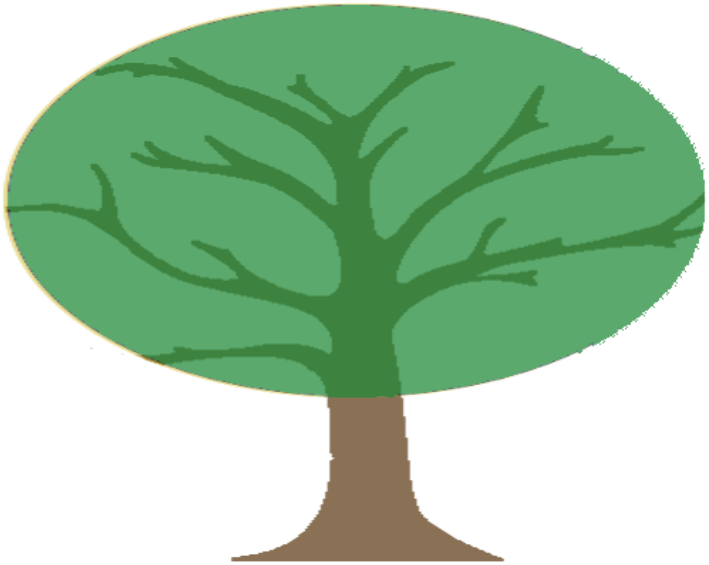
XPATH



```
//node[node[@rel="aux" and
@pt="VERB" and @lemma="het"]
and
node
node[
@pt="DE
node[
@p
and
node[@rel="hd" and
@pt="VERB" and
@lemma="lees"]]
```

KU LEUVEN

XPATH



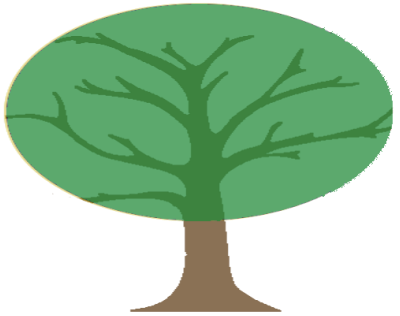
KU LEUVEN

GrETEL

2 zoekmodi:

- Voorbeeld-gebaseerd zoeken

- Zoeken met XPath



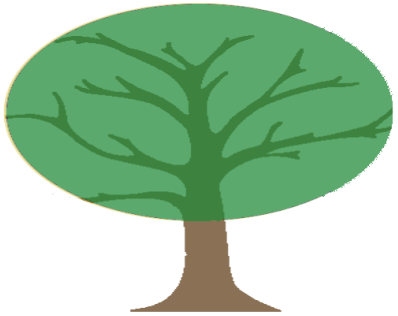
GrETEL

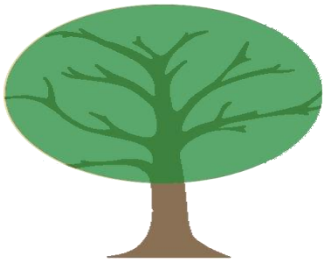
2 zoekmodi:

- Voorbeeld-gebaseerd zoeken

  voordeel: geen of beperkte kennis van de data en/of formele zoektalen (zoals XPath) vereist

- Zoeken met XPath





de gebruiker

1. Geef een voorbeeld
2. Bekijk parse
3. Geef de relevante items aan in de zin
4. Selecteer treebank
5. (Bewerk XPath)
6. Bekijk de resultaten

GrETEL

- Parst de zin (met Alpino)
- Genereert automatisch een XPath instructie
- Presenteert resultaten

KU LEUVEN

OVERZICHT

- GrETEL in een notendop
- **GrETEL demo**
  - **Case study**
  - Zoekopties
- Conclusies

KU LEUVEN

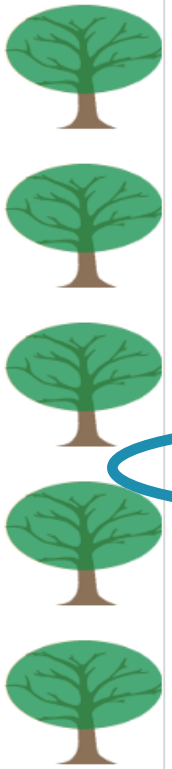
CASE STUDY

Constructies met *het* + object met een bepaald lidwoord (*die*) + werkwoord

Bv. *Ek het die boek gelees.*

KU LEUVEN

GrETEL ONLINE



GrETEL 4 Afrikaans

This website hosts GrETEL for Afrikaans, a search engine for the Afrikaans NCHLT treebank. The construction of the treebank and the adaptation of GrETEL were part of the AfriBooms project.

What is GrETEL?

GrETEL stands for **Gr**eedy **E**xtraction of **T**rees for **E**mpirical **L**inguistics. It is a user-friendly search engine for the exploitation of treebanks. It comes in two formats:

Example-based search

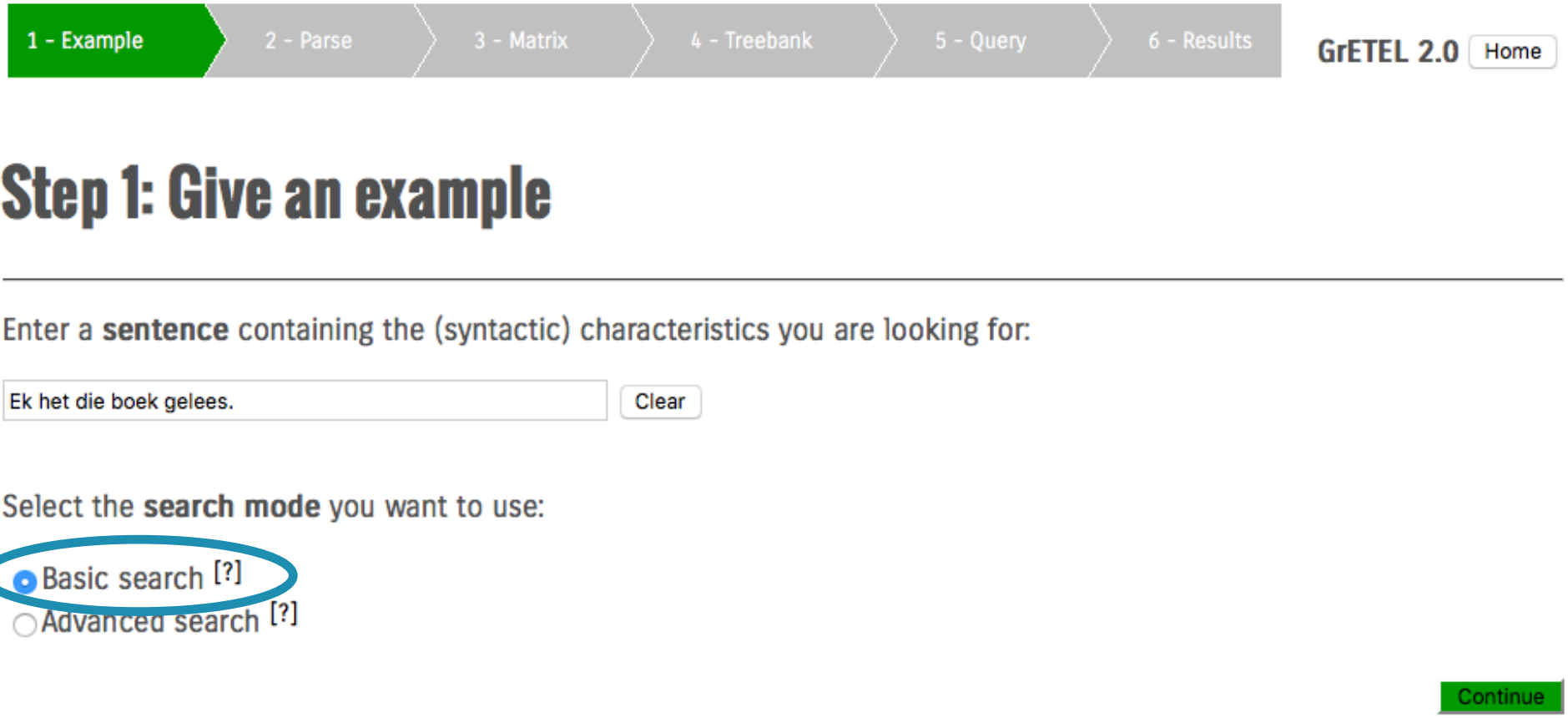
In this search mode you can use a natural language example as a starting point for searching a treebank [?] with limited knowledge about tree representations and formal query languages. [?]

XPath search

In this search mode you have to build the XPath query yourself. We strongly recommend to use the XPath search tool only when you are an experienced XPath user!

KU LEUVEN

INPUT



KU LEUVEN

INPUT PARSE

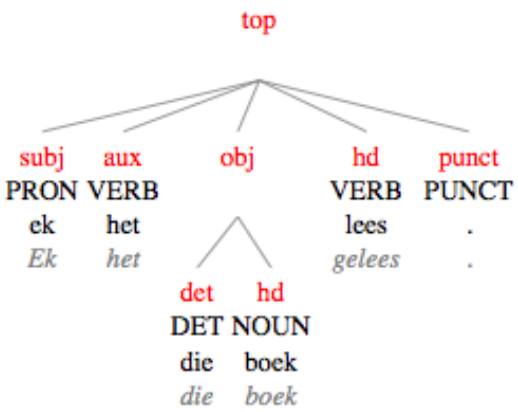
1 - Example | 2 - Parse | 3 - Matrix | 4 - Treebank | 5 - Query | 6 - Results

GrETEL 2.0 - basic search mode Home

Step 2: Input Parse

The structure of the **tagged** [?] and **parsed** [?] sentence: *Ek het die boek gelees .*

Parsed input example [full screen]



SELECTIE MATRIX

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

Step 3: Select relevant parts

Indicate the relevant[?] parts of the sentence, i.e. the parts you are interested in. [view input parse]

sentence	Ek	het	die	boek	gelees	.
word	○	●	○	○	○	○
lemma	○	○	●	○	○	○
word class	○	○	○	●	●	○
optional in search	●	○	○	○	○	●

KU LEUVEN

LEIDRAAD VOOR SELECTIE

GUIDELINES

- **word**: The exact word form. This is a case sensitive feature.
- **lemma**: Word form that generalizes over inflected forms. For example: *sin* is the lemma of *sin*, *sinne*, and *sinnetjie*; *gaan* is the lemma of *gaan* and *gegaan*. Lemma is case insensitive (except for proper names).
- **word class**: Part-of-speech tag. The different tags are: NOUN (noun), VERB (verb), ADJ (adjective), DET (article), PRON (pronoun), CONJ (conjunction), ADV (adverb), NUM (numeral), ADP (preposition), PRT (particle), X (other categories), and PUNCT (punctuation).
- **optional in search**: The word will be ignored in the search instruction. It may be included in the results, but it is not necessary.

KU LEUVEN

TREEBANK SELECTIE

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

GrETEL 2.0 - basic search mode Home

Step 4: Select a treebank

It is currently only possible to search the NCHLT treebank for Afrikaans. Which **treebank** component(s) do you want to query?

	Treebank	Contents	Sentences	Words *
☑	Part 1	First part of the Afrikaans treebank	1,000	21,086
☑	Part 2	Second part of the Afrikaans treebank	934	23,629
	NCHLT treebank	Complete treebank	1,934	44,715

* Counted by the query `//node[@pt and not(@pt="PUNCT")]`

KU LEUVEN

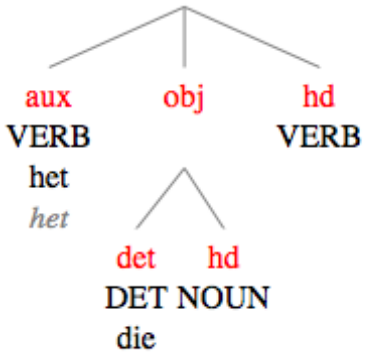
QUERY OVERZICHT

Step 5: Query Overview

Input example: *Ek het die boek gelees .*

Treebank: NCHLT
Components: 1, 2

Query tree [full screen]



KU LEUVEN

RESULTATEN

Constructies met *het* + object met een bepaald lidwoord (*die*) + werkwoord

Bv. *Ek het die boek gelees.*

→ **79 hits, verdeeld over 73 zinnen**

KU LEUVEN

RESULTATEN

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

GrETEL 2.0 - basic search mode Home

Step 6: Results

Printer-friendly version

QUERY

Input example Ek het die boek gelees.

XPath //node[node[@rel="aux" and @pt="VERB" and @word="het" and @lemma="het"] and node[@rel="obj" and node[@rel="det" and @pt="DET" and @lemma="die"] and node[@rel="hd" and @pt="NOUN"]] and node[@rel="hd" and @pt="VERB"]] Download XPath [?]

Treebank NCHLT [1, 2]

RESULTS

Hits 79 Show hits distribution

Matching sentences 73 Download [?]

Sentences in treebank 1,934

KU LEUVEN

RESULTATEN: data

Click on a sentence ID to view the tree structure. The sentence ID refers to the treebank component in which the sentence occurs, the text number, and the location within the text (page + sentence number).

Search within results:

SENTENCE ID	MATCHING SENTENCES	HITS
NCHLT.p.1125.s.1125	**Verlede week het die agbare lede die geleentheid gehad om oor hierdie aangeleenthede te besin .**	1
NCHLT.p.1126.s.1126	**Uiteraard het die aansienlike styging in die behoefte aan elektrisiteit gedurende die afgelope twee jaar die nuwe kapasiteit wat ons beskikbaar gestel het oorskry .**	1
NCHLT.p.1157.s.1157	**Hierdie situasie het die onvermydelike besef dat die tydperk van goedkoop en voldoende elektrisiteit op 'n einde is , verhaas .**	1
NCHLT.p.1173.s.1173	Ek wil weereens die Springbokke bedank wat **die aanvoorwerk gedoen het toe hulle die Rugbywêreldbeker verlede jaar gewen het .**	2
NCHLT.p.1219.s.1219	Die uitbreiding van die openbare werke-program , wat , **deur die vasgestelde doelwitte te oortref , die potensiaal getoon het om meer nuwelinge te absorbeer .**	1

KU LEUVEN

RESULTATEN: data

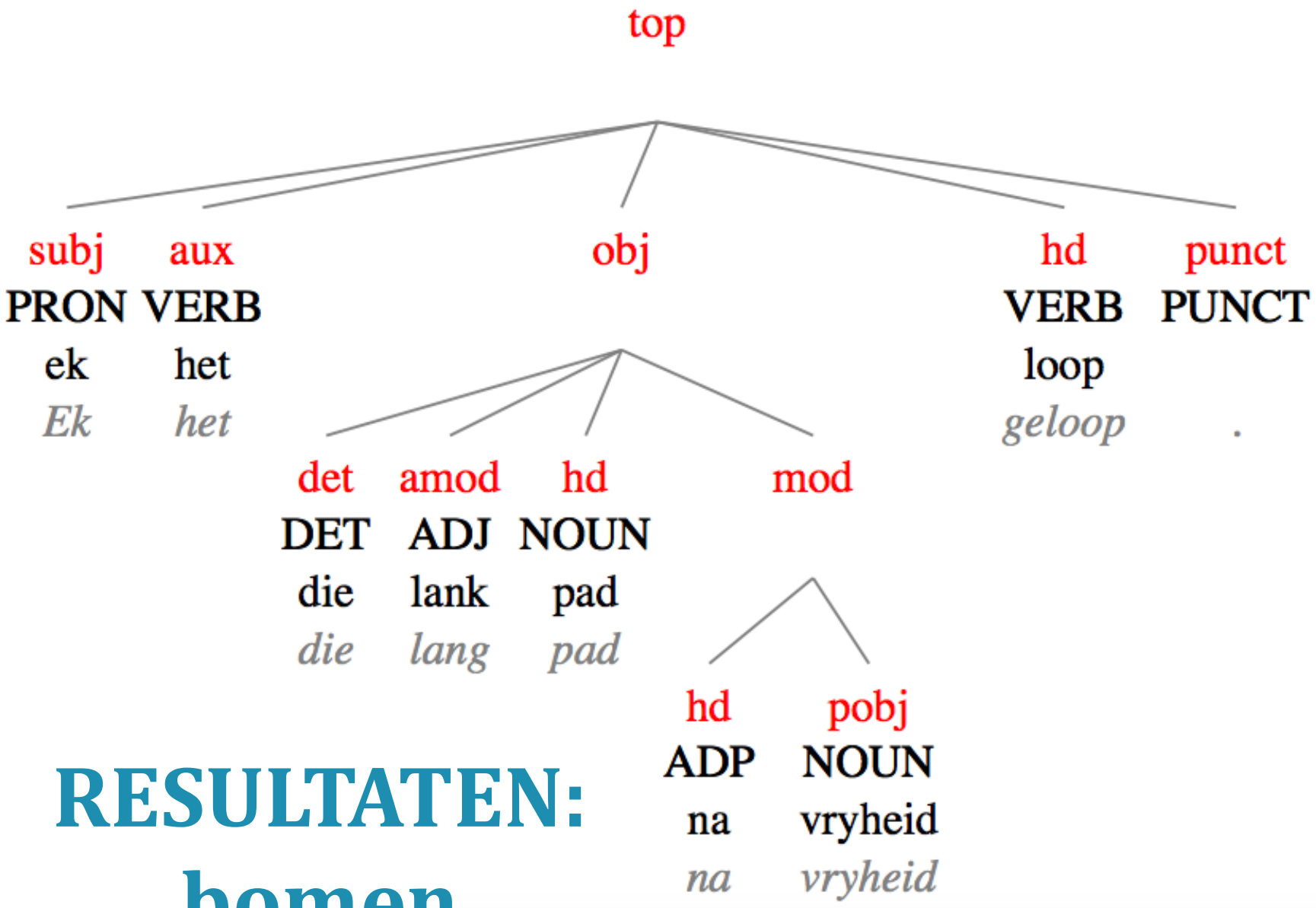
Click on a sentence ID to view the tree structure. The sentence ID refers to the treebank component in which the sentence occurs, the text number, and the location within the text (page + sentence number).

Search within results:

SENTENCE ID	MATCHING SENTENCES	HITS
NCHLT.p.1125.s.1125	**Verlede week het die agbare lede die geleentheid gehad om oor hierdie aangeleenthede te besin .**	1
NCHLT.p.1126.s.1126	**Uiteraard het die aansienlike styging in die behoefte aan elektrisiteit gedurende die afgelope twee jaar die nuwe kapasiteit wat ons beskikbaar gestel het oorskry .**	1
NCHLT.p.1157.s.1157	**Hierdie situasie het die onvermydelike besef dat die tydperk van goedkoop en voldoende elektrisiteit op 'n einde is , verhaas .**	1
NCHLT.p.1173.s.1173	Ek wil weereens die Springbokke bedank wat **die aanvoorwerk gedoen het toe hulle die Rugbywêreldbeker verlede jaar gewen het .**	2
NCHLT.p.1219.s.1219	Die uitbreiding van die openbare werke-program , wat , **deur die vasgestelde doelwitte te oortref , die potensiaal getoon het om meer nuwelinge te absorbeer .**	1

“greedy” search

KU LEUVEN



RESULTATEN: bomen

MEER RESULTATEN

Optie 1: Gebruik verschillende query's

Constructies met *het* + object met een bepaald lidwoord (*die*) + werkwoord

Bv. *Ek het die boek gelees.*

→ **79 hits**

Constructies met *het* + object met een onbepaald lidwoord (*'n*) + werkwoord

Bv. *Ek het 'n boek gelees.*

→ **42 hits**

TOTAAL: 121 hits

MEER RESULTATEN

Optie 2: Pas de XPath query aan (via “XPath Search”)

```
//node[node[@rel="aux" and @pt="VERB" and @word="het" and
@lemma="het"] and node[@rel="obj" and node[@rel="det" and
@pt="DET" and (@lemma="die" or @lemma="'n")] and
node[@rel="hd" and @pt="NOUN"]] and node[@rel="hd" and
@pt="VERB"]]
```

KU LEUVEN

MEER RESULTATEN

GrETEL 4 Afrikaans

This website hosts GrETEL for Afrikaans, a search engine for the Afrikaans NCHLT treebank. The construction of the treebank and the adaptation of GrETEL were part of the AfriBooms project.

What is GrETEL?

GrETEL stands for **Gr**eedy **E**xtraction of **T**rees for **E**mpirical **L**inguistics. It is a user-friendly search engine for the exploitation of treebanks. It comes in two formats:

Example-based search

In this search mode you can use a natural language example as a starting point for searching a treebank [?] with limited knowledge about tree representations and formal query languages. [?]

XPath search

In this search mode you have to build the XPath query yourself. We strongly recommend to use the XPath search tool only when you are an experienced XPath user!

KU LEUVEN

MEER RESULTATEN

Optie 2: Pas de XPath query aan (via “XPath Search”)

OVERZICHT

- GrETEL in een notendop
- **GrETEL demo**
  - Case study
  - **Zoekopties**
- Conclusies

KU LEUVEN

GEAVANCEERD ZOEKEN

Step 1: Give an example

Enter a **sentence** containing the (syntactic) characteristics you are looking for:

Ek het die boek gelees. Clear

Select the **search mode** you want to use:

Basic search [?]

Advanced search [?]

KU LEUVEN

GEAVANCEERD ZOEKEN

Step 3: Select relevant parts

Indicate the relevant[?] parts of the sentence, i.e. the parts you are interested in.
[view input parse]

sentence	Ek	het	die	boek	gelees	.
word	○	●	○	○	○	○
lemma	○	○	○	○	○	○
word class	○	○	○	●	●	○
optional in search	●	○	○	○	○	●
NOT in search	○	○	●	○	○	○

KU LEUVEN

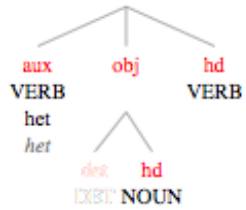
GEAVANCEERD ZOEKEN

Step 5: Query Overview

Input example: *Ek het die boek gelees .*

Treebank: NCHLT
Components: 1, 2

Query tree [full screen]



XPath expression generated from the query tree. You can adapt it if necessary. If you are dealing with a long query the XPath beautifier might come in handy.

```
//node[node[@rel="aux" and @pt="VERB" and @word="het" and @lemma="het"] and node[@rel="obj" and not(node[@rel="det" and @pt="DET"]) and node[@rel="hd" and @pt="NOUN"]] and node[@rel="hd" and @pt="VERB"]]
```

KU LEUVEN

GEAVANCEERD ZOEKEN

SENTENCE ID	MATCHING SENTENCES	HITS
NCHLT.p.1062.s.1062	**Sedert die nasie ons in 2004 'n mandaat daartoe verleen het , het ons welkome vordering gemaak in ons pogings om Suid-Afrika ten goede te verander .**	1
NCHLT.p.1080.s.1080	Ons verwelkom by hierdie geleentheid Mnr Arthur Margeman , verteenwoordiger van die veterane van die Alexandra-busboikot van 50 jaar gelede , en wat **Nelson Mandela ingesluit het .**	1
NCHLT.p.1082.s.1082	Ons verwelkom Mnr Dinilesizwe Sobukwe , seun van die uitstaande patriot en leier , Robert Sobukwe , wat ook 30 jaar gelede oorlede is nadat **hy baie jare se tronkstraf , verbanning en ander vorme van onderdrukking verduur het .**	1
NCHLT.p.1091.s.1091	Dit sal gebruik word om die doelwitte wat **die mense ons opdrag gegee het** om na te streef , nog vinniger te bereik .	1
NCHLT.p.1099.s.1099	Terwyl **ek hierdie toespraak voorberei het ,** het een van my kollegas aangevoer dat ons land geteister word deur sterk dwarswinde wat dit veral moeilik maak om te voorspel waar ons land more gaan wees .	1
NCHLT.p.1100.s.1100	Hy het voorgestel dat ek die welbekende woorde waarmee **Charles Dickens sy novelle A Tale of Two Cities geopen het ,** aanhaal om die werklikheid wat ons in die gesig staar , vas te vang .	1
NCHLT.p.1112.s.1112	Hulle is besorg dat ons land moontlik bedreig word deur die anargie wat verteenwoordig is deur die kriminele brandstigting wat **verlede maand plaasgevind het op ses passasierstreine in Tshwane .**	1

KU LEUVEN

ZOEKOPTIES

sentence	Hy	sal	die	man	help	met	sy	werk
word	○	○	○	○	○	○	○	○
lemma	○	○	○	○	○	○	○	○
word class	○	○	○	○	◉	◉	○	○
optional in search	◉	◉	◉	◉	○	○	◉	◉

OPTIONS

☑ Respect word order

KU LEUVEN

WOORDVOLGORDE

PP-over-V

- V + PP
- *Hy sal die man help met sy werk.*

- PP + V
- *Hy sal die man met sy werk help.*

WOORDVOLGORDE

PP-over-V

- V + PP
- *Hy sal die man help met sy werk*

NCHLT.p.1015.s.1015	Ons sal sorg dat die vennootskappe wat **oor die jare heen opgebou is ,** verstewig word , en dat ons verbeterde nasionale omvattende strategie teen vigs en seksueel oordraagbare siektes so gou moontlik gefinaliseer word .
NCHLT.p.1016.s.1016	Vanjaar sal ons die konkrete planne vir die implementering van die finale stadiums van ons programme wat **gerig is op die verskaffing van universele toegang tot water in 2008 , sanitasie in 2010 en elektrisiteit in 2012** afhandel .

2 576 hits in 1 530 zinnen

Maar: resultaten bevatten ook PP + V!

WOORDVOLGORDE

PP-over-V

- V + PP **+ woordvolgorde optie**
- *Hy sal die man help met sy werk*

NCHLT.p.1254.s.1254	Dit is verder baie belangrik dat die verbintenis van staatsamptenare tot hul pligte verbeter word - 'n taak wat **rus op die skouers van die leiers , staatsamptenare en die werkersuniebeweging .**
NCHLT.p.1281.s.1281	Dit is verwikkelings wat duidelik die vooruitgang kniehalter wat **ons die afgelope paar jaar gemaak het ten opsigte van die herlewing van die Afrika-kontinent .**

1 090 hits in 847 zinnen

Resultaten bevatten enkel V + PP

OVERZICHT

- GrETEL in een notendop

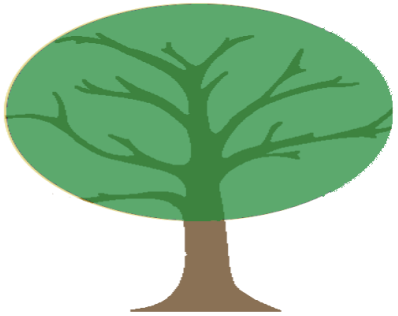
- GrETEL demo
  - Case study
  - Zoekopties

- **Conclusies**

KU LEUVEN

CONCLUSIES

- **GrETEL**: zoekmachine voor treebanks, nu ook voor Afrikaans!
- Input = natuurlijke taal
- Output = sample van gelijkaardige zinnen
- “Syntactic concordancer”
- Online beschikbaar (via *Mozilla Firefox*)
- Geen installatie nodig



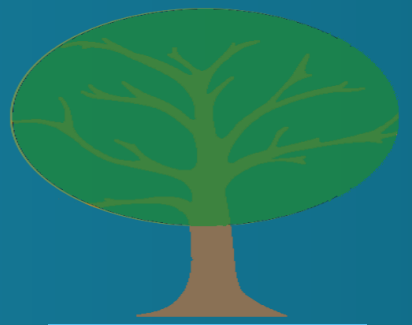
TOEKOMSTMUZIEK

- Meer data toevoegen: Taalkommissiekorpus
- Parser verbeteren

KU LEUVEN

Probeer het zelf uit!

http://gretel.ccl.kuleuven.be/afribooms



KU LEUVEN