KU LEUVEN

NEDERBOOMS

Treebank Mining for Data-based Linguistics

Liesbeth Augustinus
Vincent Vandeghinste
Ineke Schuurman
Frank Van Eynde

Dag van het Onderzoek - March 18, 2014

NEDERBOOMS

- **Exploitation of Dutch treebanks for research in linguistics**

- CLARIN-VL project
- October, 2010 – February, 2012

CLARIN

Common Language Resources and Technology Infrastructure

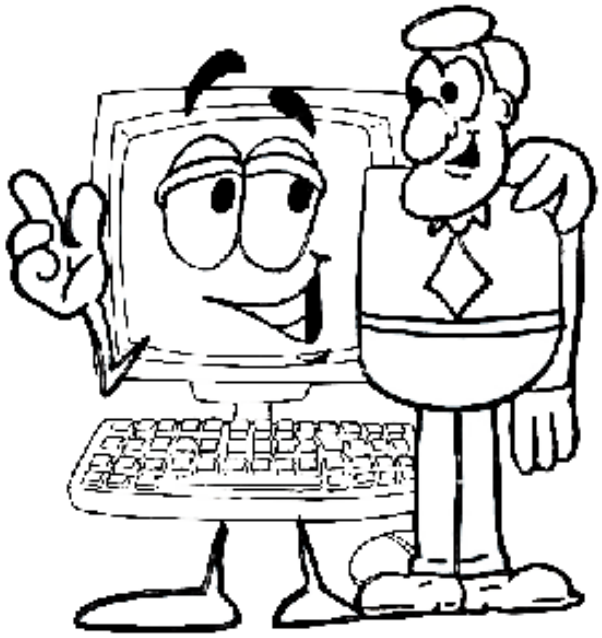
KU LEUVEN

NEDERBOOMS

- **Exploitation of Dutch treebanks for research in linguistics**

- CLARIN-VL project
- October, 2010 – February, 2012

- **Goals:**
  - User-friendly tools
  - Access to large data files
  - Fast and accurate



KU LEUVEN

GrETEL

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

- Query engine for treebanks

- **Treebank** = syntactically annotated corpus
  e.g. Penn Treebank (English), TüBa (German),
  LASSY, CGN (Dutch)

TREEBANKS

CGN treebank	LASSY small
Spoken Dutch	Written Dutch
Stylistic & regional differences conversations vs read texts NL vs VL	Stylistic differences Wikipedia vs legal texts
± 1M tokens	± 1M tokens
130k sentences	65k sentences
Manually corrected	Manually corrected

KU LEUVEN

GrETEL

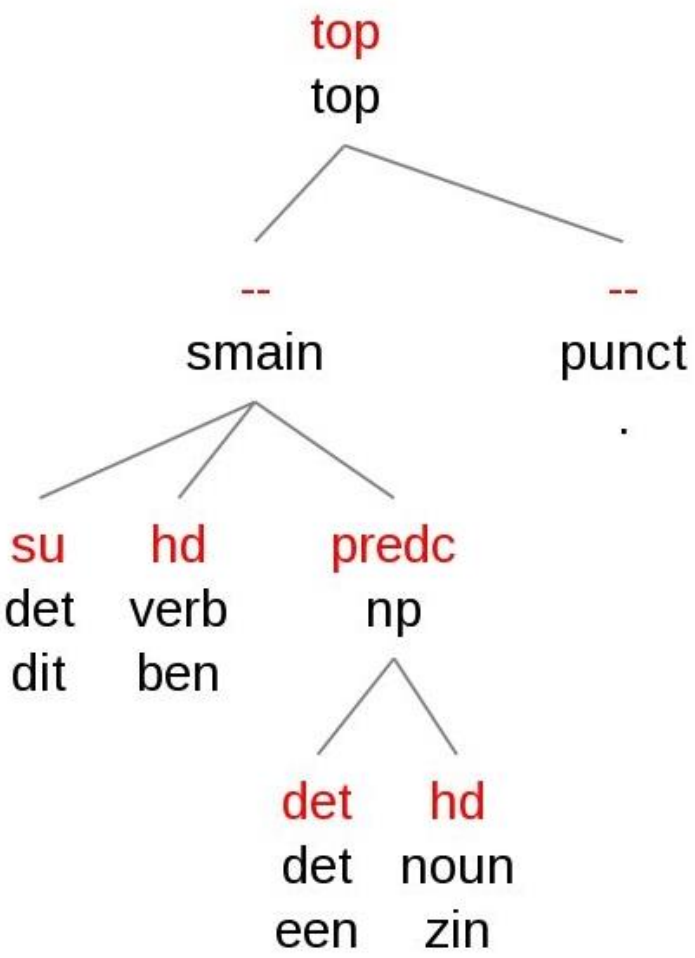
- **Gr**eedy **E**xtraction of **T**rees for **E**mpirical **L**inguistics
- Query engine for treebanks
- **Treebank** = syntactically annotated corpus
  e.g. Penn Treebank (English), TüBa (German), LASSY, CGN (Dutch)
- **Parser**
  e.g. Alpino (Van Noord 2006)

KU LEUVEN

ALPINO PARSER

Dit is een zin. >> ALPINO parser >>

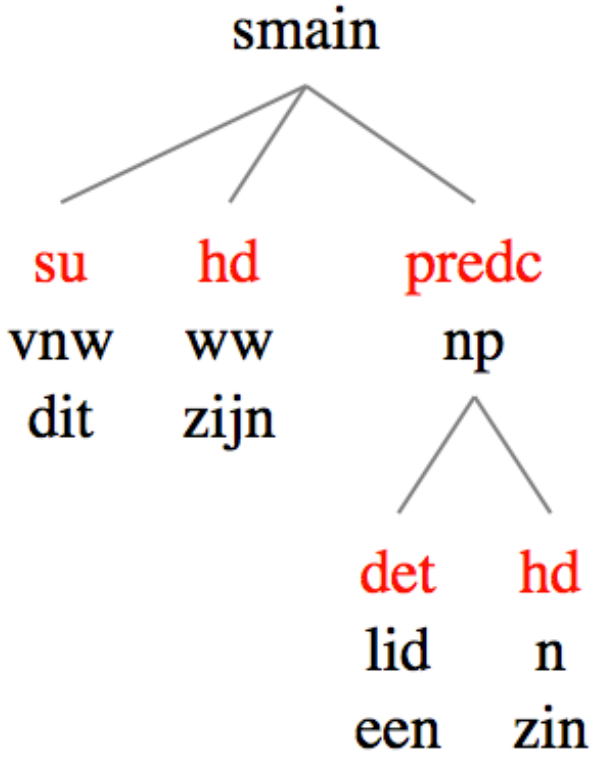
"This is a sentence."



XML trees

Query language: **XPath**

XPATH




```
//node[@cat="smain" and
node[@rel="su" and
@pt="vnw" and @lemma="dit"]
and node[@rel="hd" and
@pt="ww" and @lemma="zijn"]
and node[@rel="predc" and
@cat="np" and
node[@rel="det" and
@pt="lid" and @lemma="een"]
and node[@rel="hd" and
@pt="n" and @lemma="zin"]]]
```

XPATH

```
//node[@cat="smain" and
node[@rel="su" and
@pt="vnw" and @lemma="dit"]
and node[@rel="hd" and
@pt="ww" and @lemma="zijn"]
and node[@rel="predc" and
@cat="np" and
node[@rel="det" and
@pt="lid" and @lemma="een"]
and node[@rel="hd" and
@pt="n" and @lemma="zin"]]]
```

KU LEUVEN

XPATH

```
//node[@cat="smain" and
node[@   ="su" and
@pt=      nd @    dit"]
and
@pt="w          'zijn"]
and node      dc" and
@cat="
node
@pt=      nd @      een"]
and no    [@rel="hd      d
@pt="n" and @lemma="zin"]]]
```

KU LEUVEN

XPATH

KU LEUVEN

GrETEL

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

- **Query treebanks by example**

➔ No or limited knowledge of data structures and/or formal query languages needed

KU LEUVEN

the user

1. Example sentence
2. Indicate relevant items of the sentence
3. (Adapt XPath) Select treebank
4. Inspect results

GrETEL

- Parser (Alpino)
- Automatically generate XPath expression
- Present results

KU LEUVEN

OUTLINE

- GrETEL in a nutshell
- **GrETEL demo**
  - **Case study**
  - Search options
- Conclusions and future work

KU LEUVEN

CASE STUDY

Number agreement in quantifying noun constructions

Een + noun + plural noun + singular/plural verb form

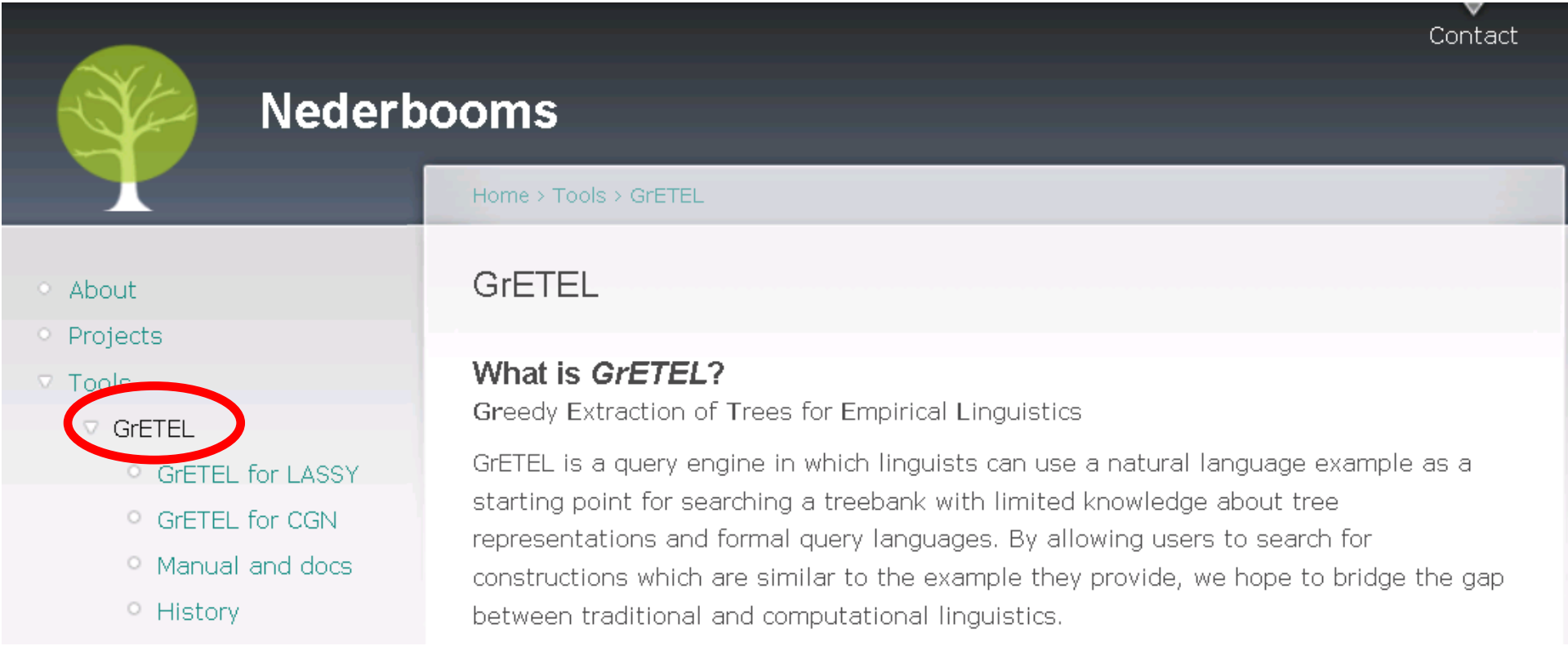
- *Een aantal treinen heeft vertraging.*
  A number trains has-SG delay

- *Een aantal treinen hebben vertraging.*
  A number trains have-PL delay

"A number of trains are running late."

KU LEUVEN

GrETEL ONLINE



KU LEUVEN

INPUT

KU LEUVEN

ANNOTATION MATRIX

Please indicate the relevant parts of the sentence. The syntactic properties of the relevant items are automatically included.

sentence		Een	aantal	treinen	heeft	vertraging	.
relevant nodes	pos	○	◉	○	○	○	○
	extended pos	○	○	◉	◉	○	○
	lemma	◉	○	○	○	○	○
	token	○	○	○	○	○	○
optional nodes		○	○	○	○	◉	◉

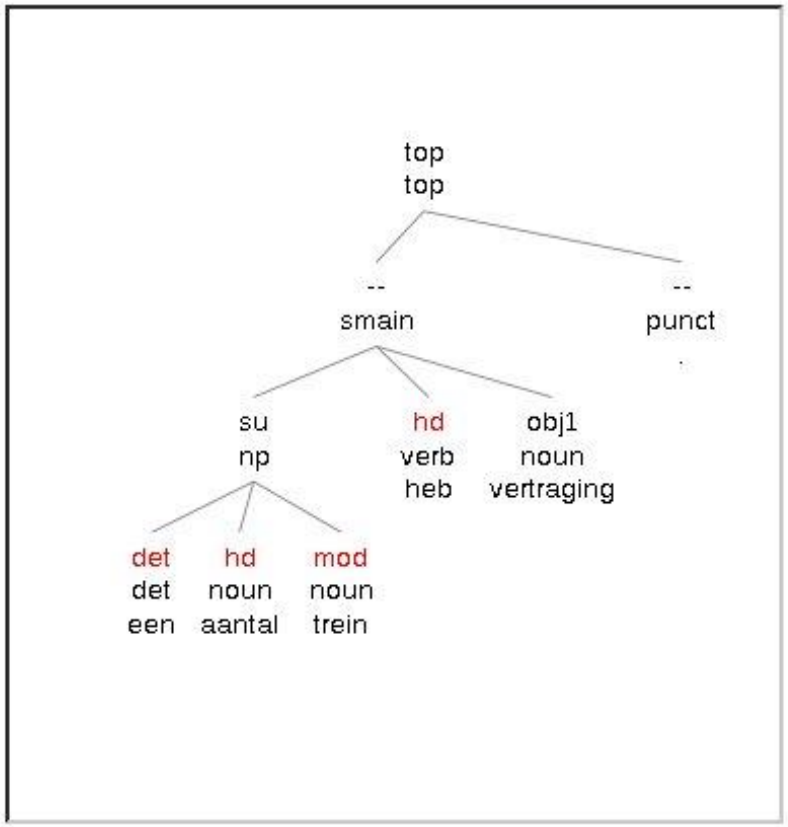
KU LEUVEN

ANNOTATION GUIDELINES

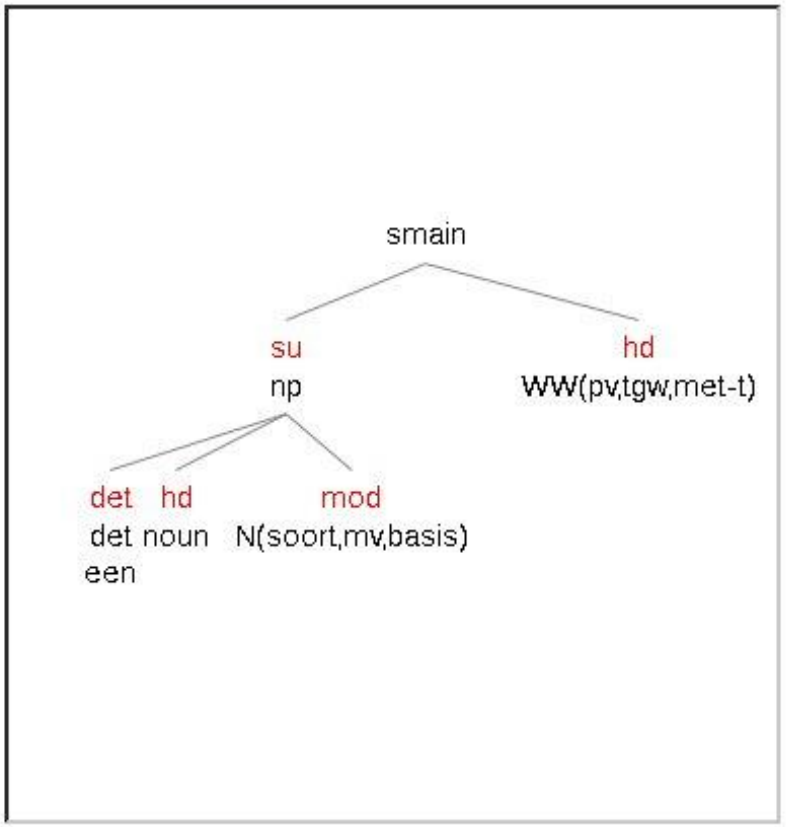
GUIDELINES

- **pos**: Short part-of-speech tag (e.g. noun, verb, prep) [list of all pos tags]
- **extended pos**: Long part-of-speech tag. Besides the part-of-speech, the tags contain more detailed information on person, gender, number, tense, agreement etc. For example: `N(soort,mv,basis)`, `WW(pv,tgw,ev)`, `VNW(pers,pron,nomin,vol,2v,ev)`. [list of all pos tags]
- **lemma**: Word form that generalizes over inflected forms. For example: *zin* is the lemma of *zin*, *zinnen*, and *zinnetje*; *gaan* is the lemma of *ga*, *gaat*, *gaan*, *ging*, *gingen*, and *gegaan*. Lemma is case insensitive (except for proper names). Note that the short part-of-speech tag will be taken into account as well.
- **token**: The exact word form. This is a case sensitive feature. Note that the short part-of-speech tag and the lemma will be taken into account as well.

Alpino parse of the input example [full screen]



Query tree [full screen]



XPATH GENERATOR

XPath query generated from the input example. You can adapt it if necessary. [download original XPath]

```
//node[@cat="smain" and node[@rel="su" and @cat="np" and node[@rel="det" and @root="een" and @pos="det"] and node[@rel="hd" and @pos="noun"] and node[@rel="mod" and @postag="N(soort,mv,basis)" and @pos="noun"]] and node[@rel="hd" and @postag="WW(pv,tgw,met-t)" and @pos="verb"]]
```

KU LEUVEN

TREEBANK SELECTION

LASSY Small

☑	Treebank	Contents	# Sentences	# Words
☑	DPC	Dutch Parallel Corpus	11,716	193,029
☑	Wikipedia	Dutch Wikipedia pages	7,341	83,360
☑	WR-P-E	E-magazines, newsletters, teletext pages, web sites, Wikipedia	14,420	232,631
☑	WR-P-P	Books, brochures, guides and manuals, legal texts, newspapers, periodicals and magazines, policy documents, proceedings, reports, surveys	17,691	281,424
☑	WS-U	Auto cues, news scripts, text for the visually impaired	14,032	184,611
	LASSY Small	Complete treebank	65,200	975,055

KU LEUVEN

RESULTS

***Een* + noun + plural noun + singular verb form**

- *Een aantal treinen heeft vertraging.*
  A number trains has-SG delay
  → **18 matches**

"A number of trains are running late."

RESULTS: table

RESULTS: 18 matches in 18 sentences (out of 65,200 sentences)

Download results[csv]

Show/hide detailed results

TREEBANK	HITS	MATCHING SENTENCE	SENTENCES IN TREEBANK
DPC	1	1	11,716
Wikipedia	0	0	7,341
WR-P-E	5	5	14,420
WR-P-P	8	8	17,691
WS-U	4	4	14,032
TOTAL	**18**	**18**	**65,200**

KU LEUVEN

RESULTS: data

SENTENCE ID	MATCHING SENTENCE	HITS	DISPLAY OPTIONS
WR-P-E-I-0000020972.p.4.s.59.2	**Een kolonie hommels sterft elk najaar** , alleen de bevruchte jonge koninginnen blijven in leven en overwinteren alleen .	1	[full screen] [XML]
WR-P-P-H-0000000076.p.1.s.7	**Een swingende groep kleuter-groupies probeert ondertussen uitgelaten het podium te beklimmen** .	1	[full screen] [XML]
WR-P-P-I-0000000111.p.2.s.1	**Een tiental commissies in het Amerikaanse Congres begint eveneens aan zijn onderzoek naar het failliet** .	1	[full screen] [XML]
WS-U-E-A-0000000206.p.18.s.7	**Eenmaal geland verslindt een wolk sprinkhanen in korte tijd de oogst van een heel seizoen** .	1	[full screen] [XML]
WS-U-E-A-0000000011.p.24.s.1	**Er dient zich een nieuwe generatie schaatsters aan** , dat bewees Mark Tuitert afgelopen weekend in Heerenveen : want hij was na vier afstanden de sterkste : en is de nieuwe Europese Kampioen allround .	1	[full screen] [XML]
WR-P-E-I-0000039352.p.3.s.259	**Gelukkig voor de talloze toeschouwers krijgt een aantal schutters weer snel een ritme te pakken** .	1	[full screen] [XML]
WS-U-E-A-0000000215.p.25.s.2	**In een voorstadje van Boedapest in Hongarije heeft zich een serie explosies voorgedaan in een vuurwerkfabriek** .	1	[full screen] [XML]

KU LEUVEN

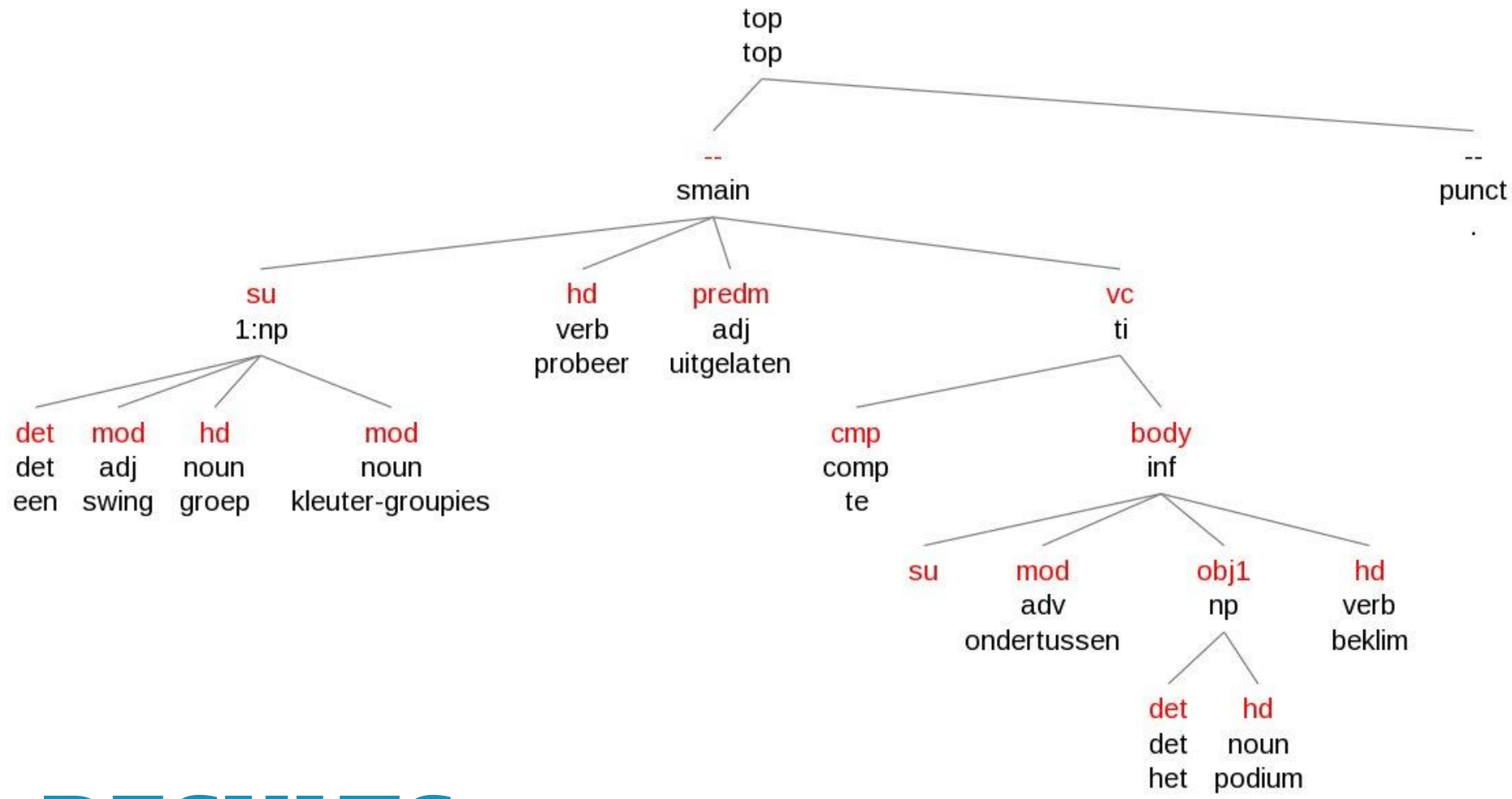
RESULTS: data

SENTENCE ID	MATCHING SENTENCE	HITS	DISPLAY OPTIONS
WR-P-E-I-0000020972.p.4.s.59.2	**Een kolonie hommels sterft elk najaar** , alleen de bevruchte jonge koninginnen blijven in leven en overwinteren alleen .	1	[full screen] [XML]
WR-P-P-H-0000000076.p.1.s.7	**Een swingende groep kleuter-groupies probeert ondertussen uitgelaten het podium te beklimmen** .	1	[full screen] [XML]
WR-P-P-I-0000000111.p.2.s.1	**Een tiental commissies in het Amerikaanse Congres begint eveneens aan zijn onderzoek naar het failliet** .	1	[full screen] [XML]
WS-U-E-A-0000000206.p.18.s.7	**Eenmaal geland verslindt een wolk sprinkhanen in korte tijd de oogst van een heel seizoen** .	1	[full screen] [XML]
WS-U-E-A-0000000011.p.24.s.1	**Er dient zich een nieuwe generatie schaatsters aan** , dat bewees Mark Tuitert afgelopen weekend in Heerenveen : want hij was na vier afstanden de sterkste : en is de nieuwe Europese Kampioen allround .	1	[full screen] [XML]
WR-P-E-I-0000039352.p.3.s.259	**Gelukkig voor de talloze toeschouwers krijgt een aantal schutters weer snel een ritme te pakken** .	1	[full screen] [XML]
WS-U-E-A-0000000215.p.25.s.2	**In een voorstadje van Boedapest in Hongarije heeft zich een serie explosies voorgedaan in een vuurwerkfabriek** .	1	[full screen] [XML]

“greedy” search

KU LEUVEN

Een swingende groep kleuter-groupies probeert ondertussen uitgelaten het podium te beklimmen .



RESULTS: trees

KU LEUVEN

RESULTS

Een + noun **+ singular/plural verb form**

- *Een aantal treinen <u>heeft</u> vertraging.*
  A number trains has-SG delay
  → **18 matches**

- *Een aantal treinen <u>hebben</u> vertraging.*
  A number trains have-PL delay
  → **30 matches**

"A number of trains are running late."

KU LEUVEN

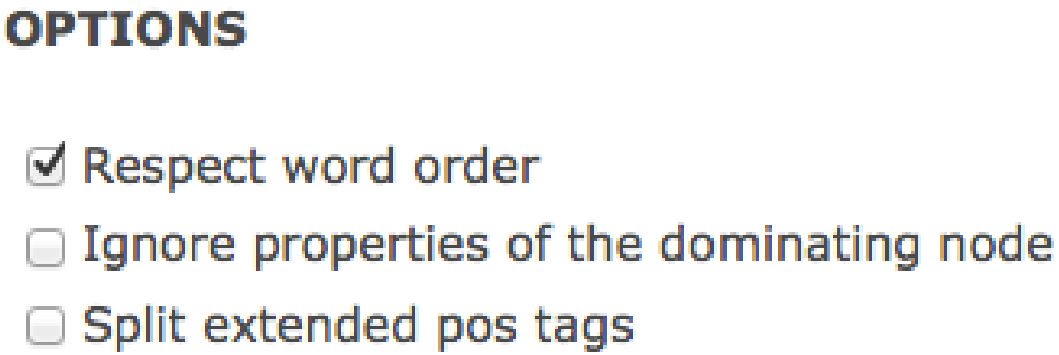
OUTLINE

- GrETEL in a nutshell
- **GrETEL demo**
  - Case study
  - **Search options**
- Conclusions and future work

KU LEUVEN

SEARCH OPTIONS

→ Below annotation matrix



KU LEUVEN

SEARCH OPTIONS

PP-over-V

- V + PP
- … *dat hij <u>opstond</u> <u>met een kater</u>.*
  … that he woke-up with a hangover

- PP + V
- … *dat hij <u>met een kater</u> <u>opstond</u>*.
  … that he with a hangover woke-up

"… that he woke up with a hangover."

KU LEUVEN

SEARCH OPTIONS

PP-over-V

- V + PP
- … *dat hij opstond met een kater.*
  … that he woke-up with a hangover

WR-P-E-I-0000039352.p.3.s.39	De verscheidenheid qua bordjes is groot , zo loopt bijvoorbeeld Schutterij 't Zandakker Gilde Sint Jan Venray met een prachtig roodkoperen geslagen draagbord , terwijl **andere verenigingen met een houten draagbord lopen** .	1	[full screen] [XML]
dpc-vla-001171-nl-sen.p.29.s.6	Het was in het Kaaitheater dat **Anne Teresa De Keersmaeker een podium vond in het begin van haar carrière** .	1	[full screen] [XML]
dpc-eli-000943-nl-sen.p.24.s.1	Het Comité voor geneesmiddelen voor menselijk gebruik (CHMP) heeft geconcludeerd dat **Ariclaim een bescheiden effect vertoonde in studies** maar dat dit effect een voordeel kan zijn voor vrouwen met matige tot ernstige stress-urine-incontinentie , daar er momenteel geen alternatieve farmacologische behandeling voor deze aandoening bestaat .	1	[full screen] [XML]
dpc-vla-001161-nl-sen.p.54.s.1	Een goed jaar na de start van het nieuwe financieringsinstrument kunnen we besluiten dat **ARKimedes na een zwakke start nu toch wel goed op gang komt** .	1	[full screen] [XML]

2,895 matches in 2,769 sentences

But: results include PP + V as well!

KU LEUVEN

SEARCH OPTIONS

PP-over-V

- V + PP **+ word order option**
- … *dat hij opstond met een kater.*
  … that he woke-up with a hangover

WS-U-E-A-0000000218.p.21.s.3	Het ging destijds eigenlijk al meteen mis : het voorprogramma met ondermeer André van Duin en de Fouryo's was veel te rustig voor een publiek **dat kwam voor rok-en-rol** .	1	[full screen] [XML]
dpc-med-000686-nl-sen.p.26.s.3	Het belangrijkste voordeel van het product is de associatie van een lage dosis ethyniloestradiol met een progestativum van de tweede generatie (cfr de polemiek van het einde van vorige eeuw over het thrombo-embolisch risico **dat lager is met tweede generatie progestativum in vergelijking met deze van de derde generatie** .)	1	[full screen] [XML]
WR-P-E-I-0000024561.p.4.s.68.3	Brels eerste album op het Barclay-label bestaat voor een groot deel uit studio-versies van materiaal **dat luisteraars al kenden van het Olympia-concert** .	1	[full screen] [XML]
dpc-rou-000983-nl-sen.p.8.s.11	Het verhaal van Rodenbach , **dat net als wijn rijpt op eikenhouten vaten** , past daar perfect bij , maar was nog niet verteld . "	1	[full screen] [XML]

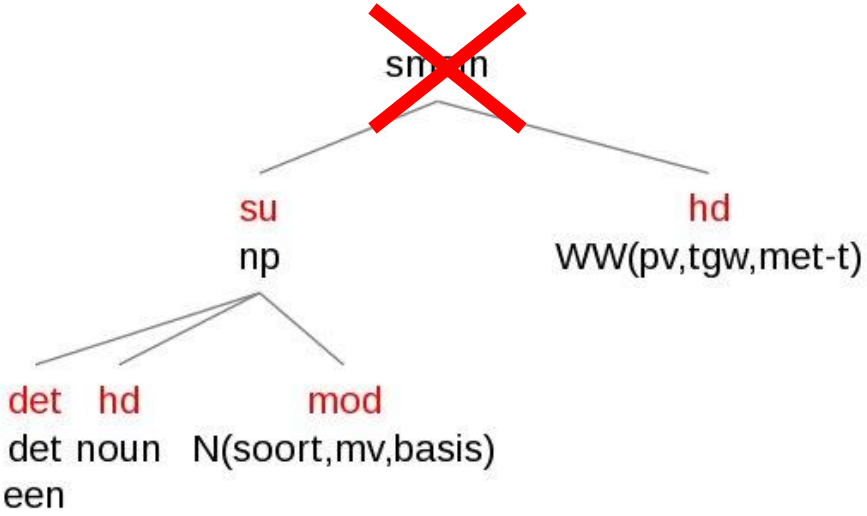
789 matches in 777 sentences
Results only include V + PP

KU LEUVEN

SEARCH OPTIONS

OPTIONS

- [ ] Respect word order
- [x] Ignore properties of the dominating node
- [ ] Split extended pos tags



XPath query generated from the input example. You can adapt it if necessary. [download original XPath]

```
//node[@cat and node[@rel="su" and @cat="np" and node[@rel="det" and @root="een" and @pos="det"] and
node[@rel="hd" and @pos="noun"] and node[@rel="mod" and @postag="N(soort,mv,basis)" and @pos="noun"]]
and node[@rel="hd" and @postag="WW(pv,tgw,met-t)" and @pos="verb"]]
```

KU LEUVEN

SEARCH OPTIONS

OPTIONS

- [ ] Respect word order
- [ ] Ignore properties of the dominating node
- [x] Split extended pos tags

XPath query generated from the input example. You can adapt it if necessary. [download original XPath]

```
//node[@cat="smain" and node[@rel="su" and @cat="np" and node[@rel="det" and @root="een" and
@pos="det"] and node[@rel="hd" and @pos="noun"] and node[@rel="mod" and @postag="N(soort,mv,basis)"
and @pos="noun"]] and node[@rel="hd" and @postag="WW(pv,tgw,met-t)" and @pos="verb"]]
```

KU LEUVEN

SEARCH OPTIONS

OPTIONS

- [ ] Respect word order
- [ ] Ignore properties of the dominating node
- [x] Split extended pos tags

XPath query generated from the input example. You can adapt it if necessary. [download original XPath]

```
//node[@cat="smain" and node[@rel="su" and @cat="np" and node[@rel="det" and @root="een" and
@pos="det"] and node[@rel="hd" and @pos="noun"] and node[@rel="mod" and @postag="N(soort,mv,basis)"
and @pos="noun"]] and node[@rel="hd" and @postag="WW(pv,tgw,met-t)" and @pos="verb"]]
```

XPath query generated from the input example. You can adapt it if necessary. [download original XPath]

```
//node[@cat="smain" and node[@rel="su" and @cat="np" and node[@rel="det" and @root="een" and
@pos="det"] and node[@rel="hd" and @pos="noun"] and node[@rel="mod" and @getal="mv" and
@graad="basis" and @pos="noun" and @ntype="soort"]] and node[@rel="hd" and @pvtijd="tgw" and
@wvorm="pv" and @pos="verb" and @pvagr="met-t"]]
```

KU LEUVEN

SEARCH OPTIONS

OPTIONS

- [ ] Respect word order
- [ ] Ignore properties of the dominating node
- [x] Split extended pos tags

XPath query generated from the input example. You can adapt it if necessary. [download original XPath]

```
//node[@cat="smain" and node[@rel="su" and @cat="np" and node[@rel="det" and @root="een" and
@pos="det"] and node[@rel="hd" and @pos="noun"] and node[@rel="mod" and @postag="N(soort,mv,basis)"
and @pos="noun"]] and node[@rel="hd" and @postag="WW(pv,tgw,met-t)" and @pos="verb"]]
```

XPath query generated from the input example. You can adapt it if necessary. [download original XPath]

```
//node[@cat="smain" and node[@rel="su" and @cat="np" and node[@rel="det" and @root="een" and
@pos="det"] and node[@rel="hd" and @pos="noun"] and node[@rel="mod" and @getal="mv" and
@graad="basis" and @pos="noun" and @ntype="soort"]] and node[@rel="hd" and @pvtijd="tgw" and
@wvorm="pv" and @pos="verb" and @pvagr="met-t"]]
```

KU LEUVEN

SEARCH OPTIONS

LASSY Small

☑	Treebank	Contents	# Sentences	# Words
☑	DPC	Dutch Parallel Corpus	11,716	193,029
☑	Wikipedia	Dutch Wikipedia pages	7,341	83,360
☑	WR-P-E	E-magazines, newsletters, teletext pages, web sites, Wikipedia	14,420	232,631
☑	WR-P-P	Books, brochures, guides and manuals, legal texts, newspapers, periodicals and magazines, policy documents, proceedings, reports, surveys	17,691	281,424
☑	WS-U	Auto cues, news scripts, text for the visually impaired	14,032	184,611
	LASSY Small	Complete treebank	65,200	975,055

OPTION

☑ Include context (one sentence before and after the matching sentence)

KU LEUVEN

SEARCH OPTIONS

SENTENCE ID	MATCHING SENTENCE	HITS	DISPLAY OPTIONS
WR-P-E-I-0000020972.p.4.s.59.2	Levenscyclus ***Een kolonie hommels sterft elk najaar*** , *alleen de bevruchte jonge koninginnen blijven in leven en overwinteren alleen* .	1	[full screen] [XML]
WR-P-P-H-0000000076.p.1.s.7	Tien meter lang . ' ***Een swingende groep kleuter-groupies probeert ondertussen uitgelaten het podium te beklimmen*** . Het gaat hier niet om lieve luistermuziek maar om stevige kleuterpop .	1	[full screen] [XML]
WR-P-P-I-0000000111.p.2.s.1	***Een tiental commissies in het Amerikaanse Congres begint eveneens aan zijn onderzoek naar het failliet*** . Een van de commissies wil informatie over het energiebeleid krijgen van de Amerikaanse vice-president Dick Cheney .	1	[full screen] [XML]
WS-U-E-A-0000000206.p.18.s.7	Een sprinkhaan kan per dag zijn eigen gewicht aan voedsel opeten . ***Eenmaal geland verslindt een wolk sprinkhanen in korte tijd de oogst van een heel seizoen*** . Alleen in Mali werden de afgelopen dagen al 42 springhanenzwermen waargenomen .	1	[full screen] [XML]
WS-U-E-A-0000000011.p.24.s.1	***Er dient zich een nieuwe generatie schaatsters aan*** , *dat bewees Mark Tuitert afgelopen weekend in Heerenveen : want hij was na vier afstanden de sterkste : en is de nieuwe Europese Kampioen allround* .	1	[full screen] [XML]
WR-P-E-I-0000039352.p.3.s.259	De eerste ronde valt het zwaarste , getuige het feit dat meestal meer dan de helft van de deelnemers dan reeds ' sneuvelt ' . ***Gelukkig voor de talloze toeschouwers krijgt een aantal schutters weer snel een ritme te pakken*** .	1	[full screen] [XML]

KU LEUVEN

OUTLINE

- GrETEL in a nutshell
- GrETEL demo
  - Case study
  - Search options
- **Conclusions and future work**

KU LEUVEN

CONCLUSIONS

- **GrETEL**: search engine for Dutch treebanks
- Input = natural language example
- Output = sample of similar sentences
- Syntactic concordancer
- Available online (via *Mozilla Firefox*)
- No installation required

KU LEUVEN

FUTURE WORK

- **GrETEL 2.0**
  - Query very large treebanks
  - Improve user interface
  - Include SoNaR corpus (ca 500M tokens, 42M sentences)

- **AfriBooms**
  - GrETEL for Afrikaans
  - Include other treebank formats

KU LEUVEN

Try it yourself at

http://nederbooms.ccl.kuleuven.be/eng/gretel

Questions? Comments? Suggestions?

Contact us at gretel@ccl.kuleuven.be

Thanks for your attention!

KU LEUVEN