KU LEUVEN

Lumberjacking with GrETEL

Liesbeth Augustinus
Vincent Vandeghinste
Ineke Schuurman
Frank Van Eynde

BKL taaldag - May 11, 2013

GrETEL

- **Gr**eedy **E**xtraction of **T**rees for **E**mpirical **L**inguistics
- Query engine for treebanks
- **Nederbooms project**
  Exploitation of Dutch treebanks
  for research in linguistics

CLARIN

Common Language Resources and Technology Infrastructure

KU LEUVEN

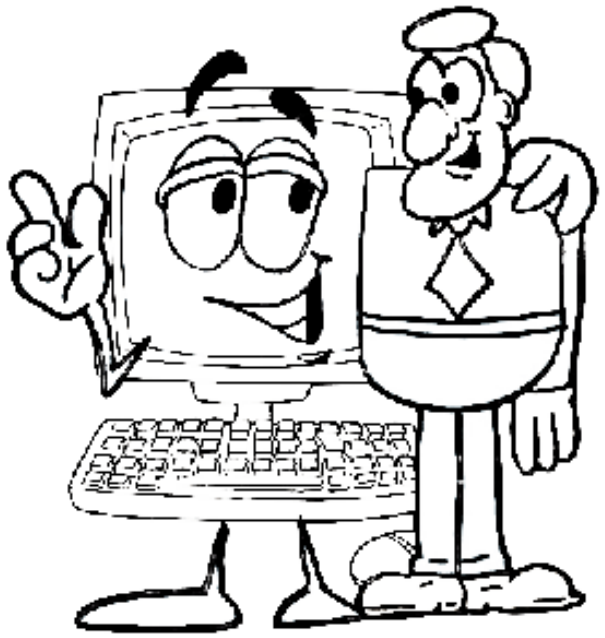
GrETEL

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

- Query engine for treebanks

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

- **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 core corpus	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

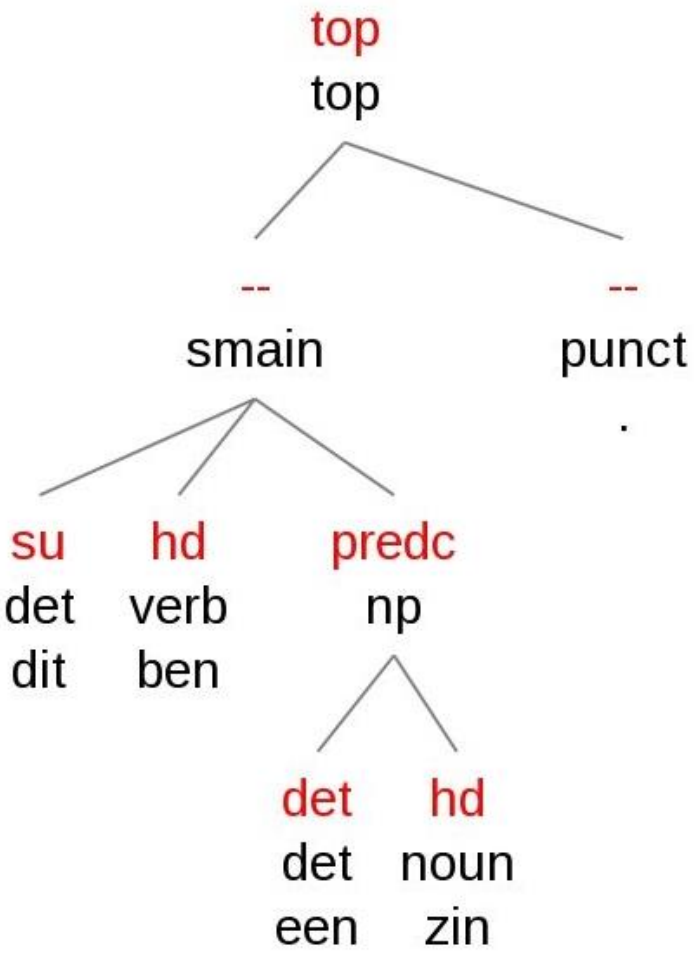
- **Treebank** = syntactically annotated corpus
  e.g. Penn Treebank (English), TüBa (German),
  LASSY, CGN (Dutch)

- **Parser**
  e.g. Alpino (Van Noord 2006)

ALPINO PARSER

Dit is een zin. >> ALPINO parser >>

"This is a sentence."

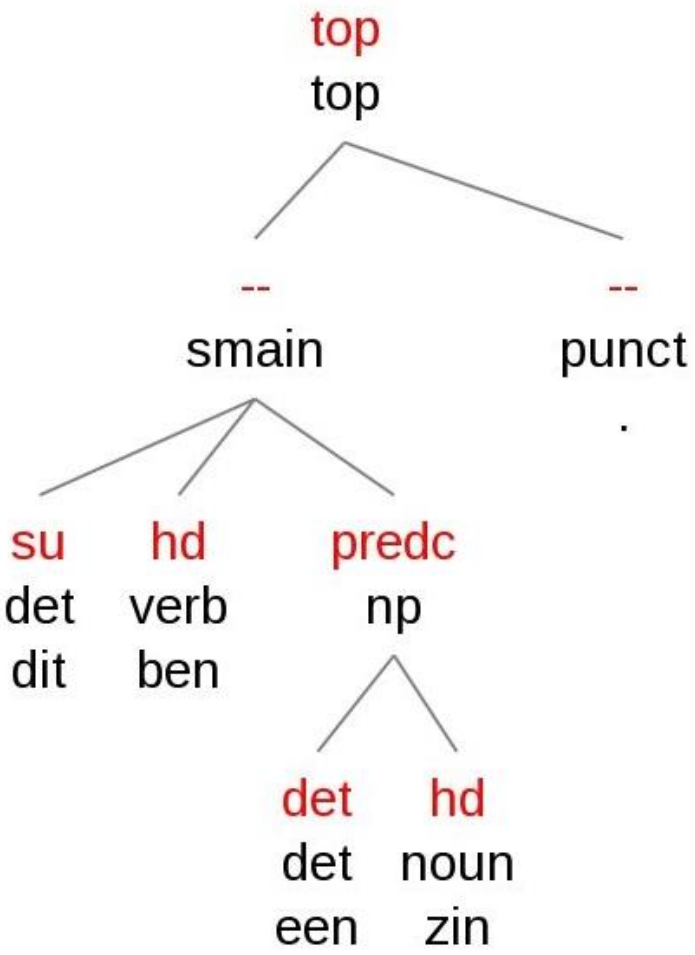


KU LEUVEN

ALPINO PARSER

Dit is een zin. >> ALPINO parser >>

"This is a sentence."

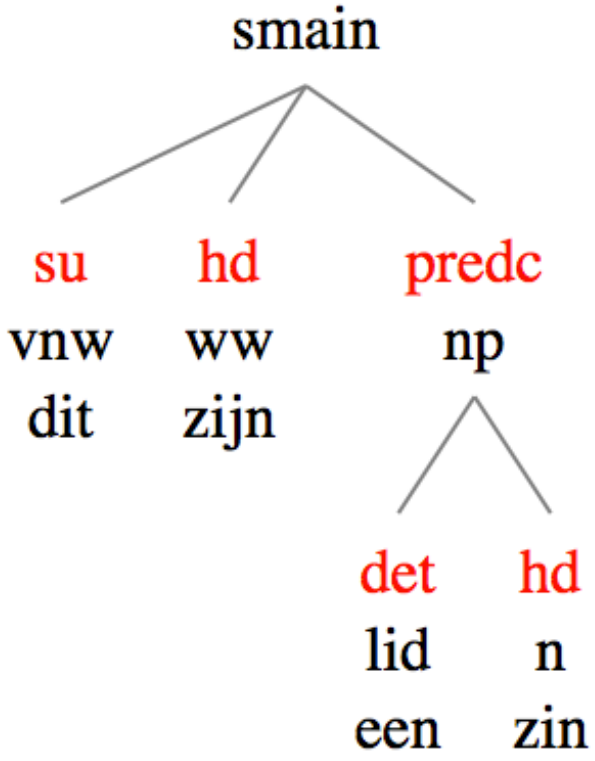


XML trees

Query language: **XPath**

KU LEUVEN

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[@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**

KU LEUVEN

GrETEL

- **Gr**eedy **E**xtraction of **T**rees for **E**mpirical **L**inguistics
- **Query treebanks by example**
- First version
  => only for LASSY treebank
- New release
  => GrETEL for CGN treebank
  => update based on user reviews

KU LEUVEN

the user

- Example sentence
- Indicate relevant items of the sentence
- (Adapt XPath)
- Select treebank
- 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

- Collective noun constructions
  - E.g. *<u>Een aantal bomen</u> zijn omgevallen.*
    'A number of trees fell down.'
  - DET + NOUN + PLURAL NOUN

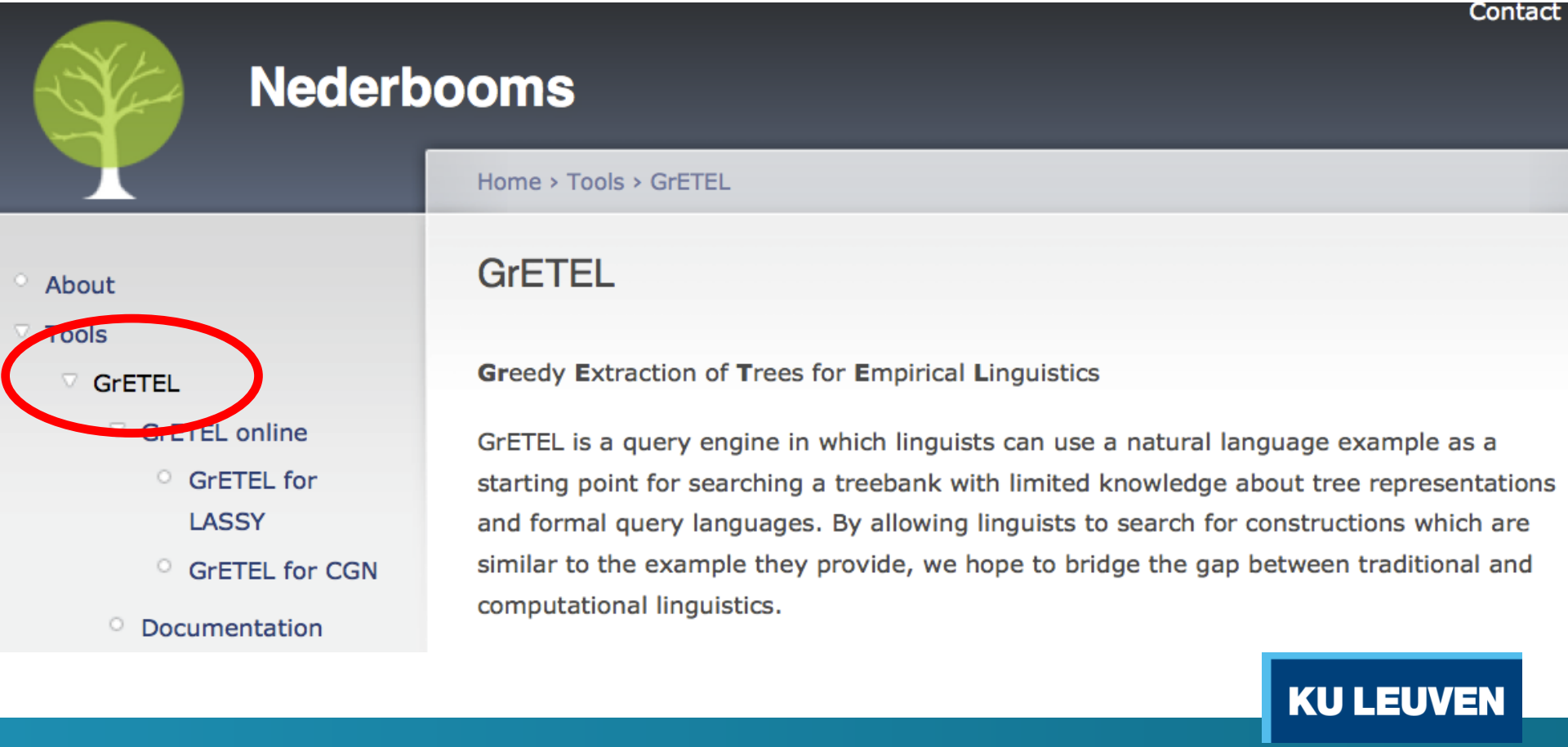
KU LEUVEN

CASE STUDY

- Collective noun constructions
  - E.g. *Een aantal bomen zijn omgevallen.*
    'A number of trees fell down.'
  - DET + NOUN + PLURAL NOUN

- Discontinuous constructions!
  - E.g. *Een groot aantal oude bomen zijn omgevallen.*
    'A large number of old trees fell down.'

GrETEL ONLINE



KU LEUVEN

INPUT


Nederbooms
Home › Tools › GrETEL › GrETEL online › GrETEL for CGN
About
Tools
GrETEL
GrETEL online
GrETEL for LASSY
GrETEL for CGN
GrETEL for CGN (v1.1)
Please provide an input example
Een aantal bomen zijn omgevallen.
Submit

ANNOTATION MATRIX

GrETEL for CGN

Please indicate the relevant parts of the sentence

sentence		Een	aantal	bomen	zijn	omgevallen	.
relevant nodes	**pos**	◉	◉	○	○	○	○
	extended pos	○	○	◉	○	○	○
	lemma	○	○	○	○	○	○
	token	○	○	○	○	○	○
optional nodes		○	○	○	◉	◉	◉

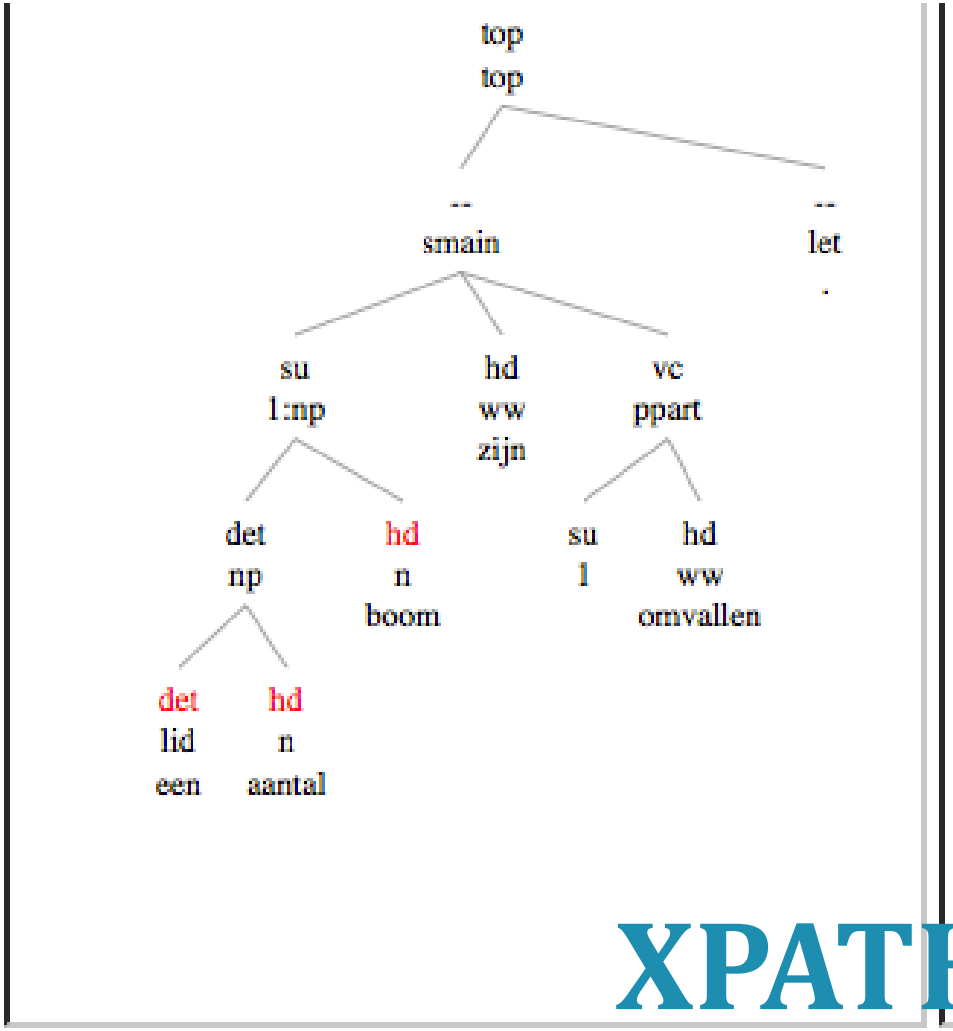
KU LEUVEN

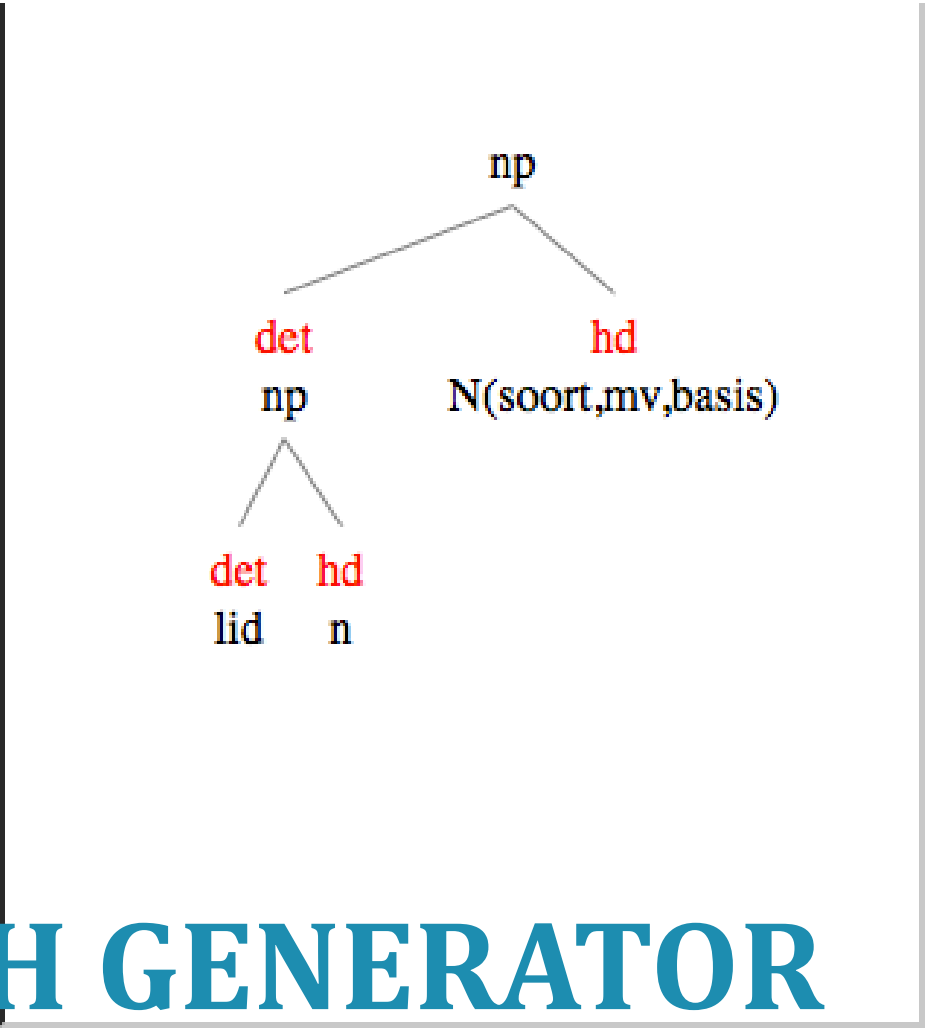
ANNOTATION GUIDELINES

GUIDELINES

- **pos**: Short part-of-speech tag. The different tags are: `n`, `ww`, `adj`, `lid`, `vnw`, `vg`, `bw`, `tw`, `vz`, `tsw`, `spec`, and `let`.
- **extended pos**: Long part-of-speech tag. 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).
- **token**: The exact word form. This is a case sensitive feature.

KU LEUVEN





XPATH GENERATOR

XPath query generated from the input example. You can adapt it if necessary.

```
//node[@cat="np" and node[@rel="det" and @cat="np" and node[@rel="det" and @pt="lid"] and node[@rel="hd" and @pt="n"]] and node[@rel="hd" and @pt="n" and @postag="N(soort,mv,basis)"]]
```

TREEBANK SELECTION

CGN core corpus

Treebank		Contents	# Sentences	# Words	# Sentences	# Words	# Sentences	# Words
☑ NL	☑ VL		NL		VL		TOTAL	
☑ NA	☑ VA	Spontaneous conversations (\'face-to-face\')	50,239	302,828	22,881	147,418	73,120	450,246
☑ NB	☑ VB	Interviews with teachers of Dutch	2,484	25,724	4,289	34,158	6,773	59,882
☑ NC	☑ VC	Telephone conversations (recorded via a switchboard)	11,649	70,084	3,142	19,984	14,791	90,068
☐ ND	☑ VD	Telephone conversations (recorded on MD)	0	0	929	6,309	929	6,309
☑ NE	☐ VE	Simulated business negotiations	3,123	25,524	0	0	3,123	25,524
☑ NF	☑ VF	Interviews/discussions/debates (broadcast)	6,290	75,167	2,617	25,122	8,907	100,289
☑ NG	☑ VG	(Political) discussions/debates /meetings (non-broadcast)	1,166	25,125	543	9,009	1,709	34,134
☑ NH	☑ VH	Lessons recorded in the classroom	3,064	26,004	1,395	10,116	4,459	36,120
☑ NI	☑ VI	Live (sports) commentaries (broadcast)	2,251	25,002	1,026	10,147	3,277	35,149

KU LEUVEN

RESULTS

Collective noun constructions

- E.g. *Een aantal bomen zijn omgevallen.*
  'A number of trees fell down.'
- DET + NOUN + PLURAL NOUN

➔ 594 matches in 576 sentences

RESULTS: table

SEARCH RESULTS

treebank	hits	matching sentences	sentences	ratio (matching sentences/sentences)
NA	92	88	50239	0.18 %
VA	79	76	22881	0.33 %
NB	29	28	2484	1.13 %
VB	24	23	4289	0.54 %
NC	28	28	11649	0.24 %
VC	9	9	3142	0.29 %
VD	2	2	929	0.22 %
NE	5	5	3123	0.16 %
NF	60	56	6290	0.89 %
VF	27	27	2617	1.03 %
NG	27	26	1166	2.23 %
VG	11	11	543	2.03 %
NH	11	11	3064	0.36 %
VH	5	5	1395	0.36 %

RESULTS: data

Sentence ID	Matching sentences	
fvb400145__607	maar xxx na zo'n xxx bijvoorbeeld we heb*a we hebben dan vroeger soaps helemaal uitgeschreven in draaiboeken **een paar scènes** hè want anders is dat veel te lang en dan die in scène gezet en gefilmd en dan moet daarna de theorie komen rond soap van wat is een soap wie kijkt er naar soaps kenmerken van ...	[full screen] [XML]
fvb400165__202	allee d'r zijn toch een een **een pak woorden waar er geen regel voor bestaat** .	[full screen] [XML]
fvb400165__357	en in die toneelvereniging we zaten daar met een heel uh bende spelers laat ons zeggen met vijfentwintig spelers waarvan er zeker twintig uh toneel gevolgd hadden .	[full screen] [XML]
fvb400165__58	die werd wel gelezen doo*a door een bepaald aa*a door een een ja **een groep studenten** eigenlijk uh .	[full screen] [XML]
fvb400165__63	hij herhaalt zichzelf uh na **een aantal bladzijden** hè .	[full screen] [XML]

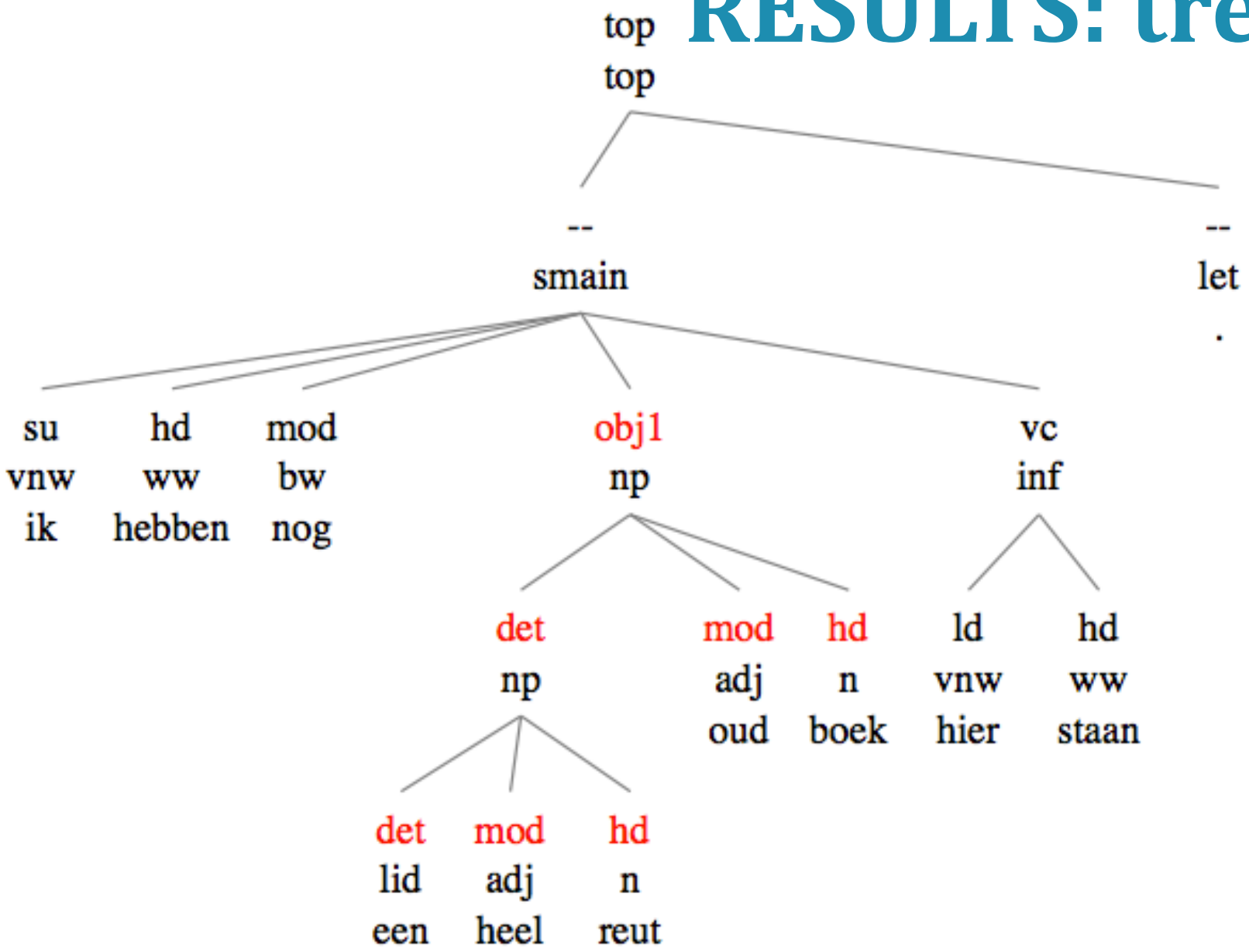
KU LEUVEN

RESULTS: data

Sentence ID	Matching sentences	
fnc008003__58	dan dan zijn jullie gewoon **een stelletje mutsen die aan de telefoon hangen** .	[full screen] [XML]
fnc008005__117	ik had vorige week had ik uh **een paar batterijen** gehaald .	[full screen] [XML]
fnc008007__272	ik bedoel uh de Queen Mum die dronk tenminste nog **een paar whisky's** elke avond om vervolgens in een diep coma weg te dommelen .	[full screen] [XML]
fnc008006__344	ik heb hier nog **een hele reut ouwe boeken** staan .	[full screen] [XML]
fnc008006__77	was een kerel en die had z'n die die timmerde z'n vrouw l*a al uh structureel in elkaar een aantal jaren .	[full screen] [XML]
fnc008007__71	ja d'r zijn toch **een aantal kinderen die 't gezien hebben** .	[full screen] [XML]
fnc008012__58	met **een aantal auto's** ?	[full screen] [XML]
fnc008013__100	want die eten in de zomer **een beetje bessen of zo** en en uh ete*a eten ze alleen maar vlees hè .	[full screen] [XML]

ik heb hier nog een hele reut ouwe boeken staan .

RESULTS: trees



LEUVEN

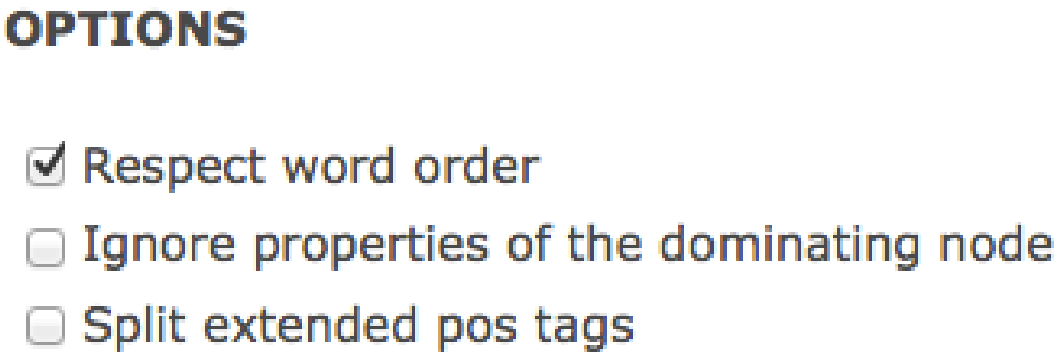
OUTLINE

- 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

Green versus red word order in Dutch

- green: past participle – auxiliary

  *De NAVO stelt dat ze er alles aan **gedaan heeft***

- red: auxiliary – past participle

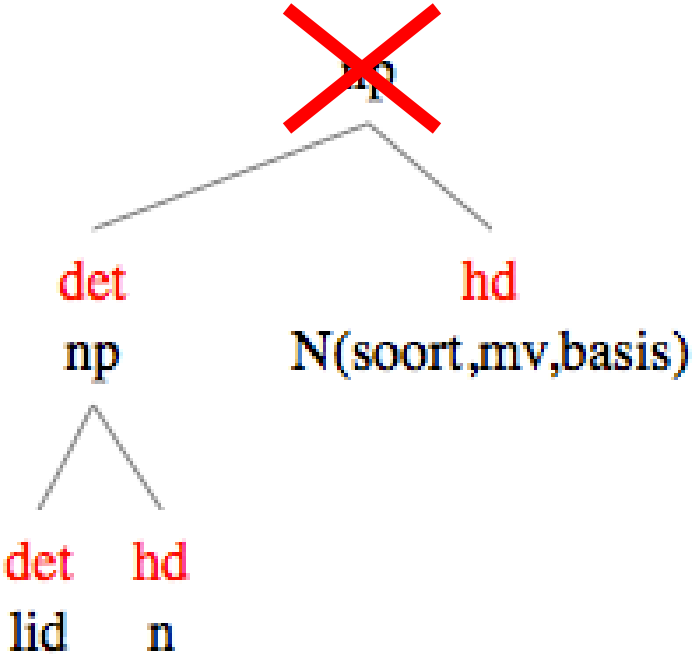
  *De NAVO stelt dat ze er alles aan **heeft gedaan***

"The NATO claim that they have done everything in their power" (deredactie.be)

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.

```
//node[@cat and node[@rel="det" and @cat="np" and node[@rel="det" and @pt="lid"] and
node[@rel="hd" and @pt="n"]] and node[@rel="hd" and @pt="n" and @postag="N(soort,mv,basis)"]]
```

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.

```
//node[@cat="np" and node[@rel="det" and @cat="np" and node[@rel="det" and @pt="lid"] and
node[@rel="hd" and @pt="n"]] and node[@rel="hd" and @pt="n" and @postag="N(soort,mv,basis)"]]
```

XPath query generated from the input example. You can adapt it if necessary.

```
//node[@cat="np" and node[@rel="det" and @cat="np" and node[@rel="det" and @pt="lid"] and
node[@rel="hd" and @pt="n"]] and node[@rel="hd" and @pt="n" and @getal="mv" and
@graad="basis" and @ntype="soort"]]
```

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.

```
//node[@cat="np" and node[@rel="det" and @cat="np" and node[@rel="det" and @pt="lid"] and
node[@rel="hd" and @pt="n"]] and node[@rel="hd" and @pt="n" and @postag="N(soort,mv,basis)"]]
```

XPath query generated from the input example. You can adapt it if necessary.

```
//node[@cat="np" and node[@rel="det" and @cat="np" and node[@rel="det" and @pt="lid"] and
node[@rel="hd" and @pt="n"]] and node[@rel="hd" and @pt="n" and @getal="mv" and
@graad="basis" and @ntype="soort"]]
```

KU LEUVEN

SEARCH OPTIONS

CGN core corpus

Treebank		Contents	# Sentences	# Words	# Sentences	# Words	# Sentences	# Words
☑ NL	☑ VL		NL		VL		TOTAL	
☑ NA	☑ VA	Spontaneous conversations (\'face-to-face\')	50,239	302,828	22,881	147,418	73,120	450,246
☑ NB	☑ VB	Interviews with teachers of Dutch	2,484	25,724	4,289	34,158	6,773	59,882
☑ NC	☑ VC	Telephone conversations (recorded via a switchboard)	11,649	70,084	3,142	19,984	14,791	90,068
☐ ND	☑ VD	Telephone conversations (recorded on MD)	0	0	929	6,309	929	6,309
☑ NE	☐ VE	Simulated business negotiations	3,123	25,524	0	0	3,123	25,524
☑ NF	☑ VF	Interviews/discussions/debates (broadcast)	6,290	75,167	2,617	25,122	8,907	100,289
☑ NG	☑ VG	(Political) discussions/debates /meetings (non-broadcast)	1,166	25,125	543	9,009	1,709	34,134
☑ NH	☑ VH	Lessons recorded in the classroom	3,064	26,004	1,395	10,116	4,459	36,120
☑ NI	☑ VI	Live (sports) commentaries (broadcast)	2,251	25,002	1,026	10,147	3,277	35,149

☑ Include context

KU LEUVEN

SEARCH OPTIONS

Sentence ID	Matching sentences	
fnc008003__58	weet ik eigenlijk niet wie dan 't meest kletst . *dan dan zijn jullie gewoon* ***een stelletje mutsen die aan de telefoon hangen*** . nou dat is helemaal niet waar .	[full screen] [XML]
fnc008005__117	wat moest jij nog in 't dorp vanmiddag ? *ik had vorige week had ik uh* ***een paar batterijen*** *gehaald* . want ik had batterijtjes nog 'ns een keer van die ouwe dinge uit één of andere koffiemolen of weet ik waar of ze in zaten maar die uh oplaadbare .	[full screen] [XML]
fnc008007__272	uh d'r zijn mensen die die die roken als als als ... *ik bedoel uh de Queen Mum die dronk tenminste nog* ***een paar whisky's*** *elke avond om vervolgens in een diep coma weg te dommelen* . ja ?	[full screen] [XML]
fnc008006__344	ik ga morgenvroeg wel effe naar 't post*a ... *ik heb hier nog* ***een hele reut ouwe boeken*** *staan* . of ouwe boeken .	[full screen] [XML]
fnc008006__77	ja dat was vandaag in dat waargebeurde maandagavonddrama ook . *was een kerel en die had z'n die die timmerde z'n vrouw l*a al uh structureel in elkaar een aantal jaren* . en toen op een gegeven moment ging ze bij d'r xxx weg en nieuw appartementje hele reutemeteut bijgestaan door d'r zuster .	[full screen] [XML]
fnc008007__71	hebben hebben andere kinderen 't gezien ? *ja d'r zijn toch* ***een aantal kinderen die 't gezien hebben*** . en dat is ook de reden waarom de leraren ge*a geattendeerd werden en uh ggg andere uh de de leraar die kwam ook aan die die zei nou ik ik rook 't meteen .	[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**
  - Include SoNaR corpus (ca 500M tokens)
  - More generic

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

KU LEUVEN

Try it yourself at

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

Thanks for your attention!