KU LEUVEN

Facilitating treebank mining of CGN with GrETEL

Liesbeth Augustinus
Vincent Vandeghinste
Frank Van Eynde

CLIN23 - January 18, 2013

GrETEL

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

- **Query treebanks by example**

- Previous versions (CLIN22, LREC 2012)
  => only for LASSY treebank

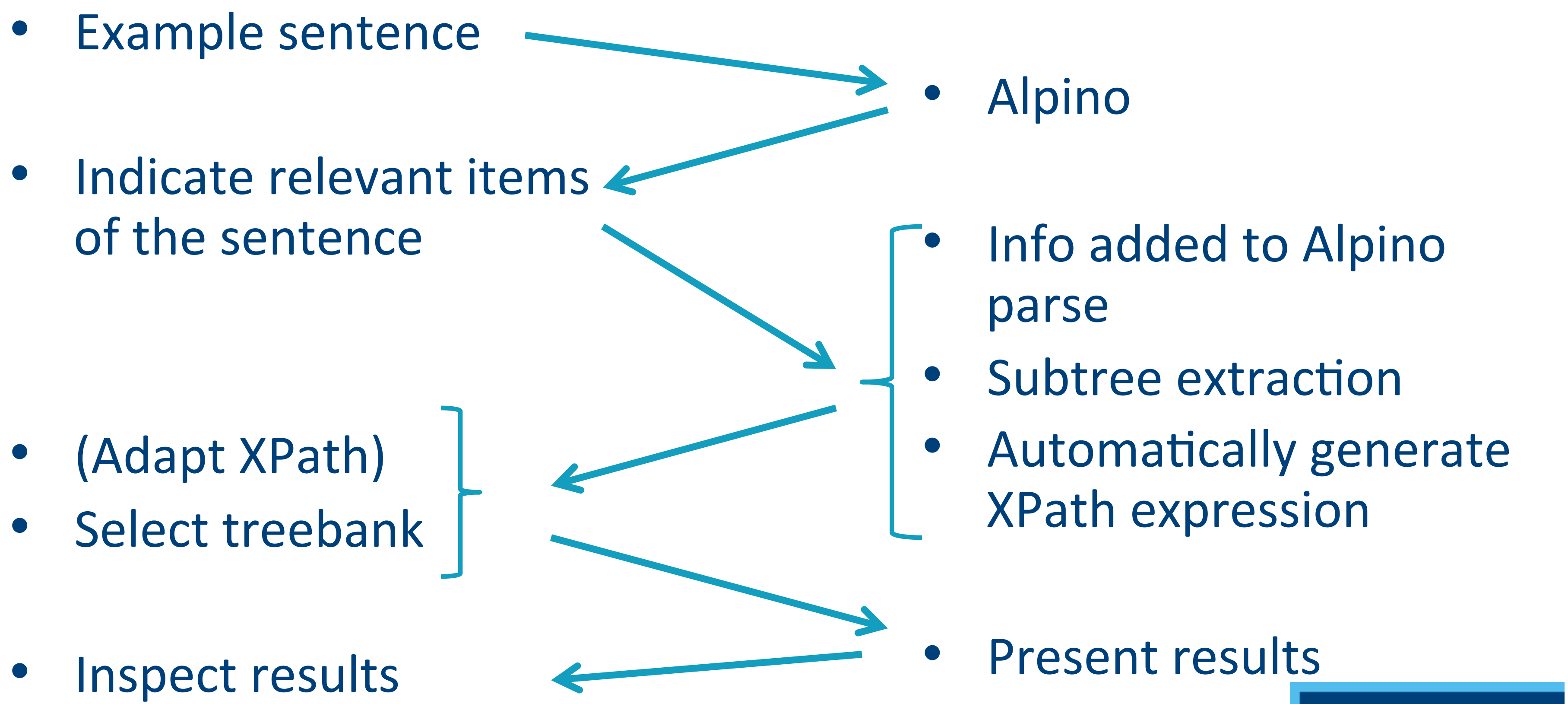
- New release
  => GrETEL for CGN treebank
  => update based on user reviews

CLARIN
Common Language Resources and Technology Infrastructure

KU LEUVEN

the user

GrETEL



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

- Alpino
- Info added to Alpino parse
- Subtree extraction
- Automatically generate XPath expression
- Present results

KU LEUVEN

OUTLINE

- GrETEL in a nutshell

- **What's new?**
  - CGN treebank
  - Search options

- Conclusions and future work

WHAT'S NEW?

1) Include CGN treebank

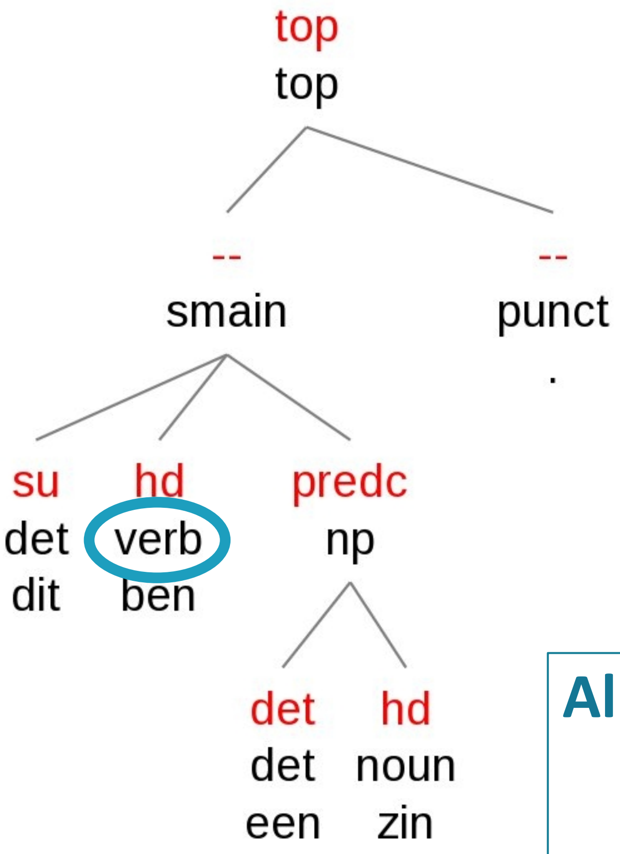
CGN core corpus
Spoken Dutch
Stylistic & regional differences
± 1M tokens
Manually corrected
TIGER > Alpino-like XML

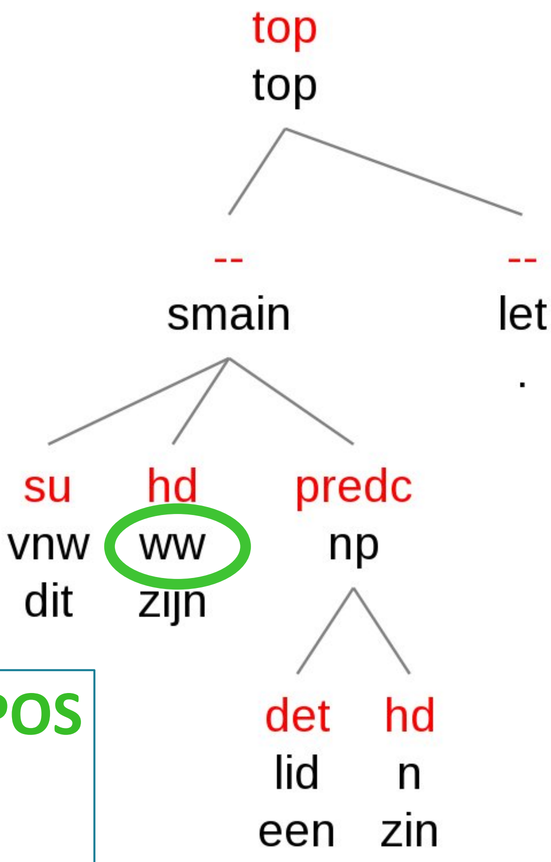
TREEBANKS

CGN core corpus	LASSY small
Spoken Dutch	Written Dutch
Stylistic & regional differences	Stylistic differences
± 1M tokens	± 1M tokens
Manually corrected	Manually corrected
TIGER > Alpino-like XML	**Alpino XML**

TREE STRUCTURES

Alpino tree vs CGN tree

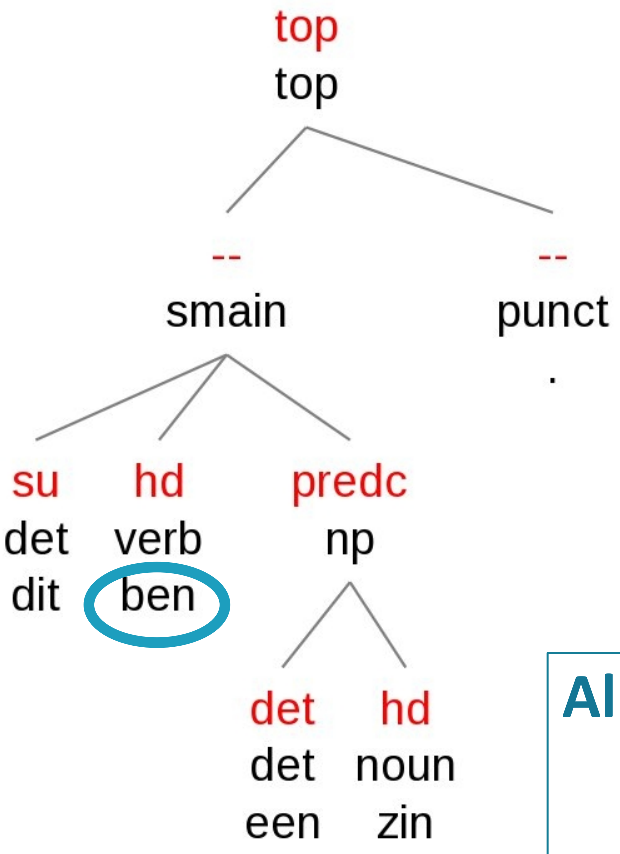




Alpino POS vs CGN POS

TREE STRUCTURES

Alpino tree vs CGN tree

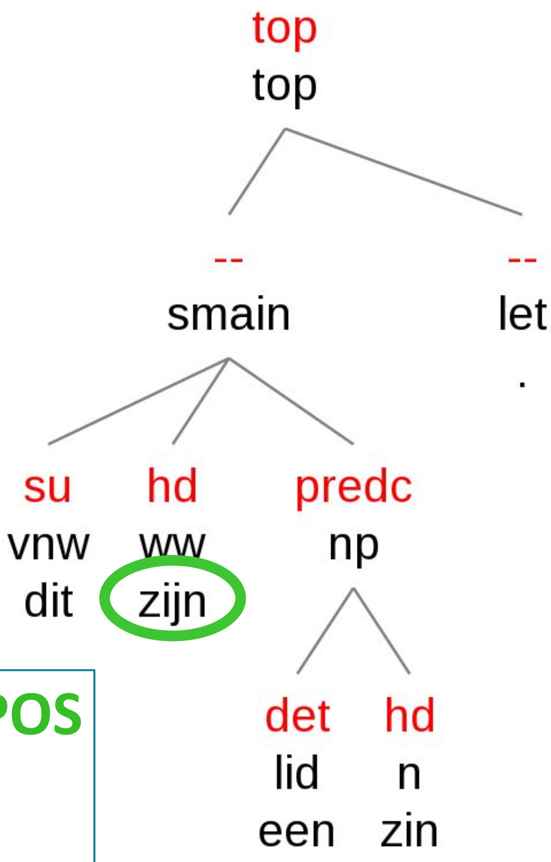


top
top

-- smain
-- punct .

su det dit
hd verb ben
predc np

det det een
hd noun zin



top
top

-- smain
-- let .

su vnw dit
hd ww zijn
predc np

det lid een
hd n zin

Alpino POS vs CGN POS

root vs lemma

TREE STRUCTURES

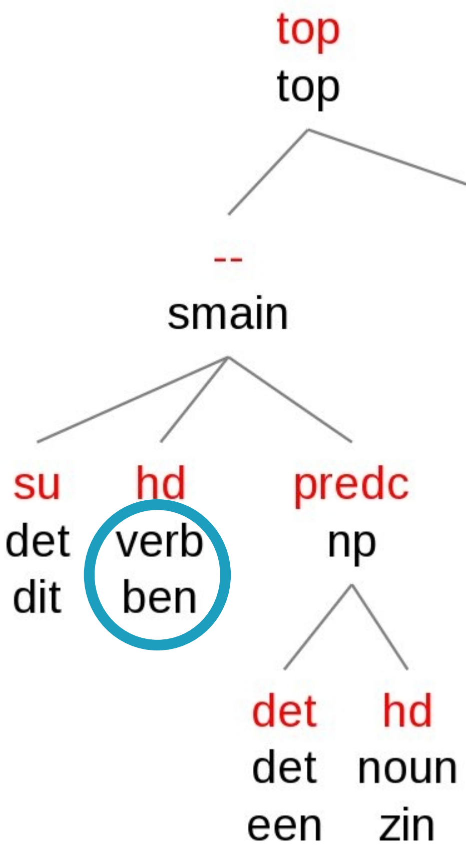
Add Alpino POS and roots to CGN treebank?

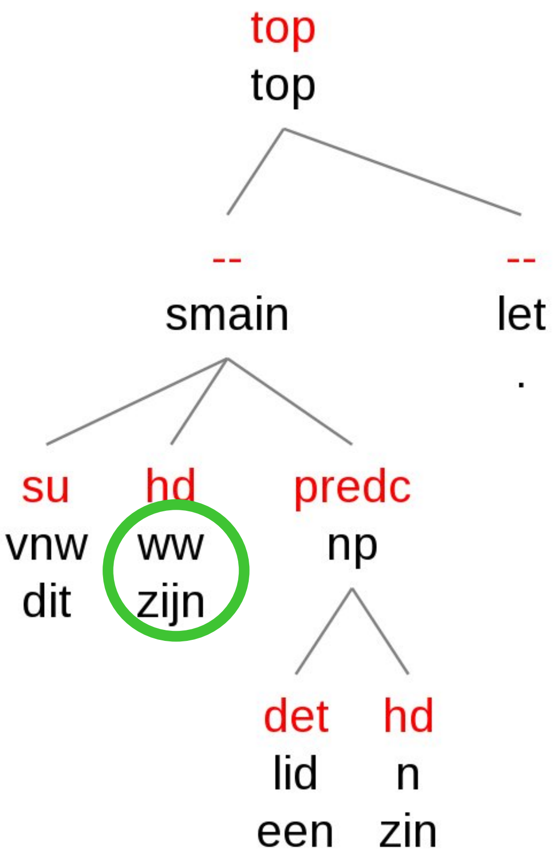
OR

Add CGN POS and lemmas to Alpino parse!

TREE STRUCTURES

Alpino tree vs CGN tree





LASSY tree

Alpino POS & CGN POS

root & lemma

TREE STRUCTURES

Add Alpino POS and roots to CGN treebank?

OR

Add CGN POS and lemmas to Alpino parse!

=> useful for querying both CGN and LASSY

TREE STRUCTURES

Add Alpino POS and roots to CGN treebank?

OR

Add CGN POS and lemmas to Alpino parse!

=> useful for querying both CGN and LASSY

WHAT'S NEW?

1) Include CGN treebank

↓

2) **More search options**

WHAT'S NEW?

1) Include CGN treebank

↓

2) NEW **search options**

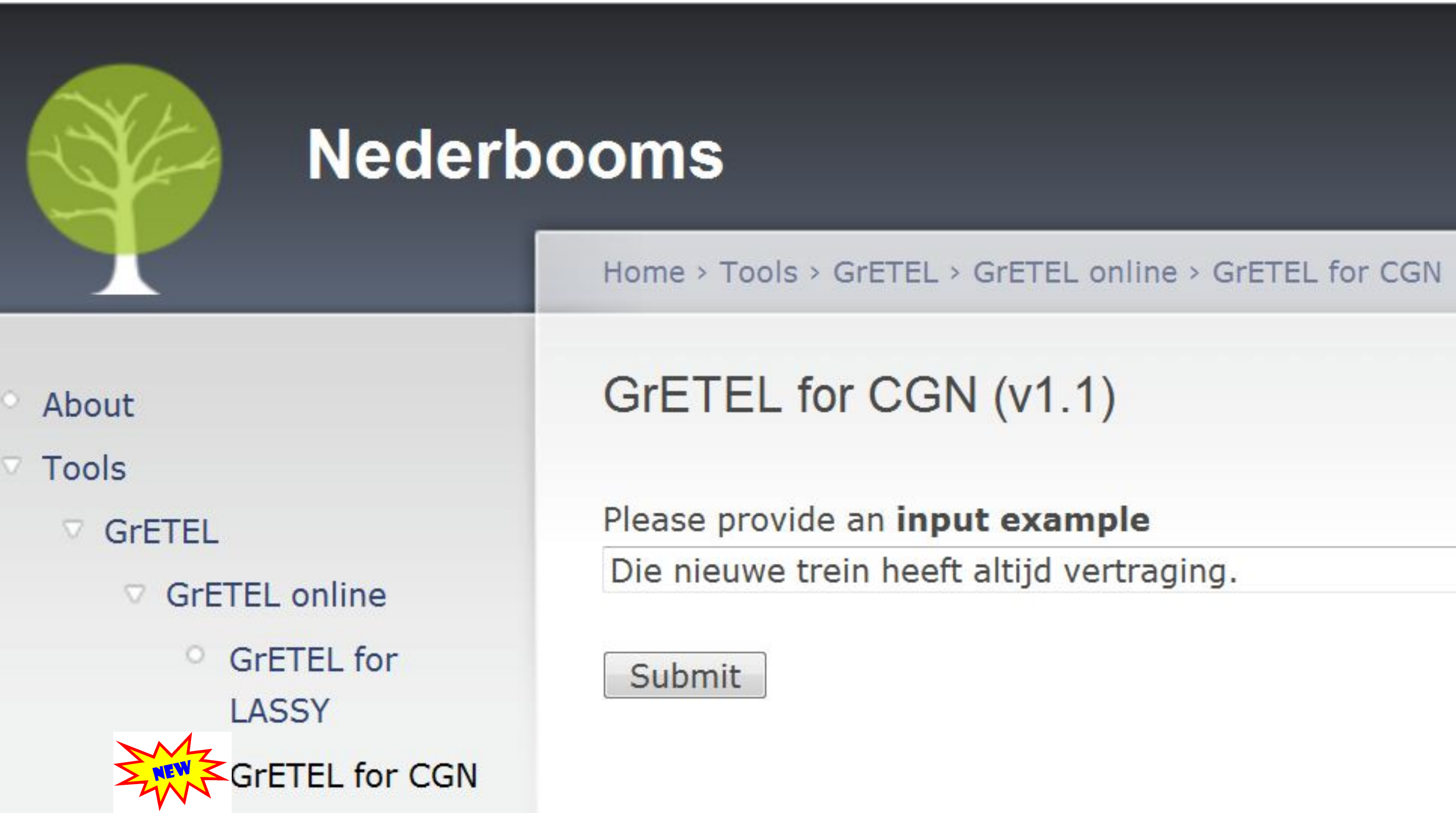
GrETEL for CGN

NEW

die + adjective (with -e) + noun

e.g. ***Die nieuwe trein*** *heeft altijd vertraging.*
“ **That new train** always has a delay. ”

INPUT



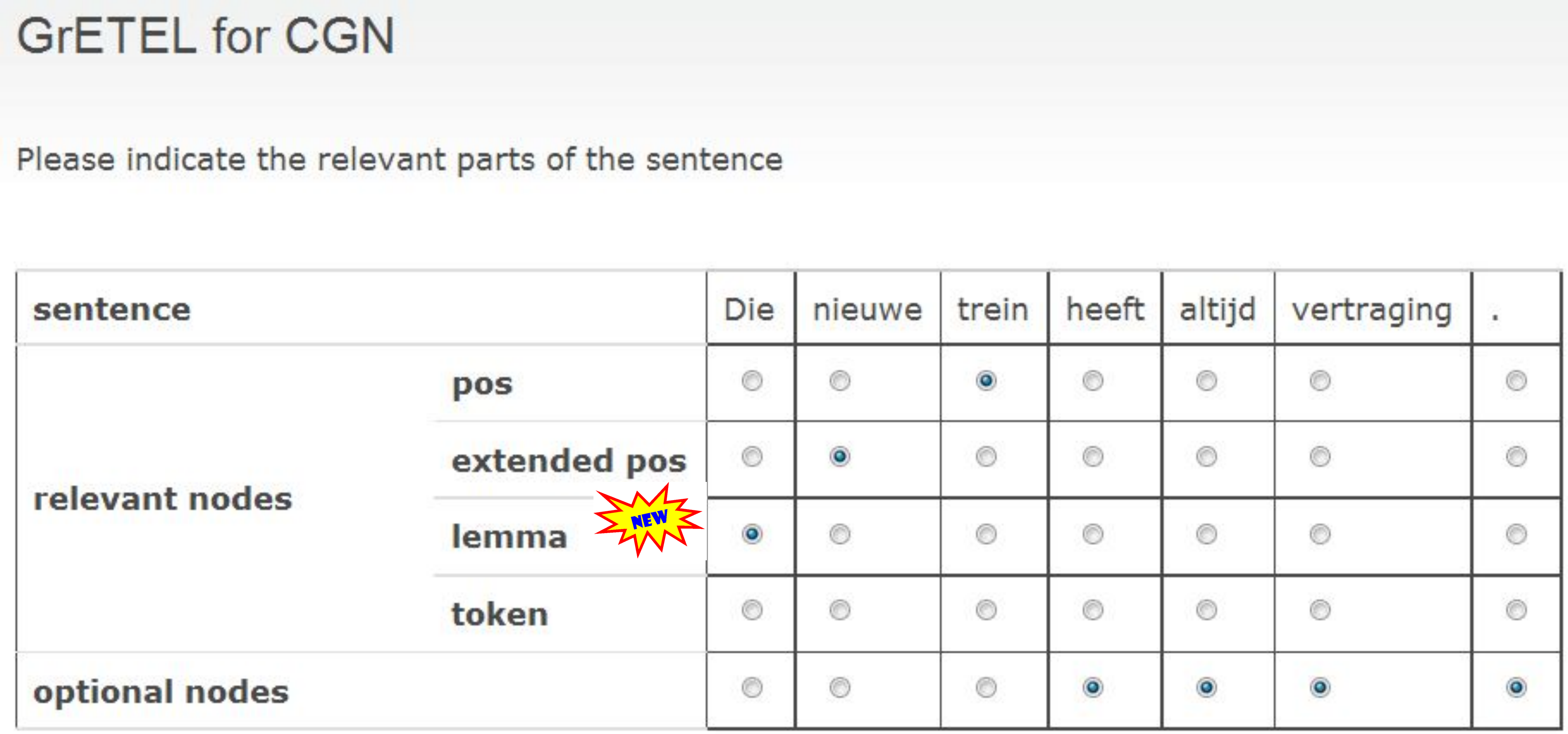
Home › Tools › GrETEL › GrETEL online › GrETEL for CGN

GrETEL for CGN (v1.1)

Please provide an **input example**

Die nieuwe trein heeft altijd vertraging.

ANNOTATION MATRIX

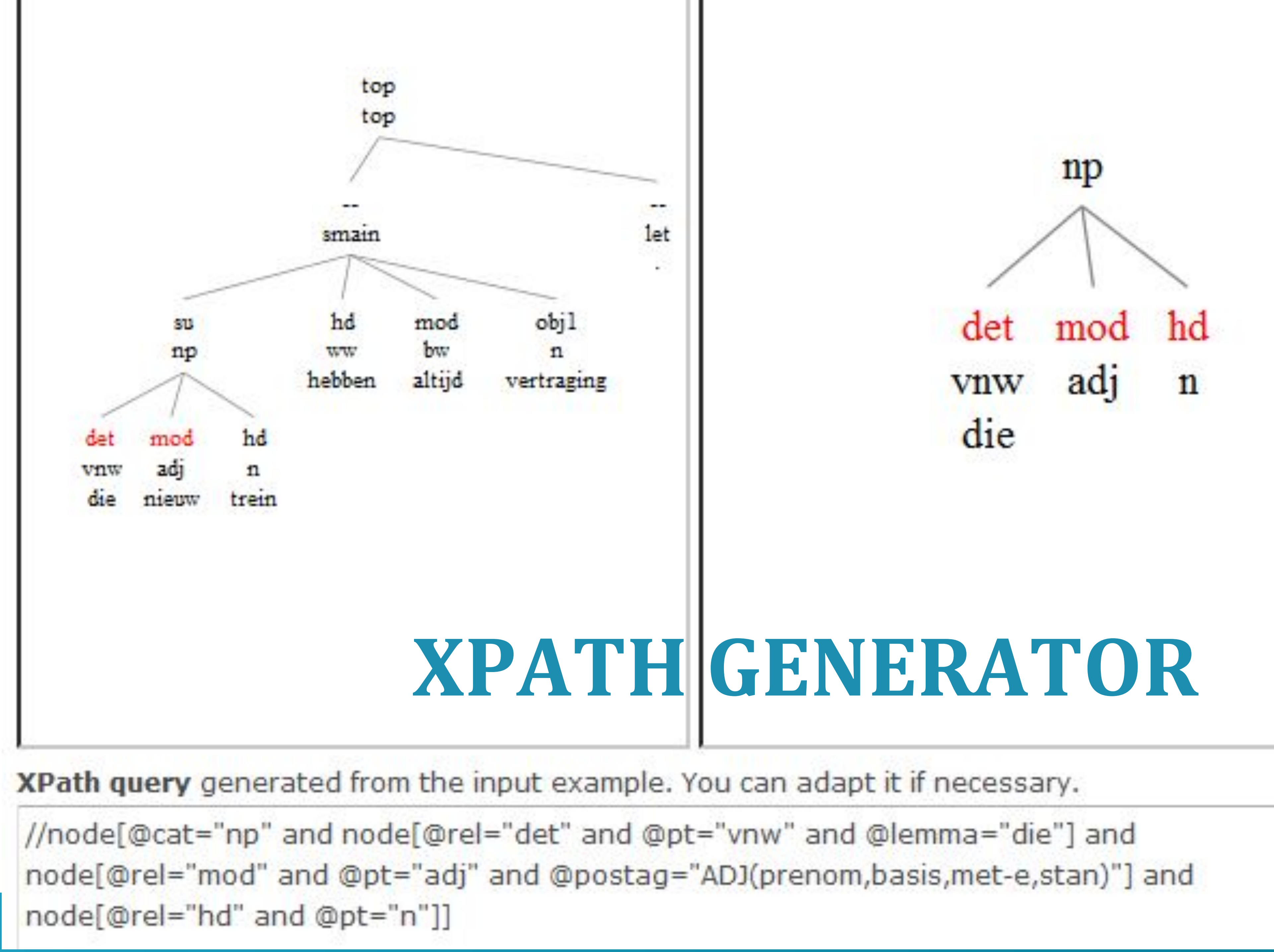


GrETEL for CGN

Please indicate the relevant parts of the sentence

sentence		Die	nieuwe	trein	heeft	altijd	vertraging	.
relevant nodes	pos	○	○	●	○	○	○	○
	extended pos	○	●	○	○	○	○	○
	lemma NEW	●	○	○	○	○	○	○
	token	○	○	○	○	○	○	○
optional nodes		○	○	○	●	●	●	●

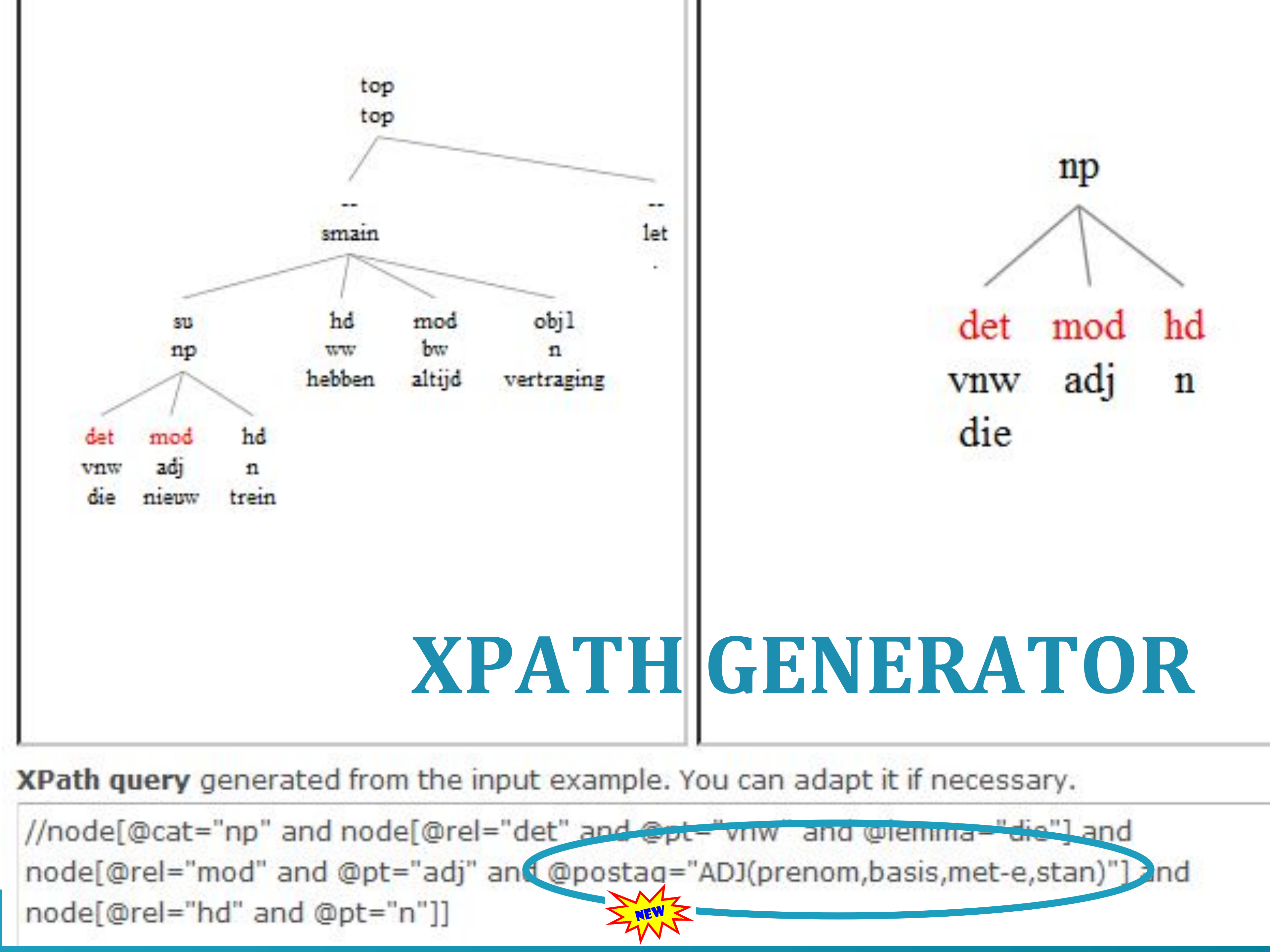
XPATH GENERATOR



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

```
//node[@cat="np" and node[@rel="det" and @pt="vnw" and @lemma="die"] and
node[@rel="mod" and @pt="adj" and @postag="ADJ(prenom,basis,met-e,stan)"] and
node[@rel="hd" and @pt="n"]]
```

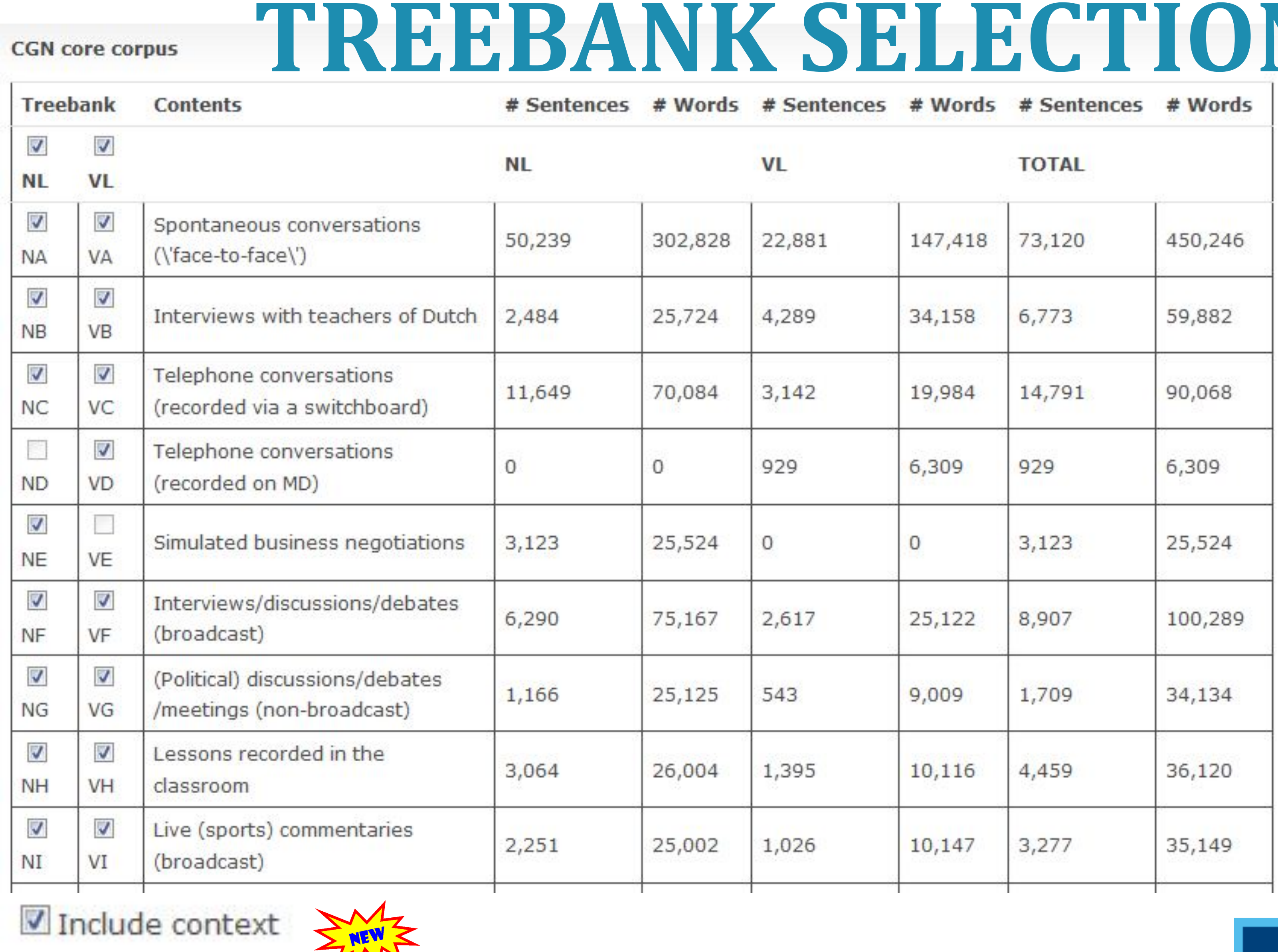
XPATH GENERATOR



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

```
//node[@cat="np" and node[@rel="det" and @pt="vnw" and @lemma="die"] and
node[@rel="mod" and @pt="adj" and @postag="ADJ(prenom,basis,met-e,stan)"] and
node[@rel="hd" and @pt="n"]]
```

TREEBANK SELECTION



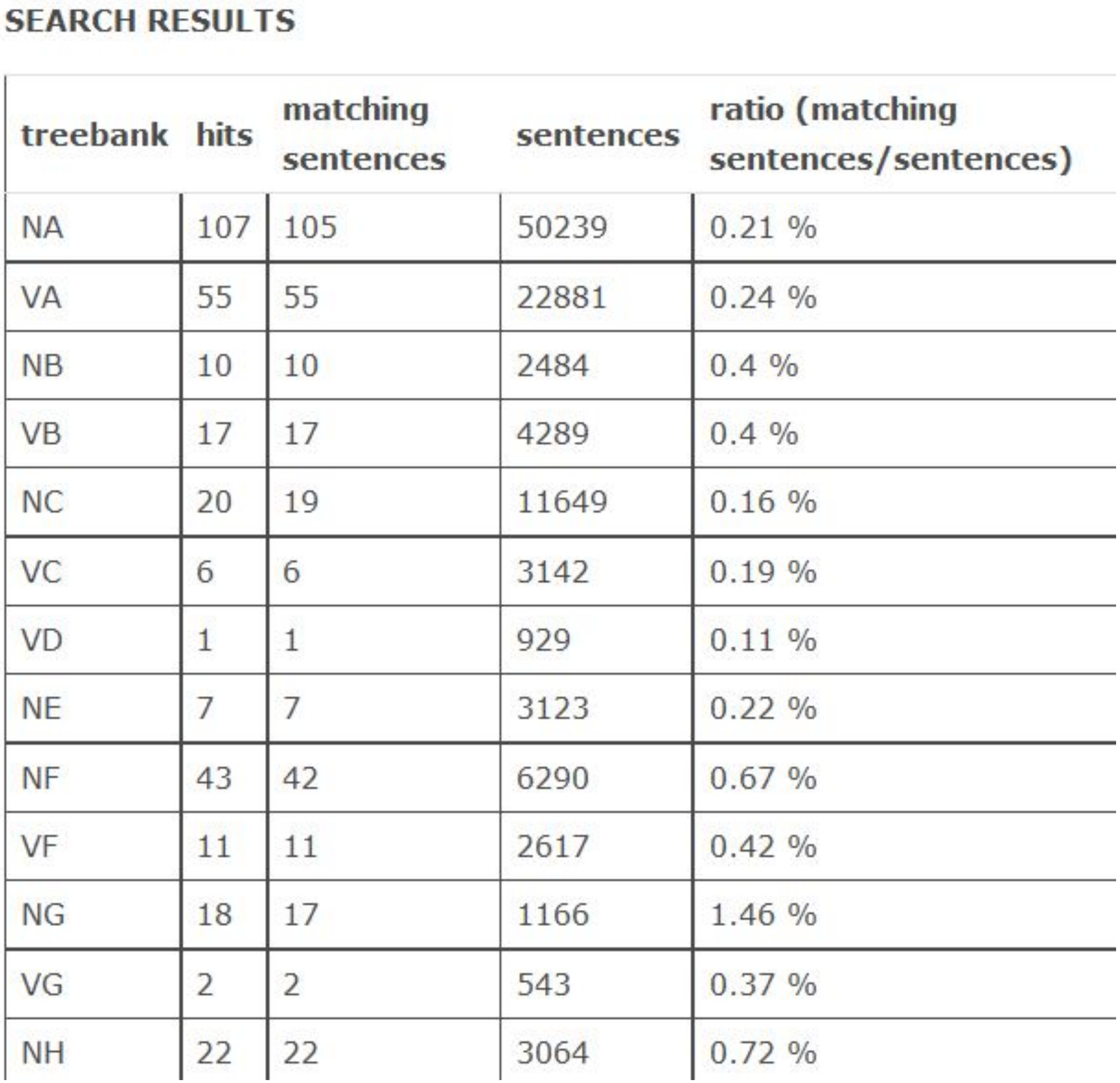
CGN core corpus

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

☑ Include context

NEW

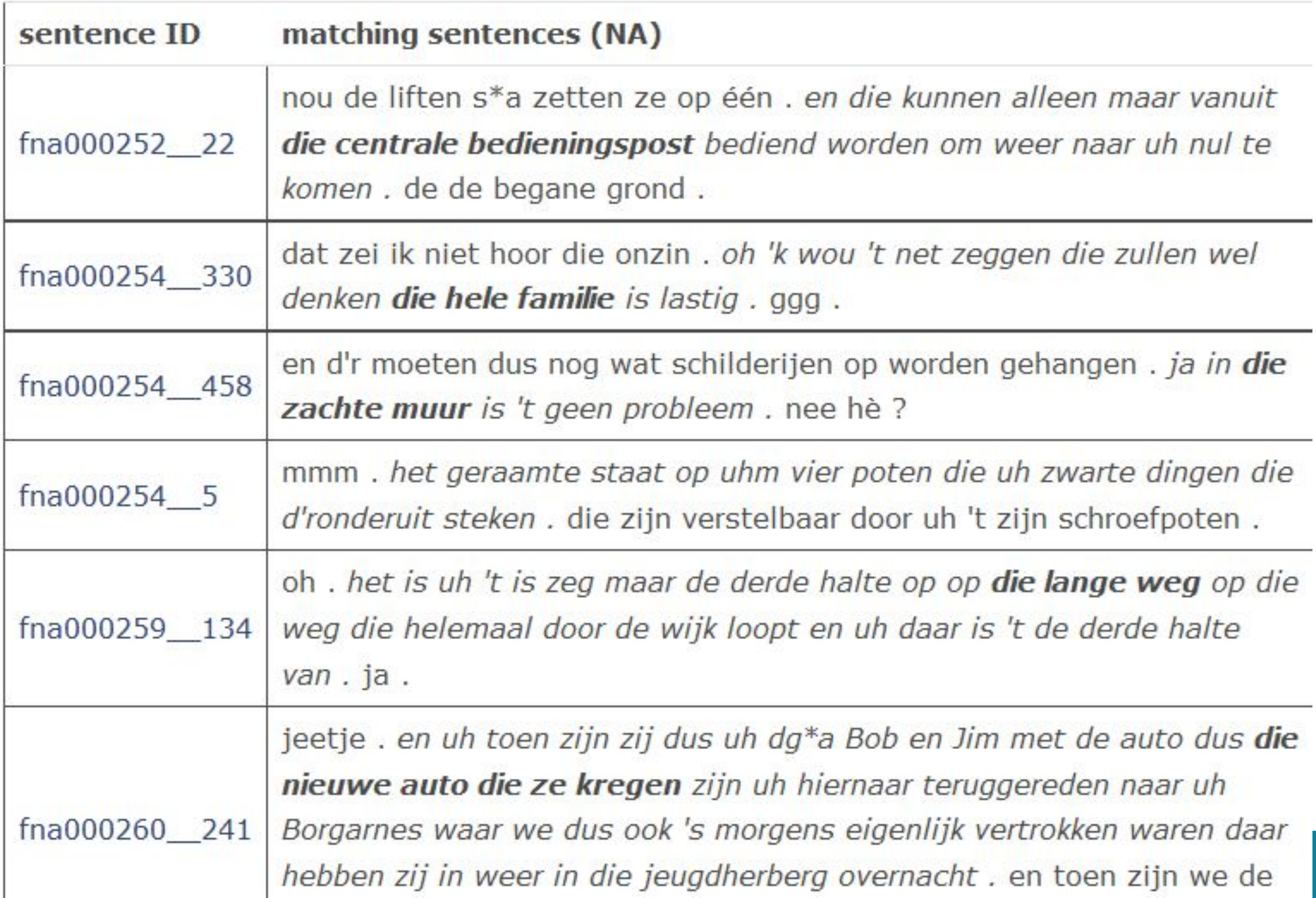
RESULTS: table



SEARCH RESULTS

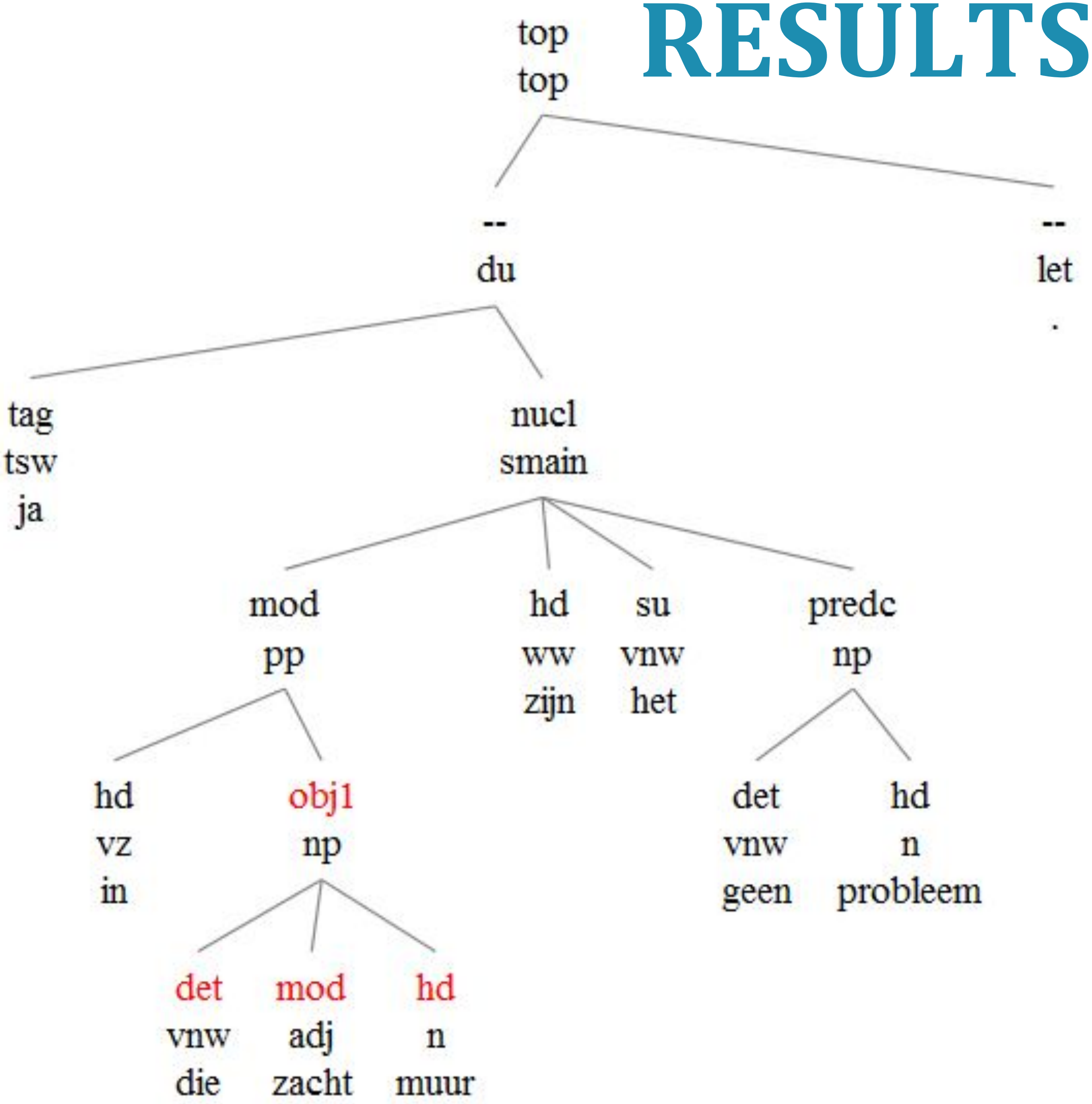
treebank	hits	matching sentences	sentences	ratio (matching sentences/sentences)
NA	107	105	50239	0.21 %
VA	55	55	22881	0.24 %
NB	10	10	2484	0.4 %
VB	17	17	4289	0.4 %
NC	20	19	11649	0.16 %
VC	6	6	3142	0.19 %
VD	1	1	929	0.11 %
NE	7	7	3123	0.22 %
NF	43	42	6290	0.67 %
VF	11	11	2617	0.42 %
NG	18	17	1166	1.46 %
VG	2	2	543	0.37 %
NH	22	22	3064	0.72 %

RESULTS: data



sentence ID	matching sentences (NA)
fna000252__22	nou de liften s*a zetten ze op één . *en die kunnen alleen maar vanuit* ***die centrale bedieningspost*** *bediend worden om weer naar uh nul te komen* . de de begane grond .
fna000254__330	dat zei ik niet hoor die onzin . *oh 'k wou 't net zeggen die zullen wel denken* ***die hele familie*** *is lastig* . ggg .
fna000254__458	en d'r moeten dus nog wat schilderijen op worden gehangen . *ja in* ***die zachte muur*** *is 't geen probleem* . nee hè ?
fna000254__5	mmm . *het geraamte staat op uhm vier poten die uh zwarte dingen die d'ronderuit steken* . die zijn verstelbaar door uh 't zijn schroefpoten .
fna000259__134	oh . *het is uh 't is zeg maar de derde halte op op* ***die lange weg*** *op die weg die helemaal door de wijk loopt en uh daar is 't de derde halte van* . ja .
fna000260__241	jeetje . *en uh toen zijn zij dus uh dg*a Bob en Jim met de auto dus* ***die nieuwe auto die ze kregen*** *zijn uh hiernaar teruggereden naar uh Borgarnes waar we dus ook 's morgens eigenlijk vertrokken waren daar hebben zij in weer in die jeugdherberg overnacht* . en toen zijn we de

RESULTS: trees



GrETEL for CGN

die + adjective (with -e) + noun

e.g. ***Die nieuwe trein** heeft altijd vertraging.*

“ **That new train** always has a delay. ”

⟹ 439 hits in 431 sentences

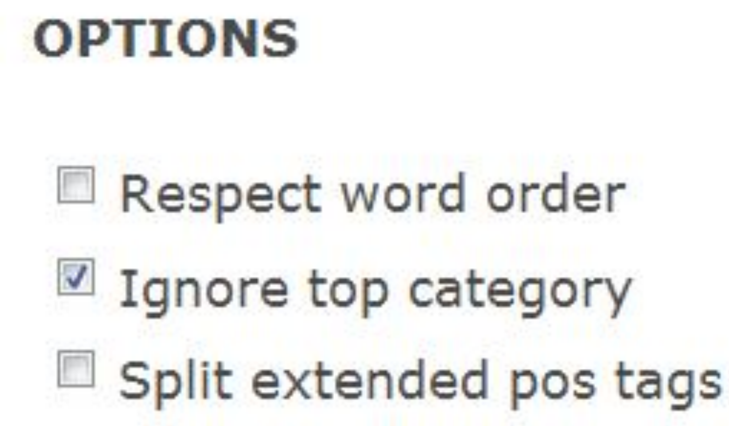
dat + adjective (without -e) + noun

e.g. ***Dat nieuw model** heeft altijd vertraging.*

“ **That new model** always has a delay. ”

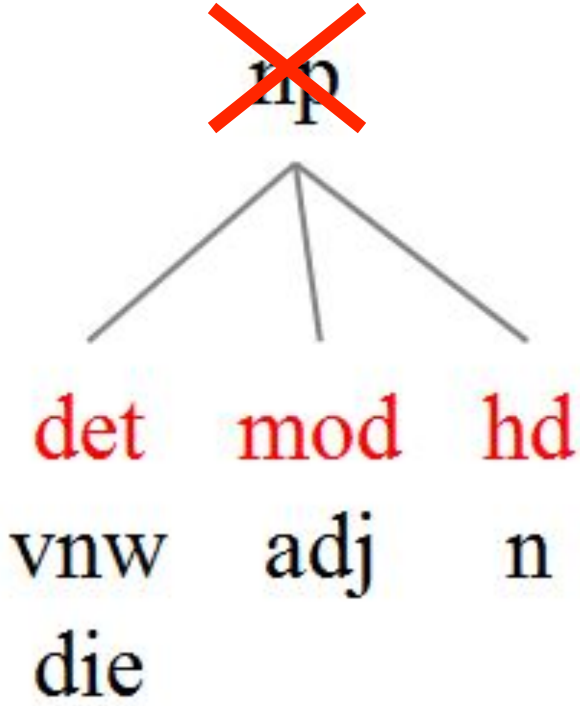
=> 37 hits in 37 sentences

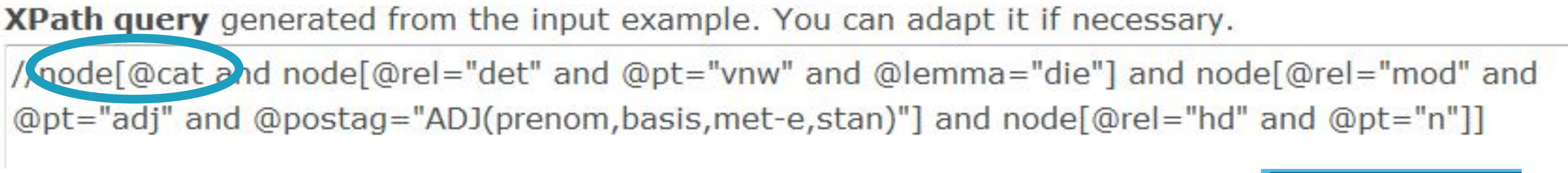
SEARCH OPTIONS



OPTIONS

- □ Respect word order
- ☑ Ignore top category
- □ Split extended pos tags

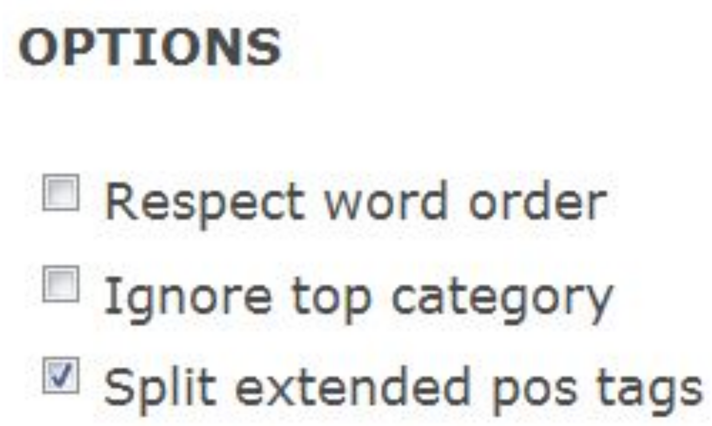




XPath query generated from the input example. You can adapt it if necessary.

```
//node[@cat and node[@rel="det" and @pt="vnw" and @lemma="die"] and node[@rel="mod" and
@pt="adj" and @postag="ADJ(prenom,basis,met-e,stan)"] and node[@rel="hd" and @pt="n"]]
```

SEARCH OPTIONS

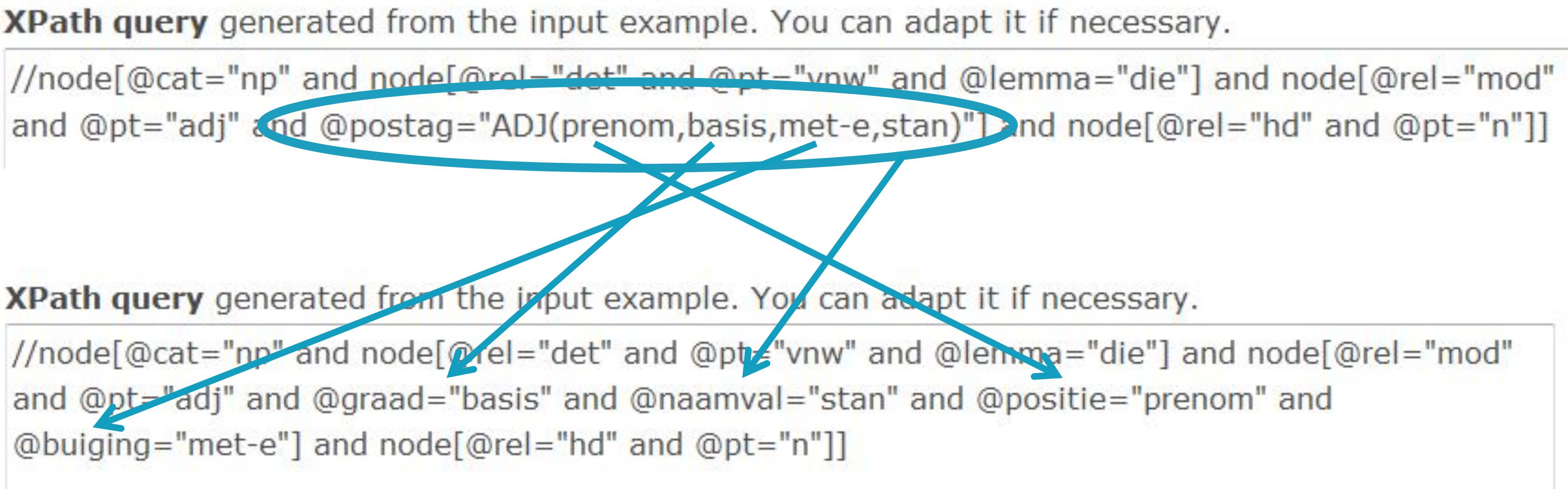


OPTIONS

□ Respect word order

□ Ignore top category

☑ 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 @pt="vnw" and @lemma="die"] and node[@rel="mod"
and @pt="adj" and @postag="ADJ(prenom,basis,met-e,stan)"] and node[@rel="hd" and @pt="n"]]
```

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

```
//node[@cat="np" and node[@rel="det" and @pt="vnw" and @lemma="die"] and node[@rel="mod"
and @pt="adj" and @graad="basis" and @naamval="stan" and @positie="prenom" and
@buiging="met-e"] and node[@rel="hd" and @pt="n"]]
```

OUTLINE

- GrETEL in a nutshell
- What’s new?
  - CGN treebank
  - Search options
- **Conclusions and future work**

CONCLUSIONS and FUTURE WORK

- GrETEL for CGN

- Lemmas and CGN POS tags added

  ⇒ more accurate results

  ⇒ more search options

  ⇒**to be included in GrETEL for LASSY**

CONCLUSIONS and FUTURE WORK

- GrETEL for CGN

- Lemmas and CGN POS tags added

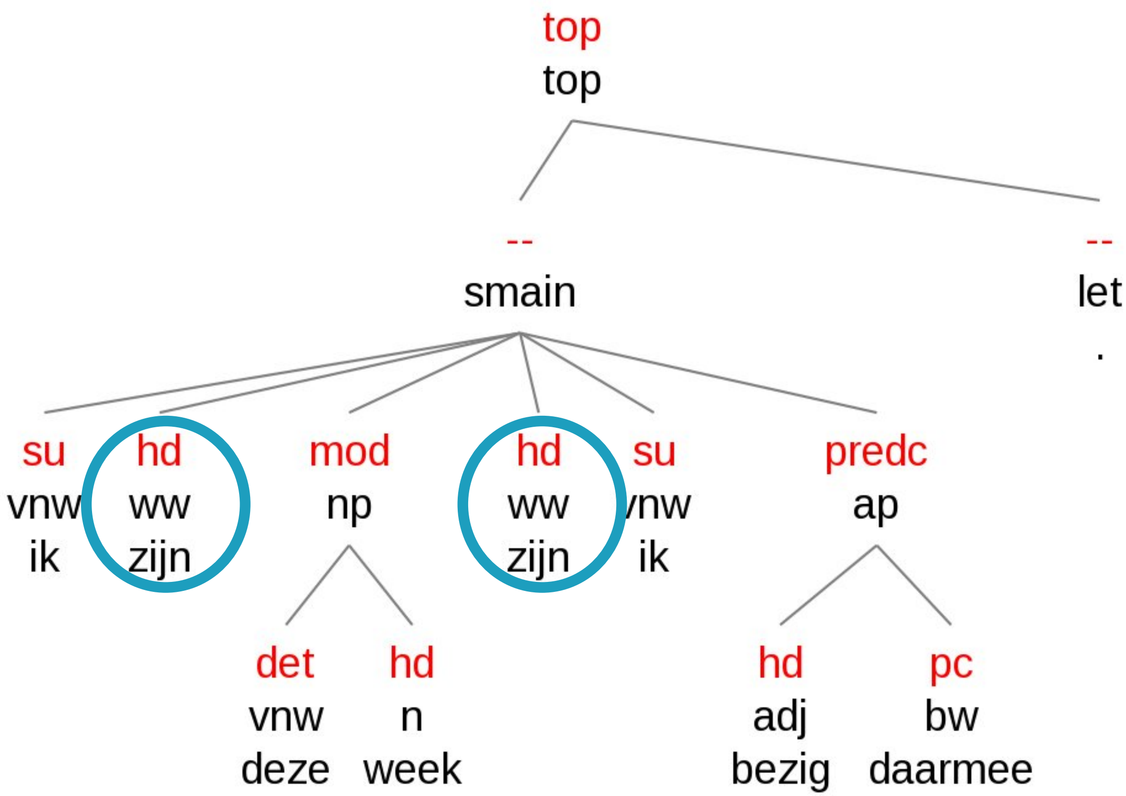
  ⇒ more accurate results

  ⇒ more search options

  ⇒ to be included in GrETEL for LASSY

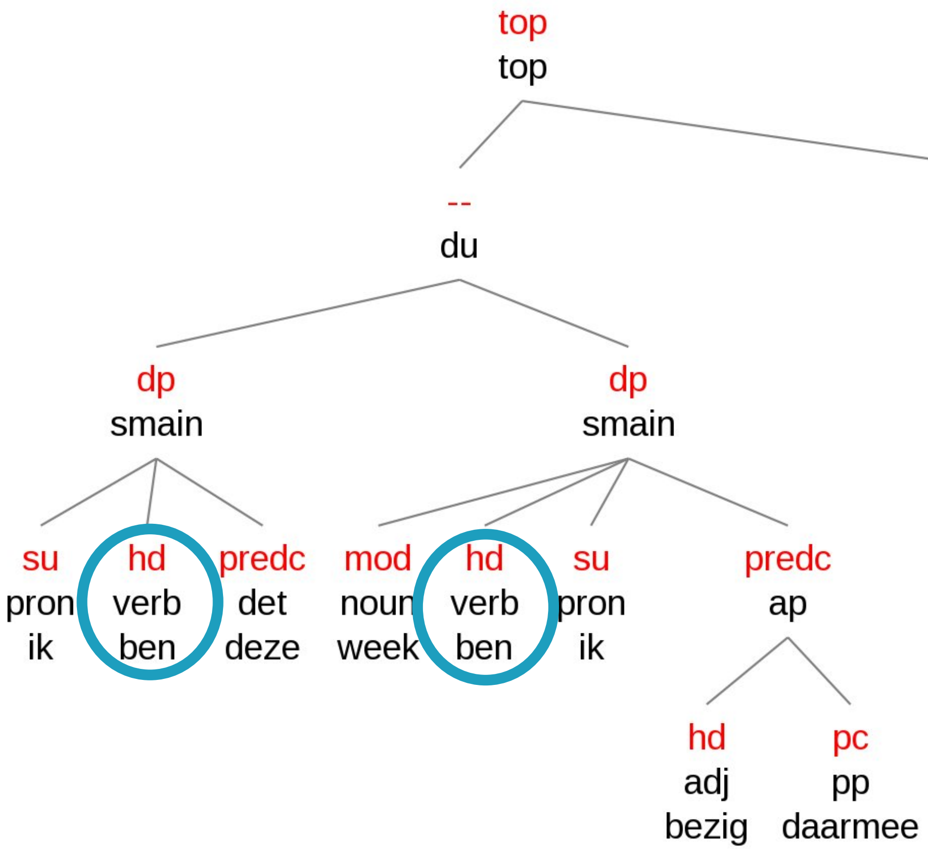
- Problem: typical spoken language phenomena

SPOKEN LANGUAGE



Ik ben deze weken ben ik bezig daarmee.
'The following weeks I'll be working on it.' (CGN, fna000276__258)

SPOKEN LANGUAGE



Ik ben deze weken ben ik bezig daarmee.

‘The following weeks I’ll be working on it.’ (Alpino parse)

CONCLUSIONS and FUTURE WORK

- GrETEL for CGN

- Lemmas and CGN POS tags added

  ⇒ more accurate results

  ⇒ more search options

  ⇒ to be included in GrETEL for LASSY

- Problem: typical spoken language phenomena

- Speed up search

Try it yourself at
http://nederbooms.ccl.kuleuven.be/eng/gretel

Thanks for your attention!