KU LEUVEN

Making a large treebank searchable online

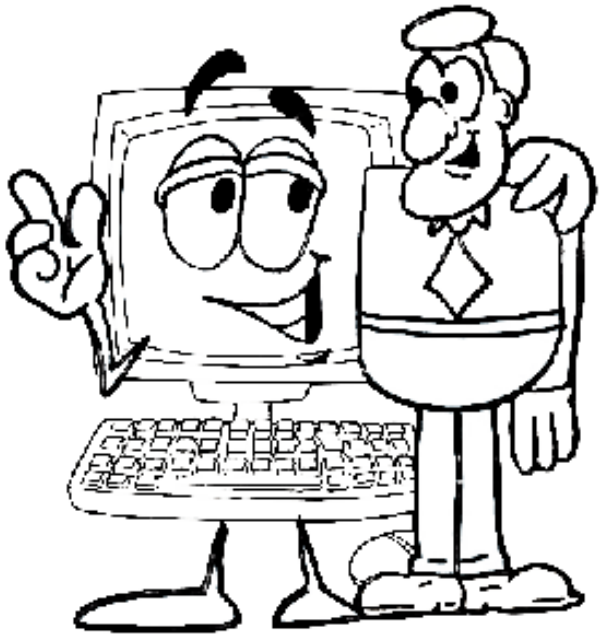
The SoNaR case

Vincent Vandeghinste
Liesbeth Augustinus

CMLC-2 - May 31, 2014

GrETEL

- **Exploitation of Dutch treebanks for research in linguistics**
- CLARIN-NTU project
- **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
- **GrETEL 1.0:** 2 treebanks, 1M words each
- **GrETEL 2.0:** SoNaR treebank, 500M tokens

Goal: scale up the search engine

KU LEUVEN

OUTLINE

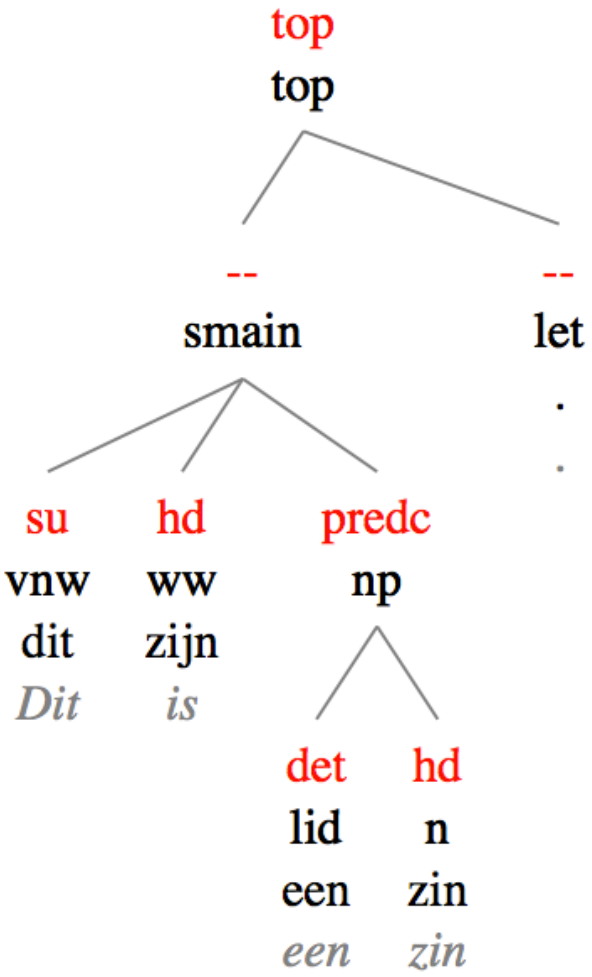
- **GrETEL in a nutshell**
- GrInd: indexing the database
- Conclusions and future work

KU LEUVEN

ALPINO PARSER

Dit is een zin. >> ALPINO parser >>

"This is a sentence."



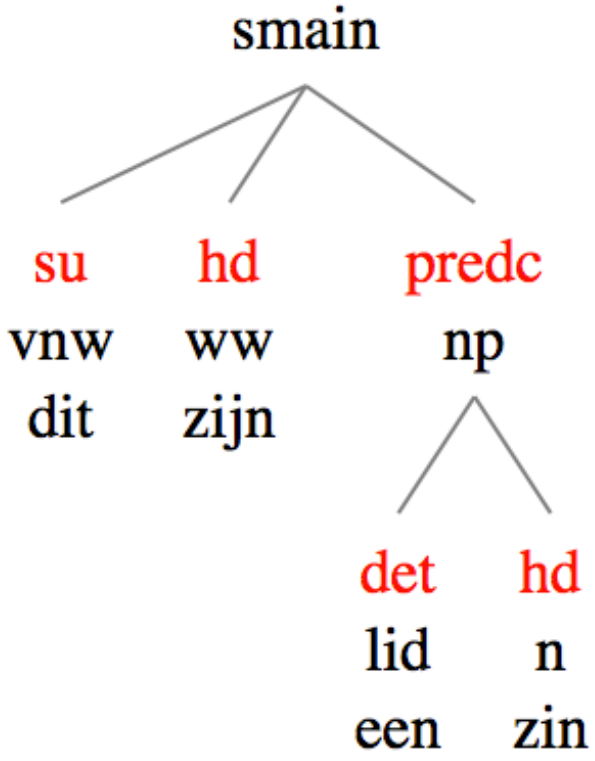
Van Noord (2006)

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"]]]
```

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

INPUT

KU LEUVEN

ANNOTATION MATRIX

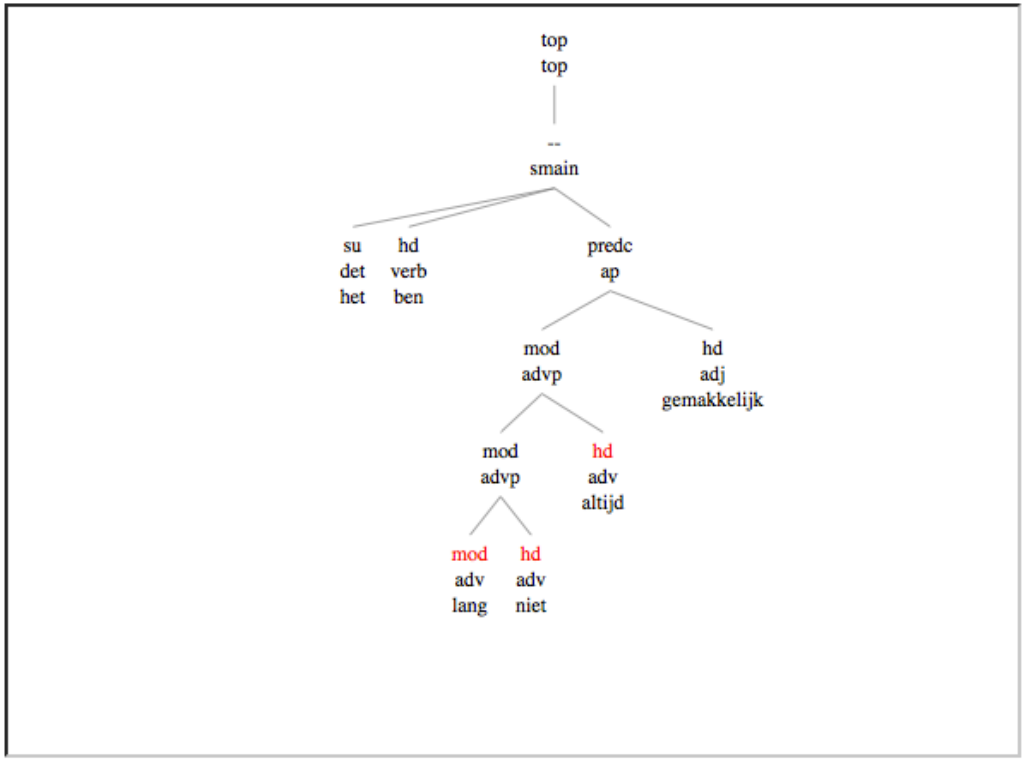
Please indicate the relevant parts of the sentence. The syntactic properties of the relevant items are automatically included.

sentence		Het	is	lang	niet	altijd	gemakkelijk
relevant nodes	**pos**	○	○	◉	◉	◉	○
	extended pos	○	○	○	○	○	○
	lemma	○	○	○	○	○	○
	token	○	○	○	○	○	○
optional nodes		◉	◉	○	○	○	◉

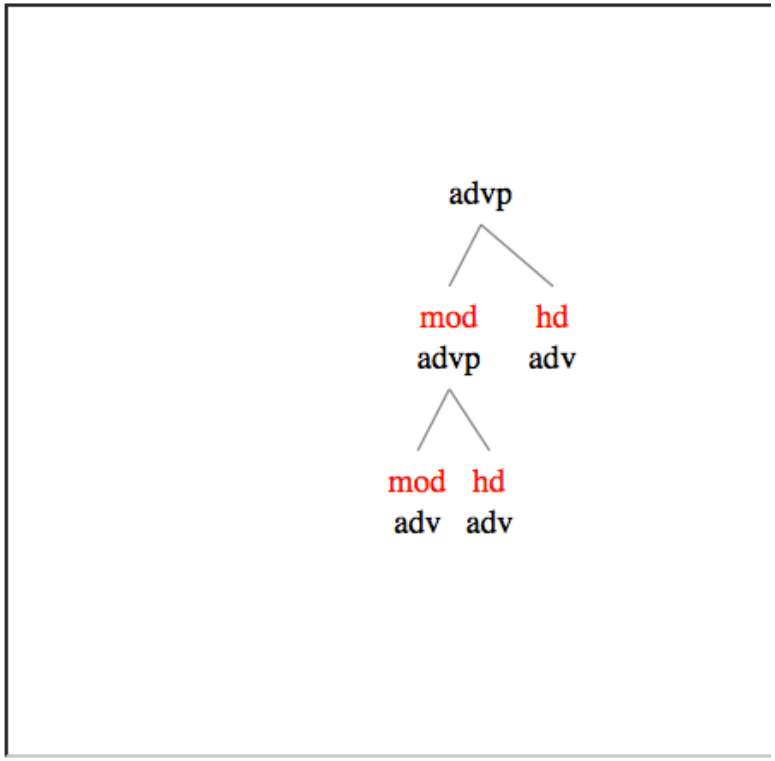
KU LEUVEN

XPATH GENERATOR

Alpino parse of the input example [full screen]



Query tree [full screen]



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

[download original XPath]

```
//node[@cat="advp" and node[@rel="mod" and @cat="advp" and node[@rel="mod" and
@pos="adv"] and node[@rel="hd" and @pos="adv"]] and node[@rel="hd" and @pos="adv"]]
```

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

Input example: *Het is **lang niet altijd** gemakkelijk.*

"It is far from easy."

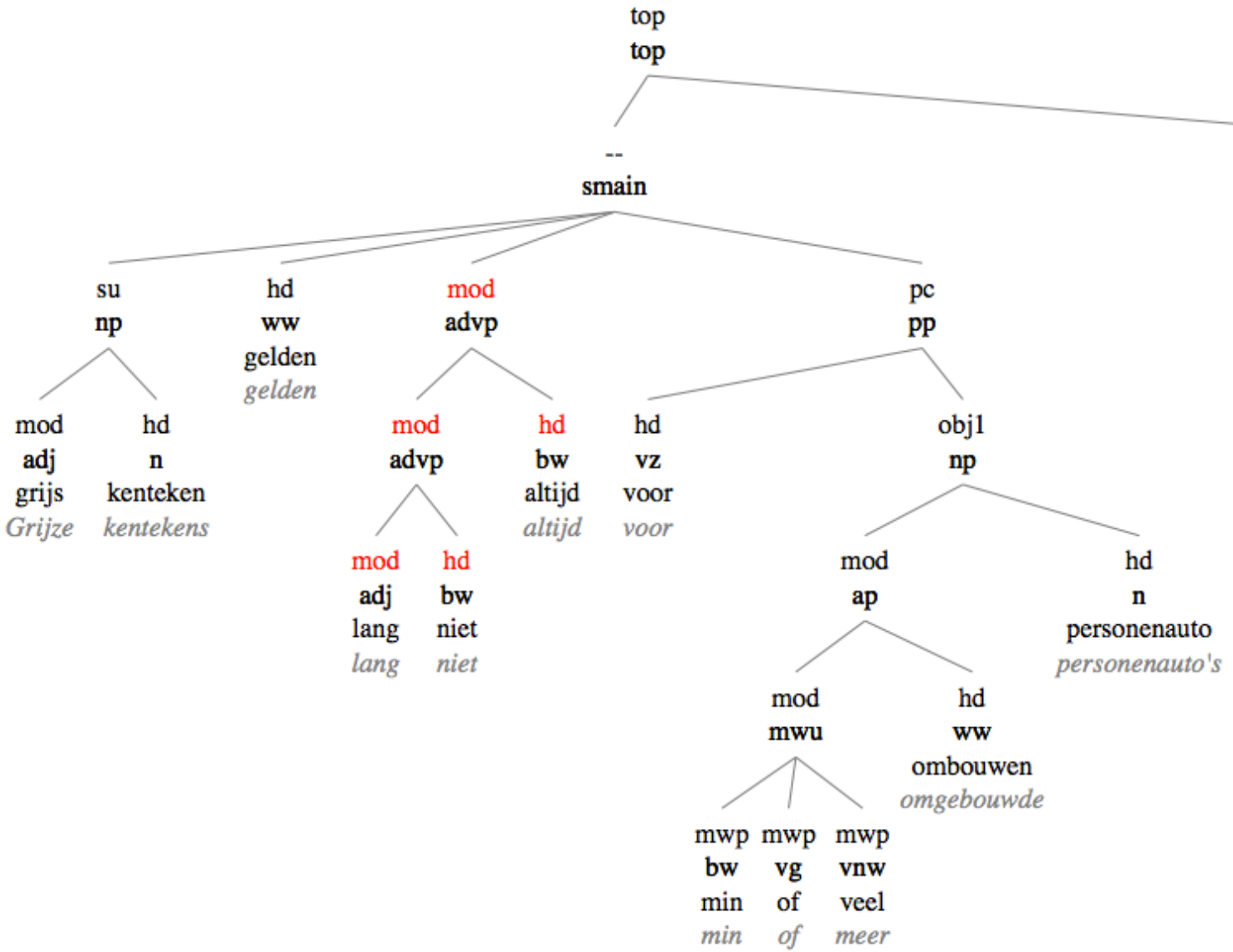
ADVP with a modifying ADVP embedded

→ 14 matches (in 65K sentences, 1M words)

RESULTS: data

SENTENCE ID	MATCHING SENTENCE	HITS	DISPLAY OPTIONS
WS-U-E-A-0000000229.p.10.s.1	Grijze kentekens gelden **lang niet altijd** voor min of meer omgebouwde personenauto's .	1	[full screen] [XML]
dpc-vla-001175-nl-sen.p.212.s.2	Deze publicaties verschijnen niet alleen in ' papieren ' vorm , maar **meestal ook elektronisch** .	1	[full screen] [XML]
WS-U-E-A-0000000206.p.34.s.3	**Niet alleen rechtdoor** , er waren zelfs hindernissen ingebouwd .	1	[full screen] [XML]
WR-P-P-H-0000000046.p.11.s.2	**Niet alleen vandaag** , maar permanent .	1	[full screen] [XML]
WR-P-P-I-0000000182.p.17.s.1	Het gaat daarbij **niet langer alleen** , of zelfs in de eerste plaats , om de recente lichte PCB-besmetting bij Hanekop , maar ook om situaties uit het verleden .	1	[full screen] [XML]
WR-P-P-I-0000000248.p.6.s.2	De partij heeft **nog altijd niet** beslist wie haar kandidaat voor het kanselierschap zal zijn .	1	[full screen] [XML]
WR-P-P-I-0000000037.p.1.s.3	Maar het is **nog altijd niet** duidelijk wat de oorzaak van het verschijnsel is .	1	[full screen] [XML]
WR-P-P-I-0000000182.p.24.s.3	In dezelfde sfeer situeer ik de vaststelling dat het voedselagentschap **nog altijd niet** functioneel is .	1	[full screen] [XML]

KU LEUVEN



RESULTS: trees

KU LEUVEN

OUTLINE

- GrETEL in a nutshell
- **GrInd: indexing the database**
- Conclusions and future work

KU LEUVEN

GrETEL 2.0

- **Goal**
  scaling up the query engine to a 500M word treebank

- **How?**
  Indexing system based on syntactic patterns (subtrees)
  = **GrETEL Indexing (GrInd)**

KU LEUVEN

GrETEL 2.0

- **Goal**
  scaling up the query engine to a 500M word treebank

- **How?**
  Indexing system based on syntactic patterns (subtrees)
  = **GrETEL Indexing (GrInd)**

  1) Preprocessing the data
  2) Querying the data

KU LEUVEN

GrInd: Preprocessing

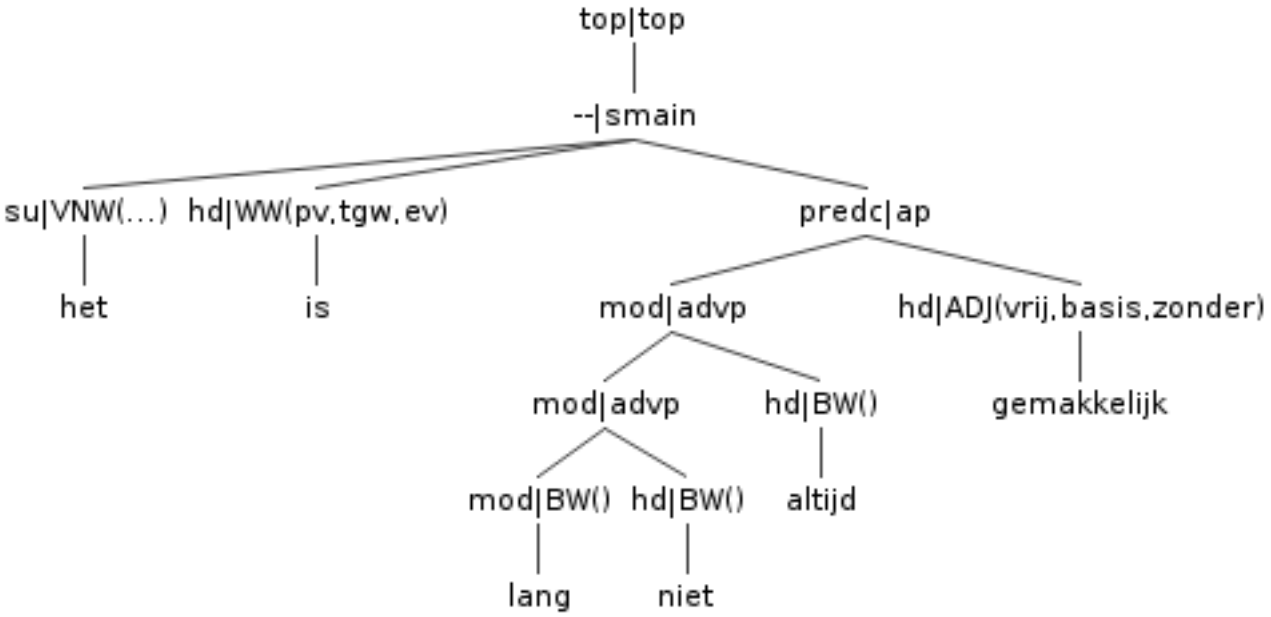
- **Step 1**
  For every node in the parse tree:
  - take all possible subtrees
  - with the node as root

- **Step 2**
  For each subtree:
  - Take top node + all children (D=1)
  - Put them into a database

GrInd: Preprocessing

For every node in the parse tree:

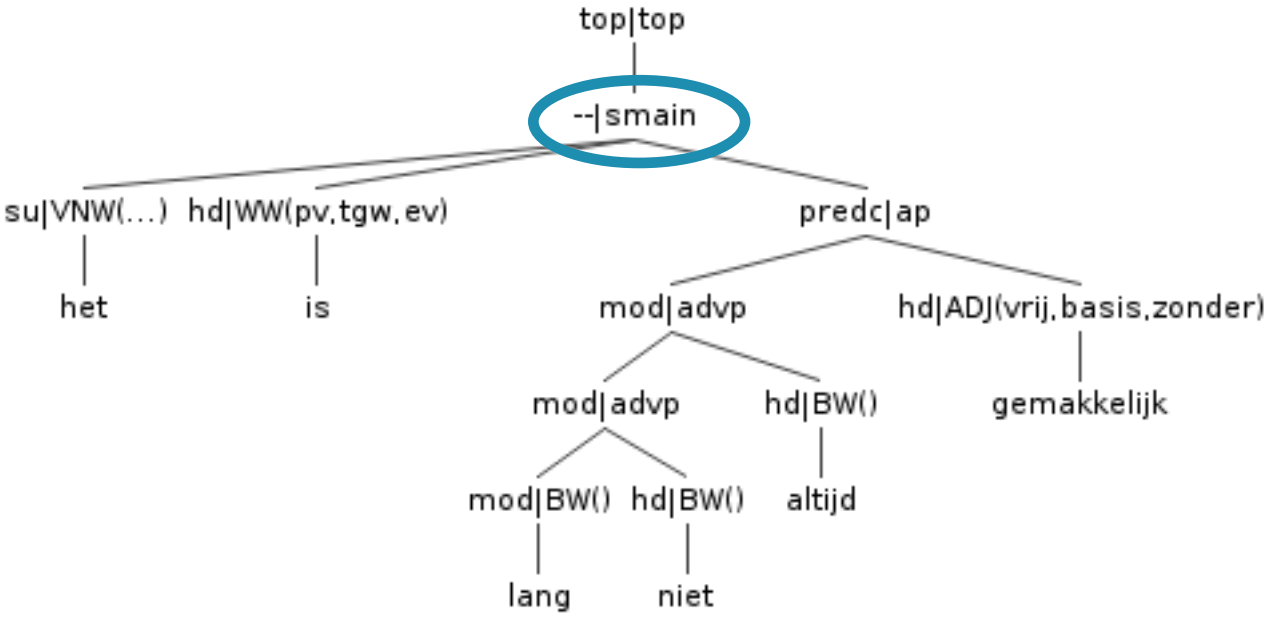
- take all possible subtrees
- with the node as root



GrInd: Preprocessing

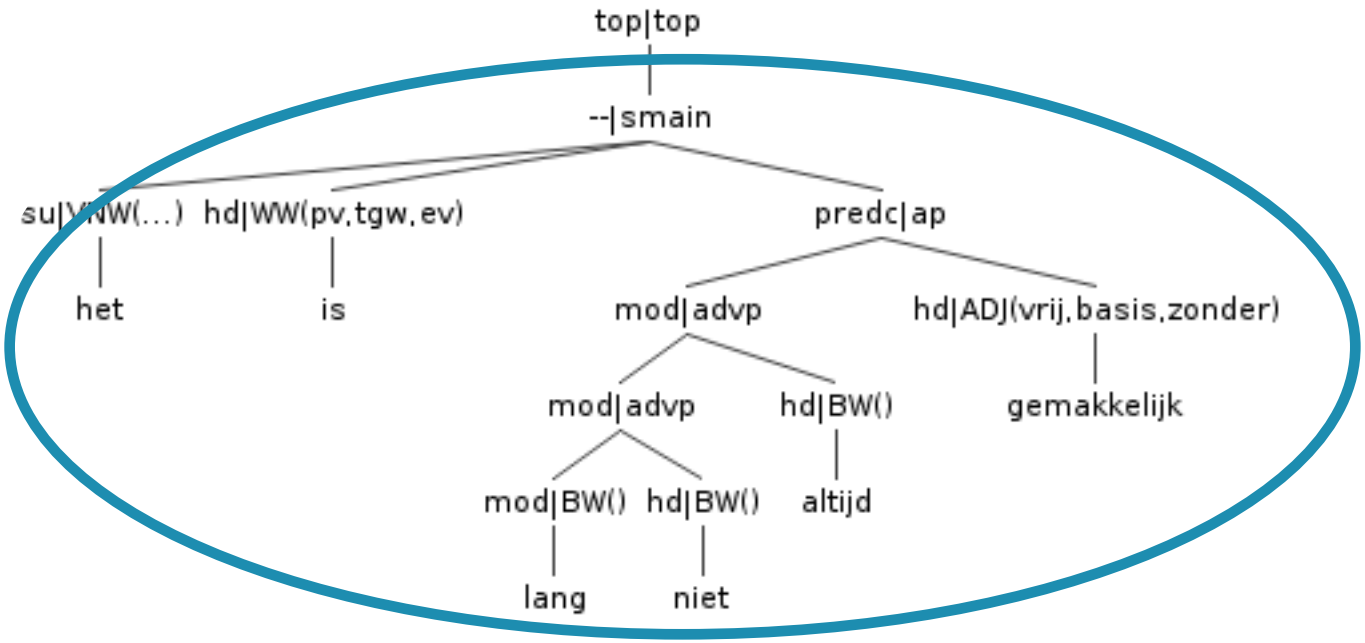
For every node in the parse tree:

- take all possible subtrees
- with the node as root



GrInd: Preprocessing

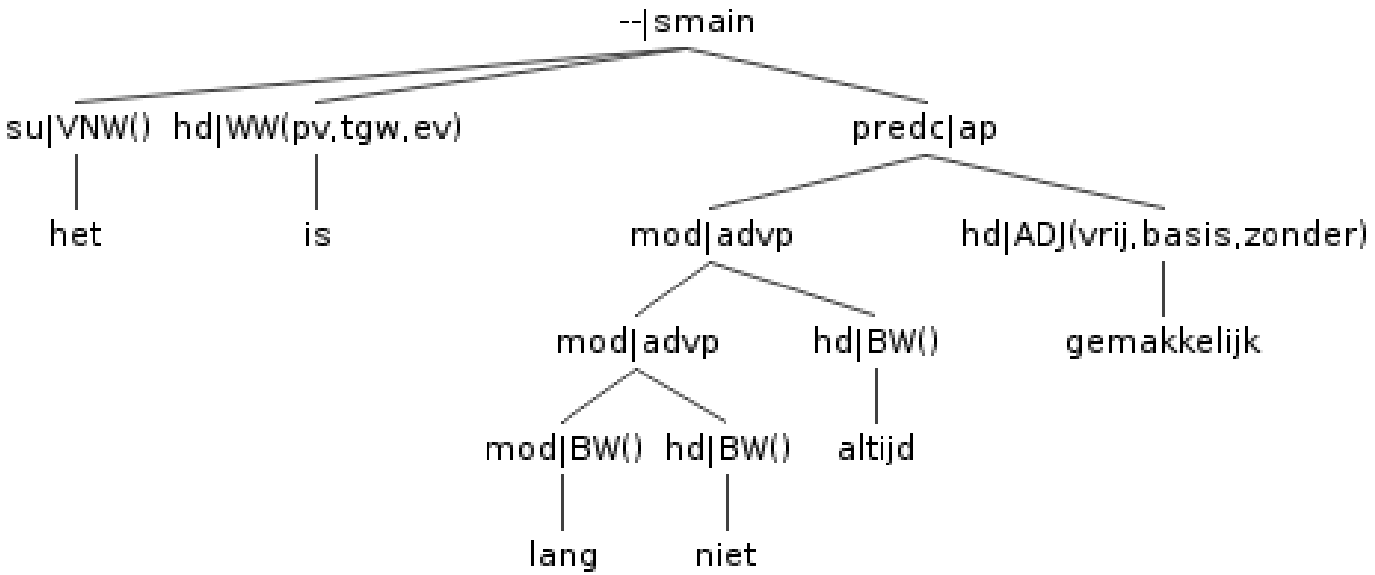
Subtree 1



KU LEUVEN

GrInd: Preprocessing

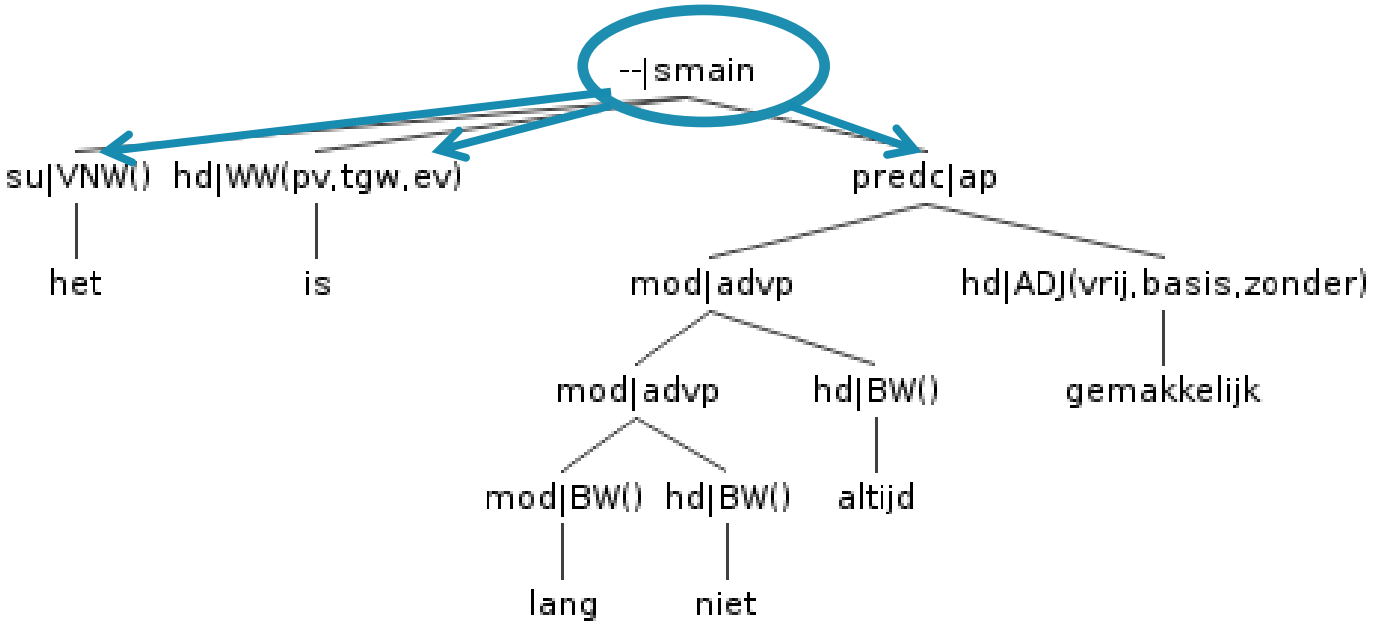
Subtree 1



KU LEUVEN

GrInd: Preprocessing

Subtree 1

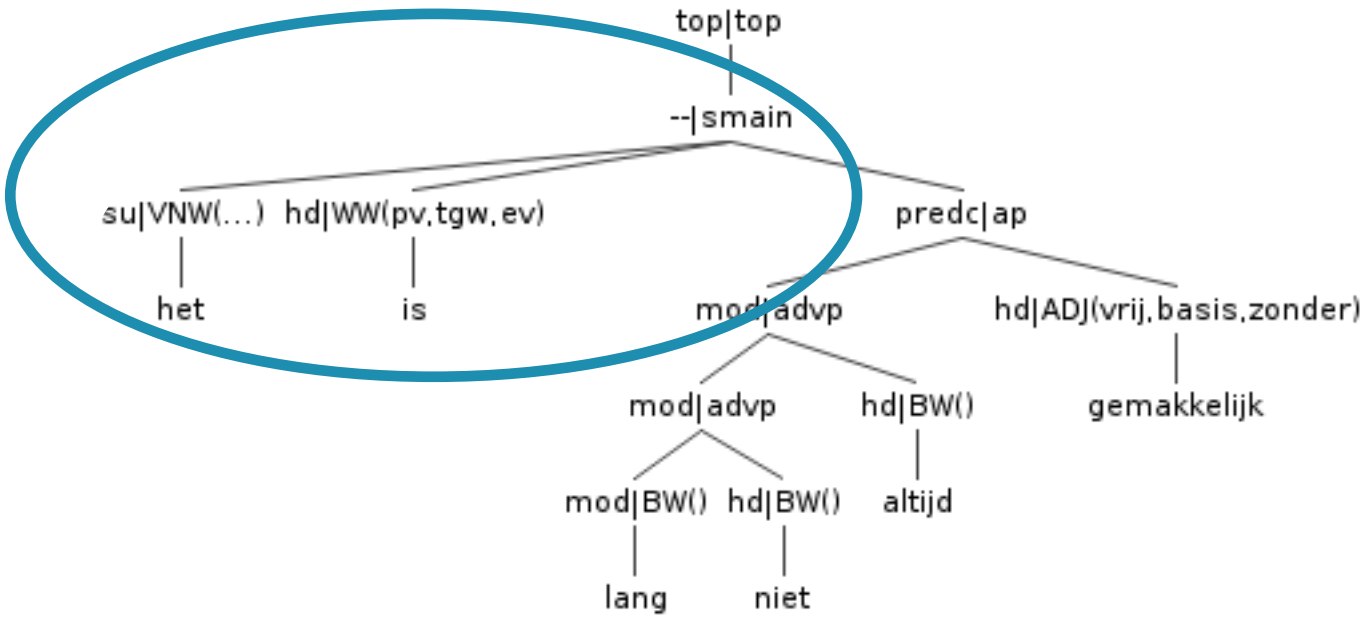


- Take top node + all children
- Put them into a database:
  SMAIN hd%ww_predc%ap_su%vnw

KU LEUVEN

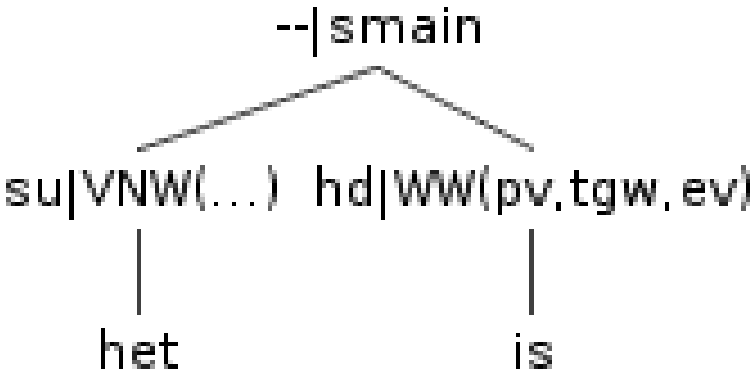
GrInd: Preprocessing

Subtree 2



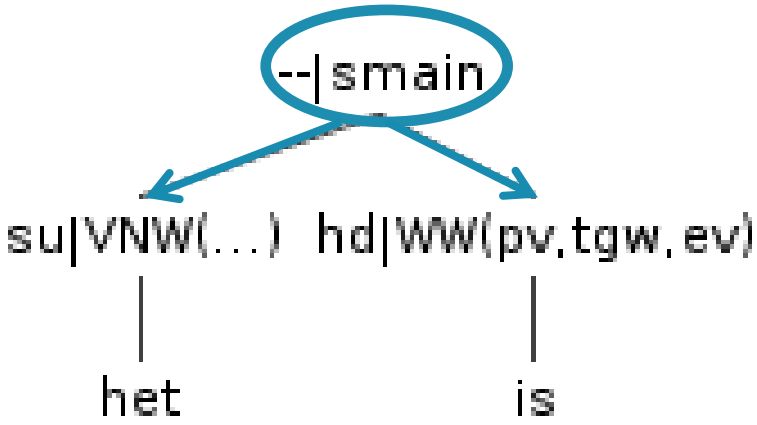
GrInd: Preprocessing

Subtree 2



GrInd: Preprocessing

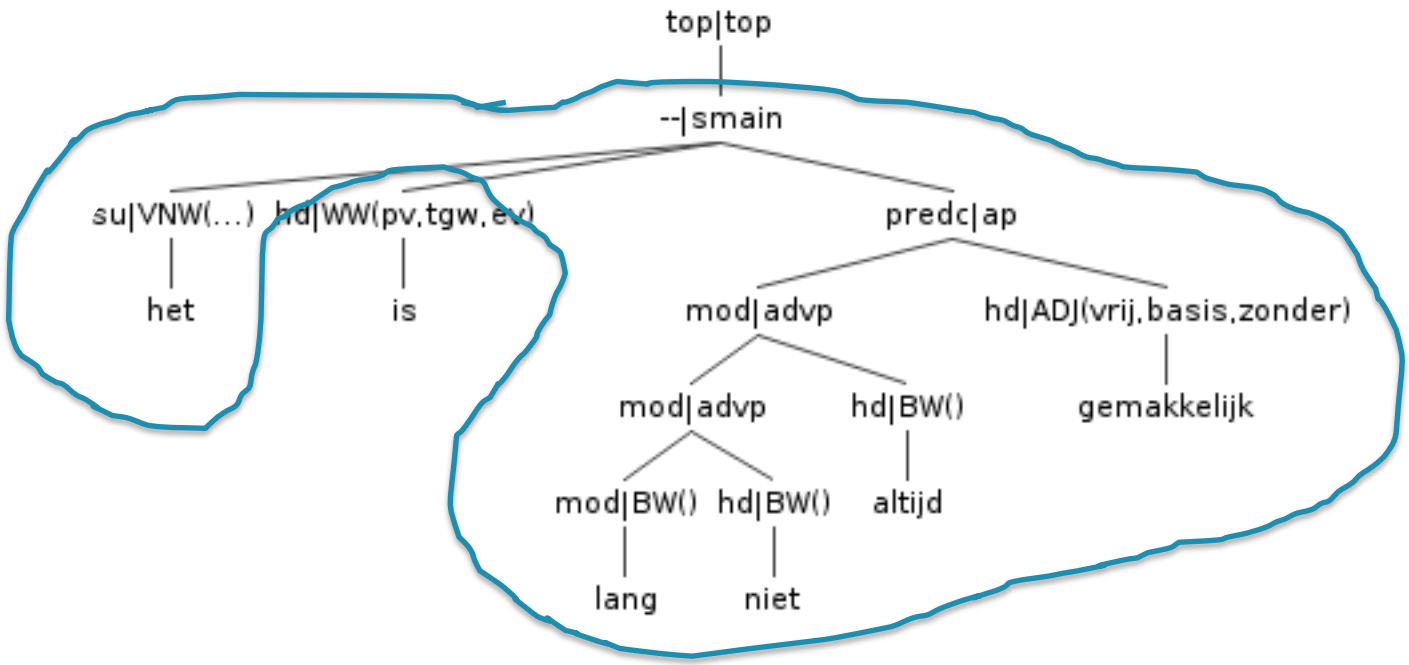
Subtree 2



- Take top node + all children
- Put them into a database:
  SMAIN hd%ww_su%vnw

GrInd: Preprocessing

Subtree 3

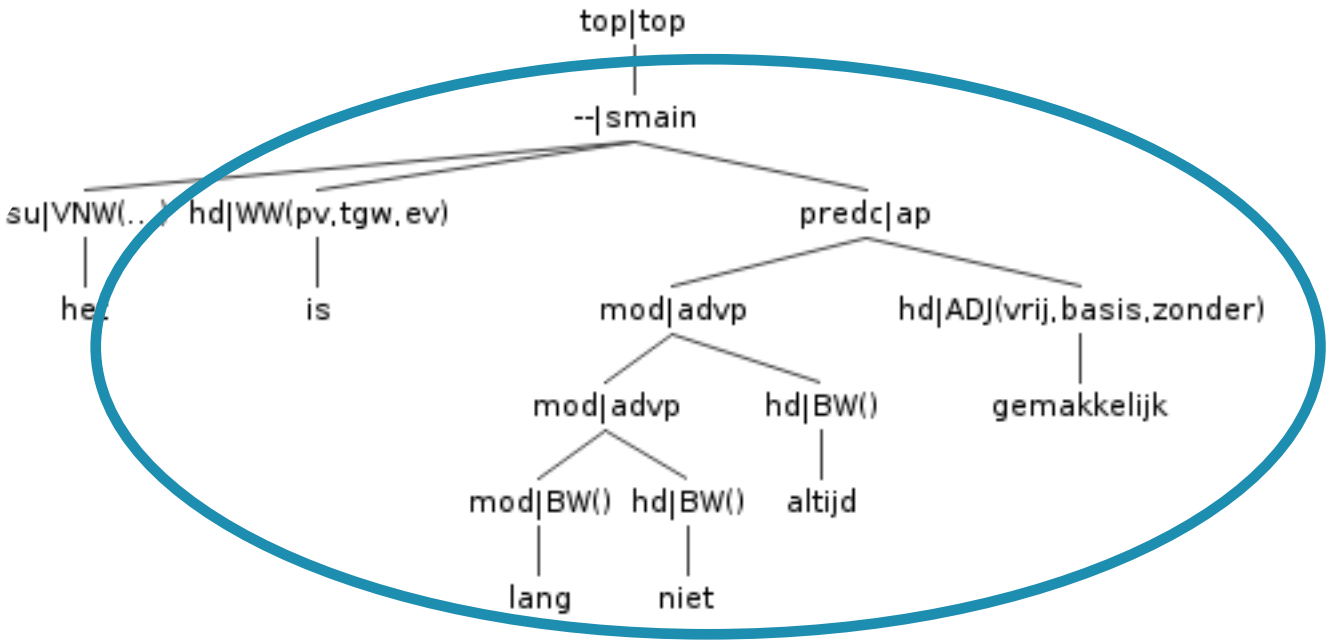


SMAIN su%vnw_predc%ap

KU LEUVEN

GrInd: Preprocessing

Subtree 4

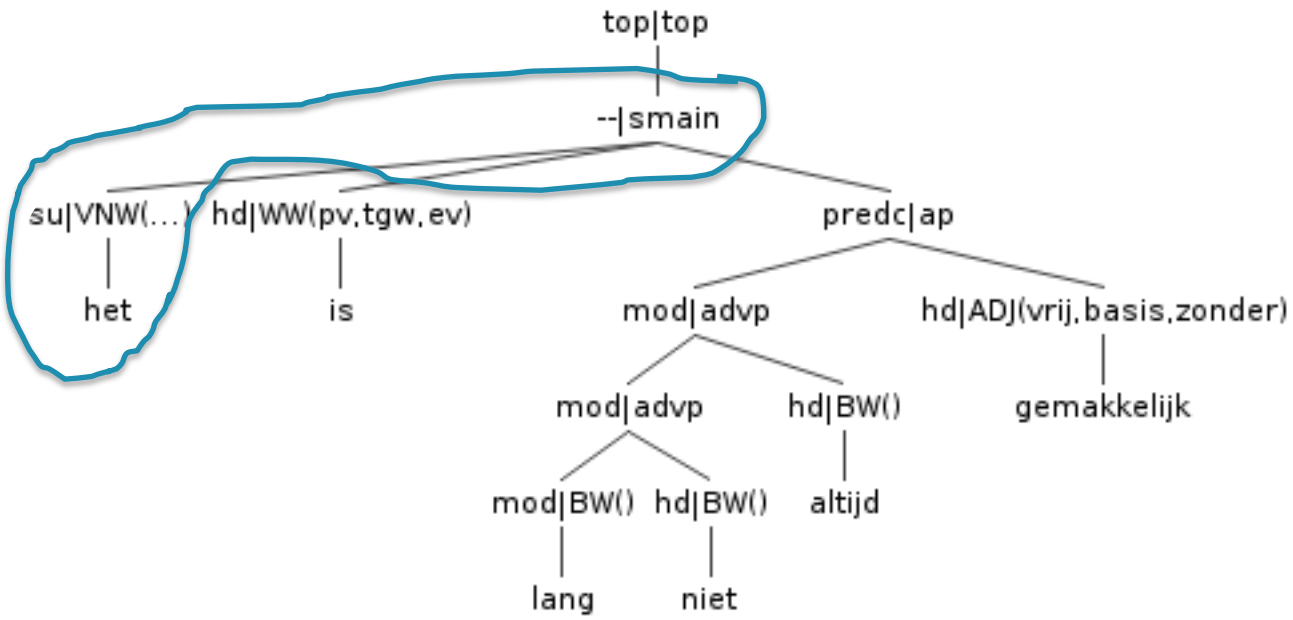


SMAIN hd%ww_predc%ap

KU LEUVEN

GrInd: Preprocessing

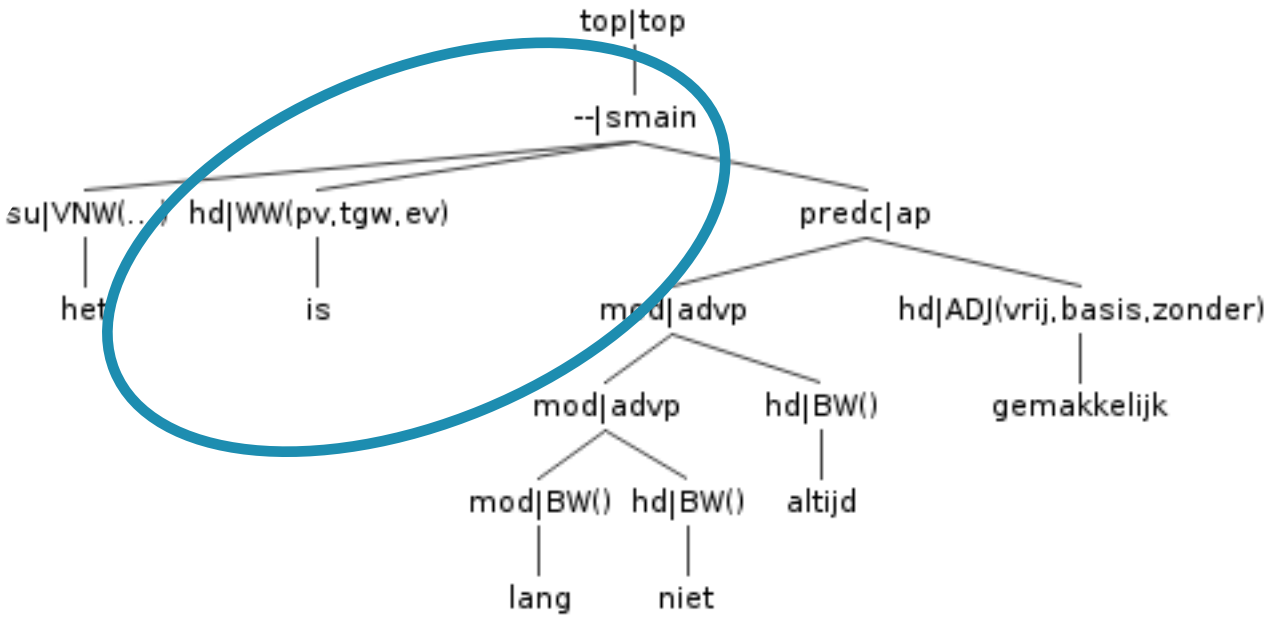
Subtree 5



SMAIN su%vnw

GrInd: Preprocessing

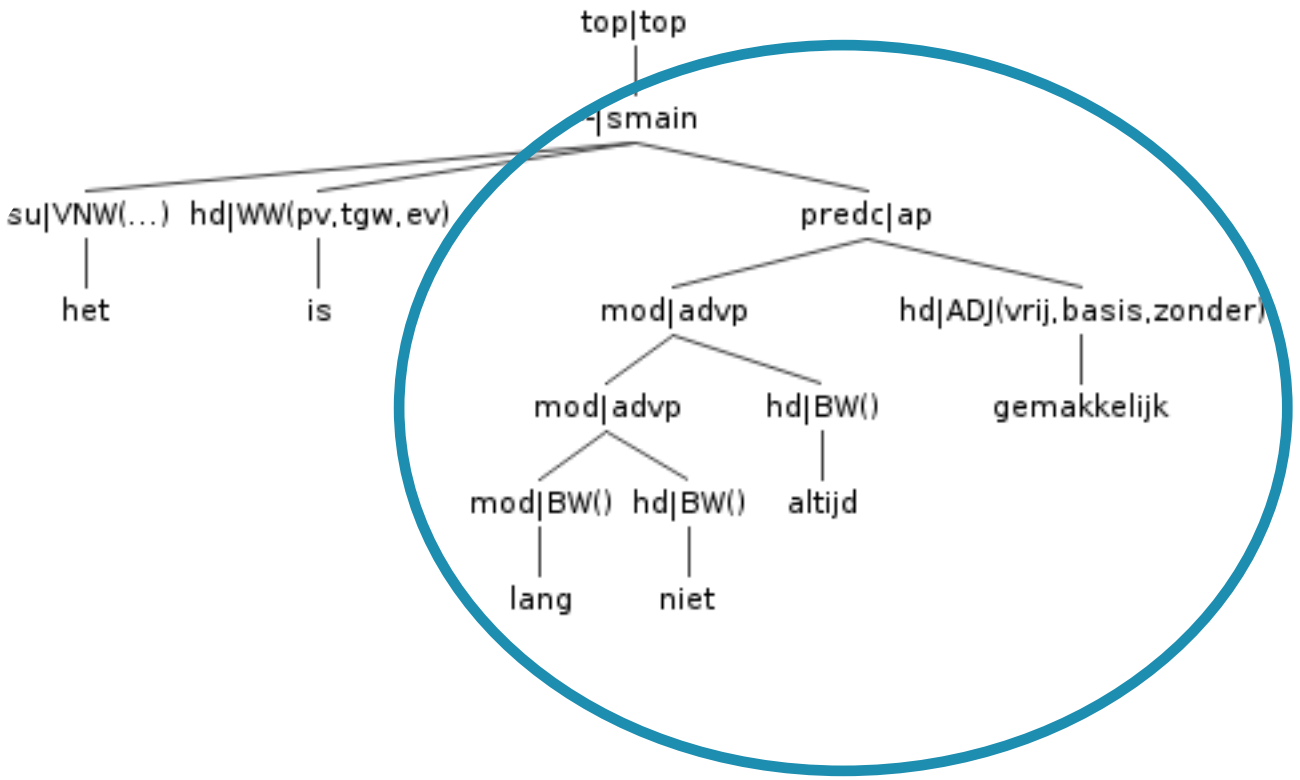
Subtree 6



SMAIN hd%ww

GrInd: Preprocessing

Subtree 7



SMAIN predc%ap

KU LEUVEN

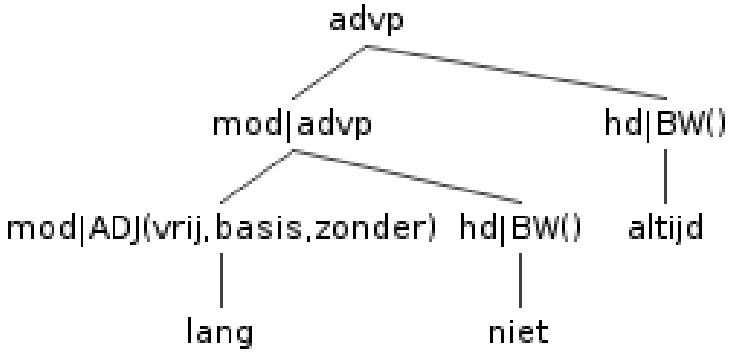
GrInd: Preprocessing

- **Step 1**
  For every node in the parse tree:
  - take all possible subtrees
  - with the node as root

- **Step 2**
  For each subtree:
  - Take top node + all children (D=1)
  - **Put them into a database**

GrInd: Preprocessing

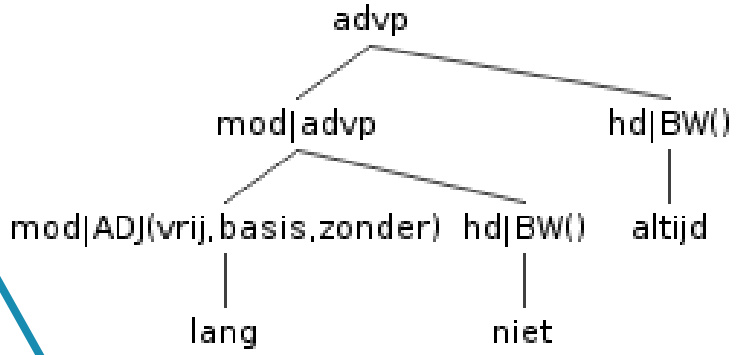
```
<treebank component="WRPEF" cat="advp"
file="mod%advp_hd%bw">
 <tree id="WR-P-E-F-0000000769.p.4.s.7" >
  <node begin="3" cat="advp" end="6" id="6"
rel="mod">
   <node begin="3" cat="advp" end="5" id="7"
rel="mod">
    <node begin="3" buiging="zonder" end="4"
frame="adverb" graad="basis" id="8" lcat="advp"
lemma="lang" pos="adv" positie="vrij"
postag="ADJ(vrij,basis,zonder)" pt="adj" rel="mod"
root="lang" sense="lang" word="lang"/>
    <node begin="4" end="5" frame="adverb" id="9"
lcat="advp" lemma="niet" pos="adv" postag="BW()"
pt="bw" rel="hd" root="niet" sense="niet"
word="niet"/>
   </node>
   <node begin="5" end="6" frame="adverb" id="10"
lcat="advp" lemma="altijd" pos="adv" postag="BW()"
pt="bw" rel="hd" root="altijd" sense="altijd"
word="altijd"/>
  </node>
 </tree>
</treebank>
```



KU LEUVEN

GrInd: Preprocessing

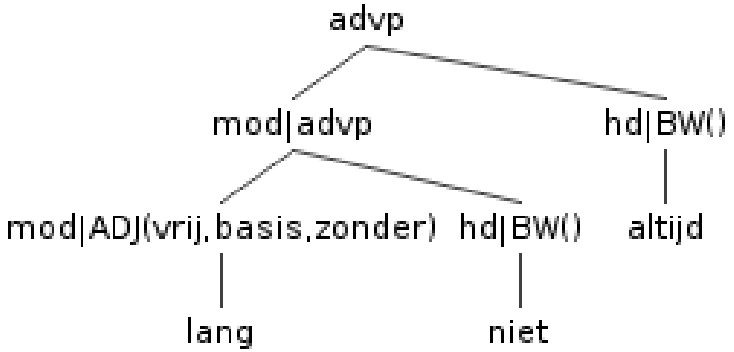
```
<treebank component="WRPEF" cat="advp"
file="mod%advp_hd%bw">
 <tree id="WR-P-E-F-0000000769.p.4.s.7" >
  <node begin="3" cat="advp" end="6" id="6"
rel="mod">
   <node begin="3" cat="advp" end="5" id="7"
rel="mod">
    <node begin="3" buiging="zonder" end="4"
frame="adverb" graad="basis" id="8" lcat="advp"
lemma="lang" pos="adv" positie="vrij"
postag="ADJ(vrij,basis,zonder)" pt="adj" rel="mod"
root="lang" sense="lang" word="lang"/>
    <node begin="4" end="5" frame="adverb" id="9"
lcat="advp" lemma="niet" pos="adv" postag="BW()"
pt="bw" rel="hd" root="niet" sense="niet"
word="niet"/>
   </node>
   <node begin="5" end="6" frame="adverb" id="10"
lcat="advp" lemma="altijd" pos="adv" postag="BW()"
pt="bw" rel="hd" root="altijd" sense="altijd"
word="altijd"/>
  </node>
 </tree>
</treebank>
```



Breadth-first pattern:
hd%bw_mod%advp

GrInd: Preprocessing

```
<treebank component="WRPEF" cat="advp"
file="mod%advp_hd%bw">
 <tree id="WR-P-E-F-0000000769.p.4.s.7" >
  <node begin="3" cat="advp" end="6" id="6"
rel="mod">
   <node begin="3" cat="advp" end="5" id="7"
rel="mod">
    <node begin="3" buiging="zonder" end="4"
frame="adverb" graad="basis" id="8" lcat="advp"
lemma="lang" pos="adv" positie="vrij"
postag="ADJ(vrij,basis,zonder)" pt="adj" rel="mod"
root="lang" sense="lang" word="lang"/>
    <node begin="4" end="5" frame="adverb" id="9"
lcat="advp" lemma="niet" pos="adv" postag="BW()"
pt="bw" rel="hd" root="niet" sense="niet"
word="niet"/>
   </node>
   <node begin="5" end="6" frame="adverb" id="10"
lcat="advp" lemma="altijd" pos="adv" postag="BW()"
pt="bw" rel="hd" root="altijd" sense="altijd"
word="altijd"/>
  </node>
 </tree>
</treebank>
```

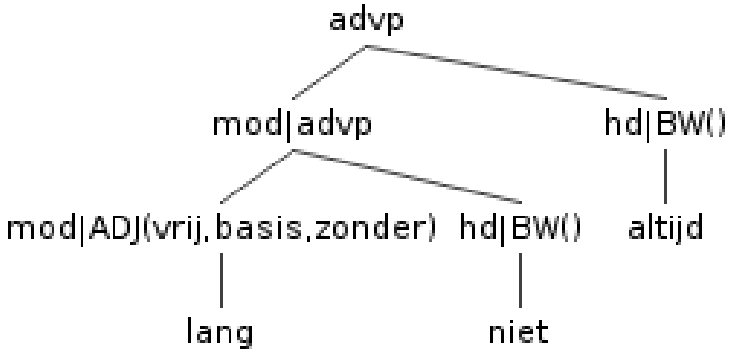


- Breadth-first pattern: hd%bw_mod%advp
- Combined with root node: advp

KU LEUVEN

GrInd: Preprocessing

```
<treebank component="WRPEF" cat="advp"
file="mod%advp_hd%bw">
 <tree id="WR-P-E-F-0000000769.p.4.s.7" >
  <node begin="3" cat="advp" end="6" id="6"
rel="mod">
   <node begin="3" cat="advp" end="5" id="7"
rel="mod">
    <node begin="3" buiging="zonder" end="4"
frame="adverb" graad="basis" id="8" lcat="advp"
lemma="lang" pos="adv" positie="vrij"
postag="ADJ(vrij,basis,zonder)" pt="adj" rel="mod"
root="lang" sense="lang" word="lang"/>
    <node begin="4" end="5" frame="adverb" id="9"
lcat="advp" lemma="niet" pos="adv" postag="BW()"
pt="bw" rel="hd" root="niet" sense="niet"
word="niet"/>
   </node>
   <node begin="5" end="6" frame="adverb" id="10"
lcat="advp" lemma="altijd" pos="adv" postag="BW()"
pt="bw" rel="hd" root="altijd" sense="altijd"
word="altijd"/>
  </node>
 </tree>
</treebank>
```

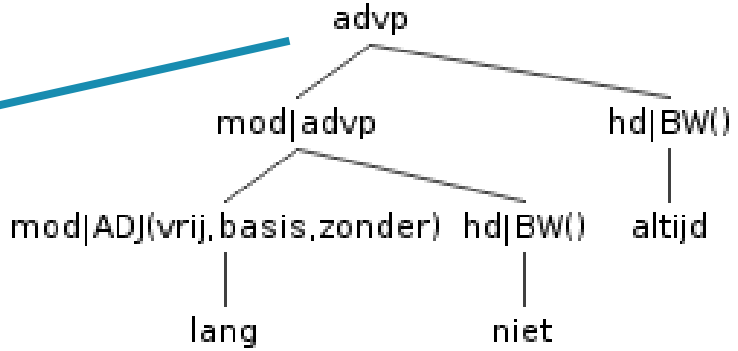


- Breadth-first pattern: hd%bw_mod%advp
- Combined with root node: advp
- Per corpus component: WRPEF

KU LEUVEN

GrInd: Preprocessing

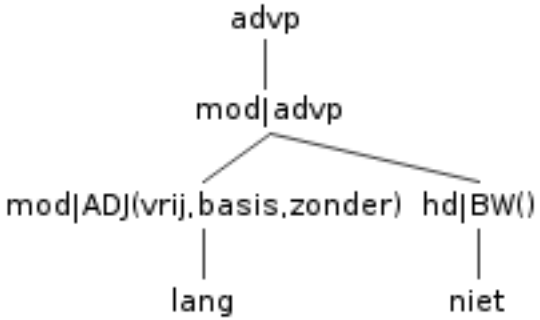
```
<treebank component="WRPEF" cat="advp"
file="mod%advp_hd%bw">
 <tree id="WR-P-E-F-0000000769.p.4.s.7" >
  <node begin="3" cat="advp" end="6" id="6"
rel="mod">
   <node begin="3" cat="advp" end="5" id="7"
rel="mod">
    <node begin="3" buiging="zonder" end="4"
frame="adverb" graad="basis" id="8" lcat="advp"
lemma="lang" pos="adv" positie="vrij"
postag="ADJ(vrij,basis,zonder)" pt="adj" rel="mod"
root="lang" sense="lang" word="lang"/>
    <node begin="4" end="5" frame="adverb" id="9"
lcat="advp" lemma="niet" pos="adv" postag="BW()"
pt="bw" rel="hd" root="niet" sense="niet"
word="niet"/>
   </node>
   <node begin="5" end="6" frame="adverb" id="10"
lcat="advp" lemma="altijd" pos="adv" postag="BW()"
pt="bw" rel="hd" root="altijd" sense="altijd"
word="altijd"/>
  </node>
 </tree>
</treebank>
```



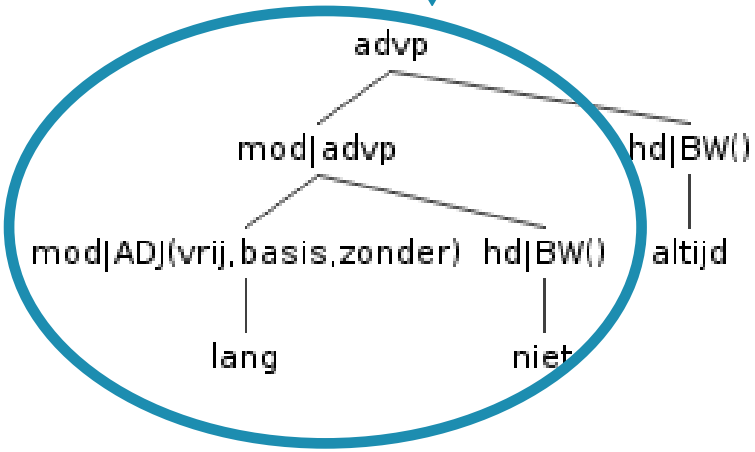
- Link to the original tree in the treebank

KU LEUVEN

GrInd: Preprocessing



included in



- <include> tags to avoid copying information
- more general patterns are included in more specific patterns

```
<treebank component="WRPEF" cat="advp"
file="mod%advp">
  <include file="WRPEFadvpmod%advp_hd%bw" />
</treebank>
```

KU LEUVEN

GrInd: Preprocessing

Component	Contents	Trees	GrInded DBs
WR-P-E-A	Discussion lists	4 396 361	3 686 409
WR-P-E-C	E-magazines	551 343	716 491
WR-P-E-E	Newsletters	115	2 283
WR-P-E-F	Press releases	18 373	72 285
WR-P-E-G	Subtitles	3 925 834	699 117
WR-P-E-H	Teletext pages	40 715	76 989
WR-P-E-I	Websites	205 037	253 921
WR-P-E-J	Wikipedia	1 355 061	1 154 753
…	…	…	…
SoNaR	**Complete treebank**	**40 384 789**	**17 389 801**

GrETEL 2.0

- **Goal**
  scaling up the query engine to a 500M word treebank

- **How?**
  Indexing system based on syntactic patterns (subtrees)
  = **GrETEL Indexing (GrInd)**

  1) Preprocessing the data
  2) Querying the data

KU LEUVEN

GrInd: Querying

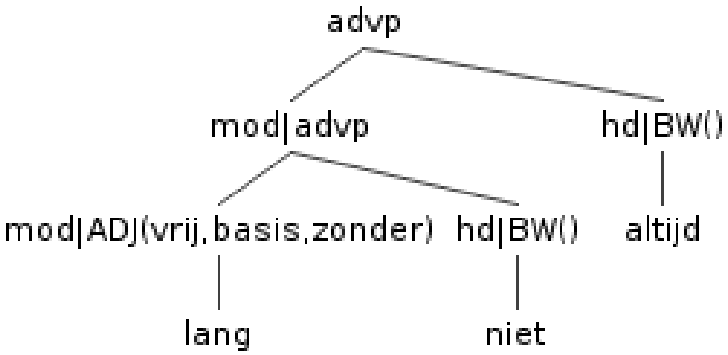
After preprocessing steps:

Put GrInded DBs into BaseX XML database
(Holupirek & Scholl 2008)

= XPath/XQuery engine

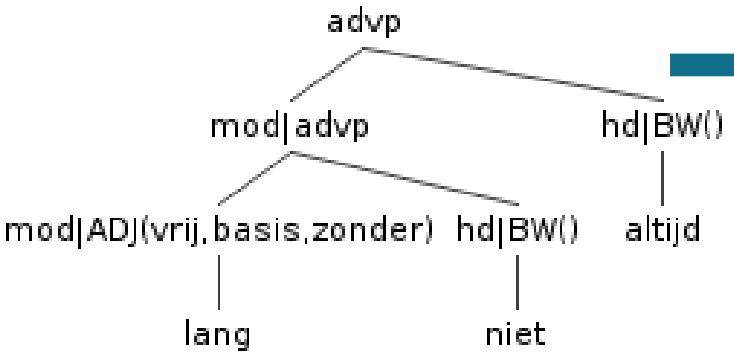
KU LEUVEN

GrInd: Querying



KU LEUVEN

GrInd: Querying



Converted to BF

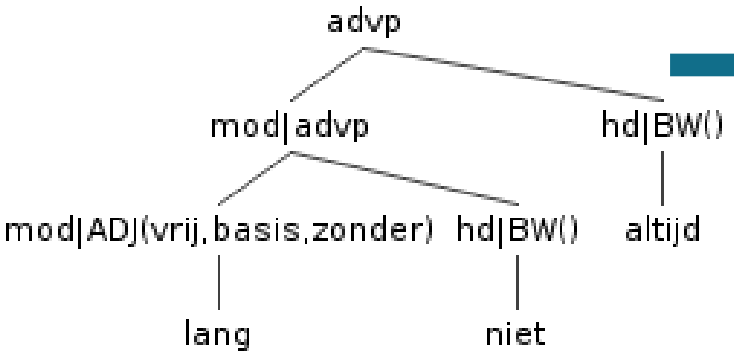
ADVP mod%advp_hd%bw

Converted into XPath

```
//node[@cat="advp" and node[@rel=" and
@cat="advp" and node[@rel="mod @pt="bw"
and @lemma="lang"] and node[@rel="hd"
and @pt="bw" and @lemma="niet"]] and
node[@rel="hd" @pt="bw" and
@lemma="altijd"]]
```

KU LEUVEN

GrInd: Querying



Converted to BF

ADVP mod%advp_hd%bw

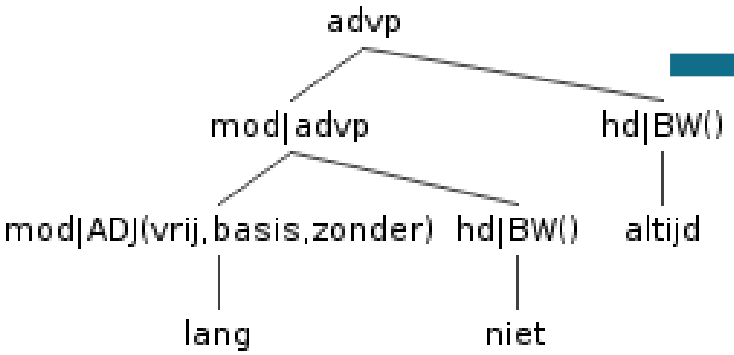
XPath applied to GrInded DB

Converted into XPath

```
//node[@cat="advp" and node[@rel=" and
@cat="advp" and node[@rel="mod @pt="bw"
and @lemma="lang"] and node[@rel="hd"
and @pt="bw" and @lemma="niet"]] and
node[@rel="hd" @pt="bw" and
@lemma="altijd"]]
```

KU LEUVEN

GrInd: Querying



Converted to BF

ADVP mod%advp_hd%bw

XPath applied to GrInded DB

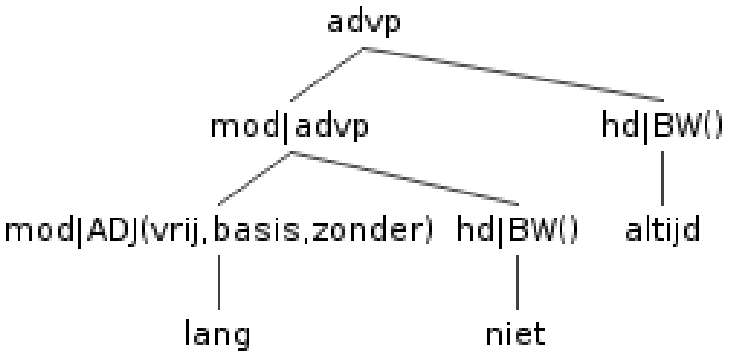
Converted into XPath

```
//node[@cat="advp" and node[@rel=" and
@cat="advp" and node[@rel="mod @pt="bw"
and @lemma="lang"] and node[@rel="hd"
and @pt="bw" and @lemma="niet"]] and
node[@rel="hd" @pt="bw" and
@lemma="altijd"]]
```

Results: set of similar sentences

KU LEUVEN

GrInd: Querying



GrETEL 1.0

XPath applied to complete treebank

Converted into XPath

```
//node[@cat="advp" and node[@rel=" and
@cat="advp" and node[@rel="mod @pt="bw"
and @lemma="lang"] and node[@rel="hd"
and @pt="bw" and @lemma="niet"]] and
node[@rel="hd" @pt="bw" and
@lemma="altijd"]]
```

Results: set of similar sentences

KU LEUVEN

OUTLINE

- GrETEL in a nutshell
- GrInd: indexing the database
- **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

CONCLUSIONS

- **GrETEL 2.0** 500M SoNaR treebank
- Query treebank in *reasonable* time
- GrInd: index based on syntactic patterns
  - Works well for specific language patterns
  - Needs improvement for general patterns
    → too many <include> tags slow down querying

KU LEUVEN

FUTURE WORK

- **Benchmarking**
- **XPath search**
  - XPath2Tree
- **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!

Nederlandse Taalunie

CLARIN

KU LEUVEN