KATHOLIEKE UNIVERSITEIT
LEUVEN

QUERYING TREEBANKS BY EXAMPLE

Liesbeth Augustinus
Vincent Vandeghinste

KU LEUVEN

CLIN 22 – January 20th, 2012

NEDERBOOMS

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

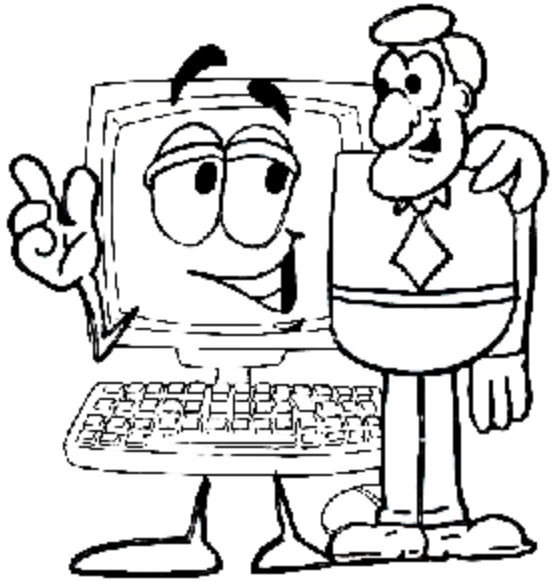
- September 2010 – February 2012
- Frank Van Eynde (CCL) and Hans Smessaert (NGTG)
- Liesbeth Augustinus and Vincent Vandeghinste (CCL)
- CLARIN Project

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

CLARIN
Common Language Resources and Technology Infrastructure

NEDERBOOMS

How can we combine the **data-oriented approach** of **treebank mining** with the **knowledge-oriented** method of **theoretical and descriptive linguistics?**



OUTLINE

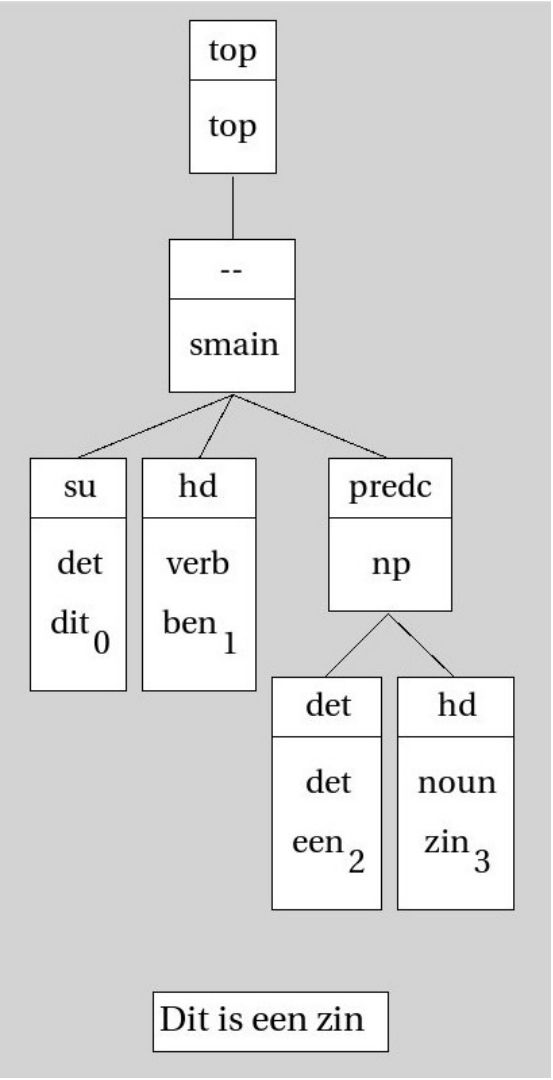
- The Nederbooms project
- **Querying LASSY**
  - **Existing tool & querying issues**
  - GrETEL
- Conclusion and future research

LASSY

- Written texts
  - Wikipedia, books, newspapers, reports, websites, law texts...
- ALPINO parser (van Noord)
  - Automatic syntactic annotation, XML-trees
- LASSY small
  - 1 million words, 65200 sentences
  - Manually corrected
- LASSY large
  - 1.5 billion words
  - Not corrected

ALPINO

“Dit is een zin” >> ALPINO parser >>



QUERYING LASSY

- Existing search tool: **dtsearch** (Kloosterman 2007)
  - Query language **XPath**

    = standard query language for xml trees

QUERYING LASSY

- Existing search tool: **dtsearch** (Kloosterman 2007)
  - Query language **XPath**

    = standard query language for xml trees

- Some examples
  - “look for all nodes in which a noun is modified by the adjective 'politiek' e.g. *politieke discussies*”

```
//node/node[@root='politiek' and @pos='adj']/../
node[@pos='noun']
```

  - “look for all subclauses where the participle occurs before the finite verb e.g. *... dat het gebeurd is*”

```
//node[@cat='ssub' and ((./node[@rel='vc' and
@cat='ppart']/node[@rel='hd' and @pos='verb']/@begin <
./node[@rel='hd' and @pos='verb']/@begin)) and (./node[@rel='vc'
and @cat='ppart']/node[@rel='hd' and
@pos='verb']/../../node[@rel='hd' and @pos='verb'])]
```

QUERYING LASSY

XPath

- Not user-friendly
- Knowledge of Alpino grammar necessary

= problematic for non-technical linguists

- Verify theory through data with corpus or treebank examples
- Time consuming, requires some effort

How to make interaction between computational linguistics and theoretical linguistics possible?

OUTLINE

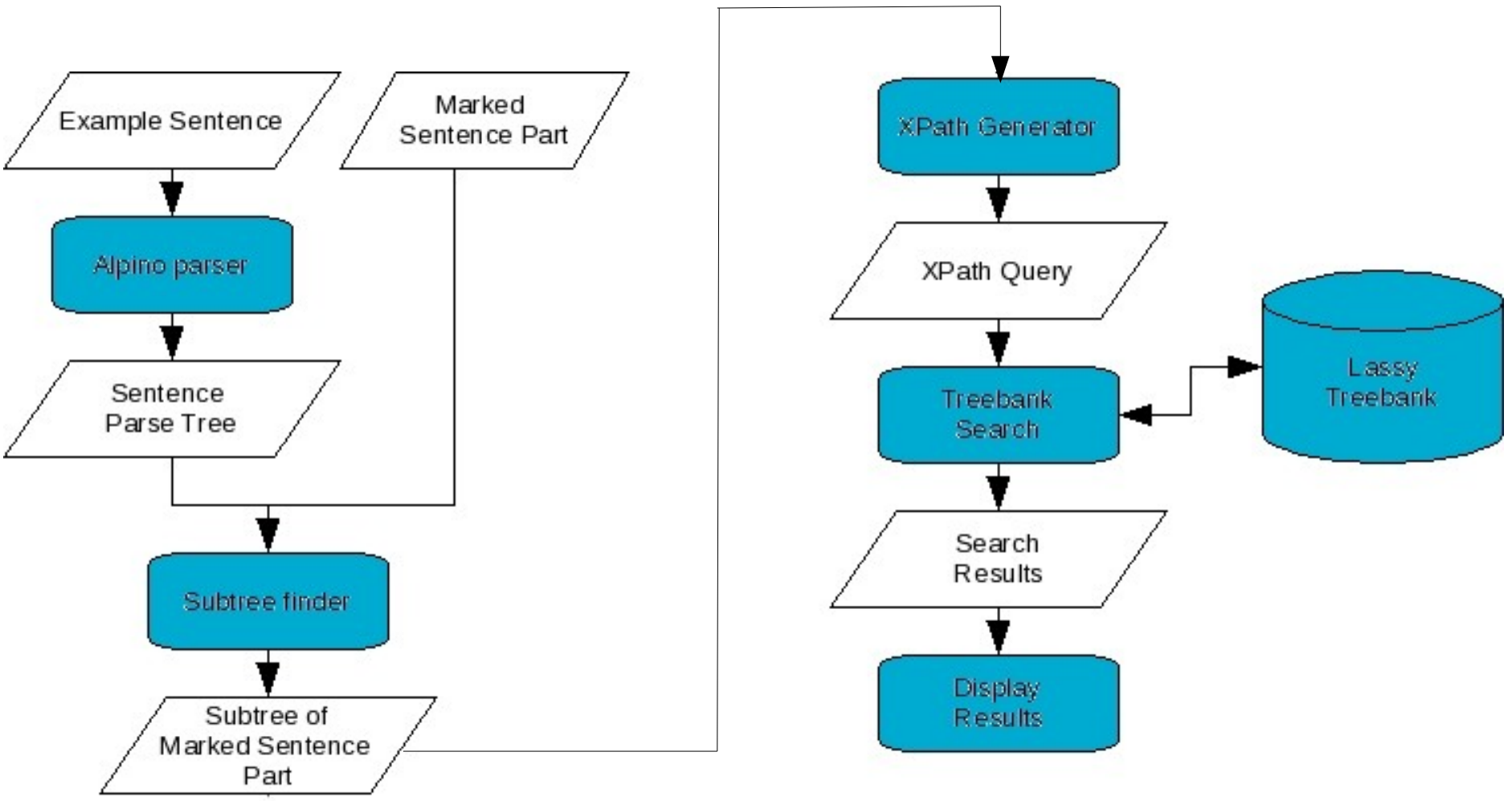
- The Nederbooms project
- **Querying LASSY**
  - Existing tool & querying issues
  - **GrETEL**
- Conclusion and future research

GrETEL

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

- Search tool based on example sentences
- Input = natural language
- No explicit knowledge of formal query language nor Alpino grammar required
- Bridge gap between descriptive and computational linguistics

GrETEL - architecture



GrETEL – input

Zowel [NP] als [NP] + singular or plural?

bv. Zowel de politie als de brandweer is/zijn ter plaatse. (Taaladvies.net)

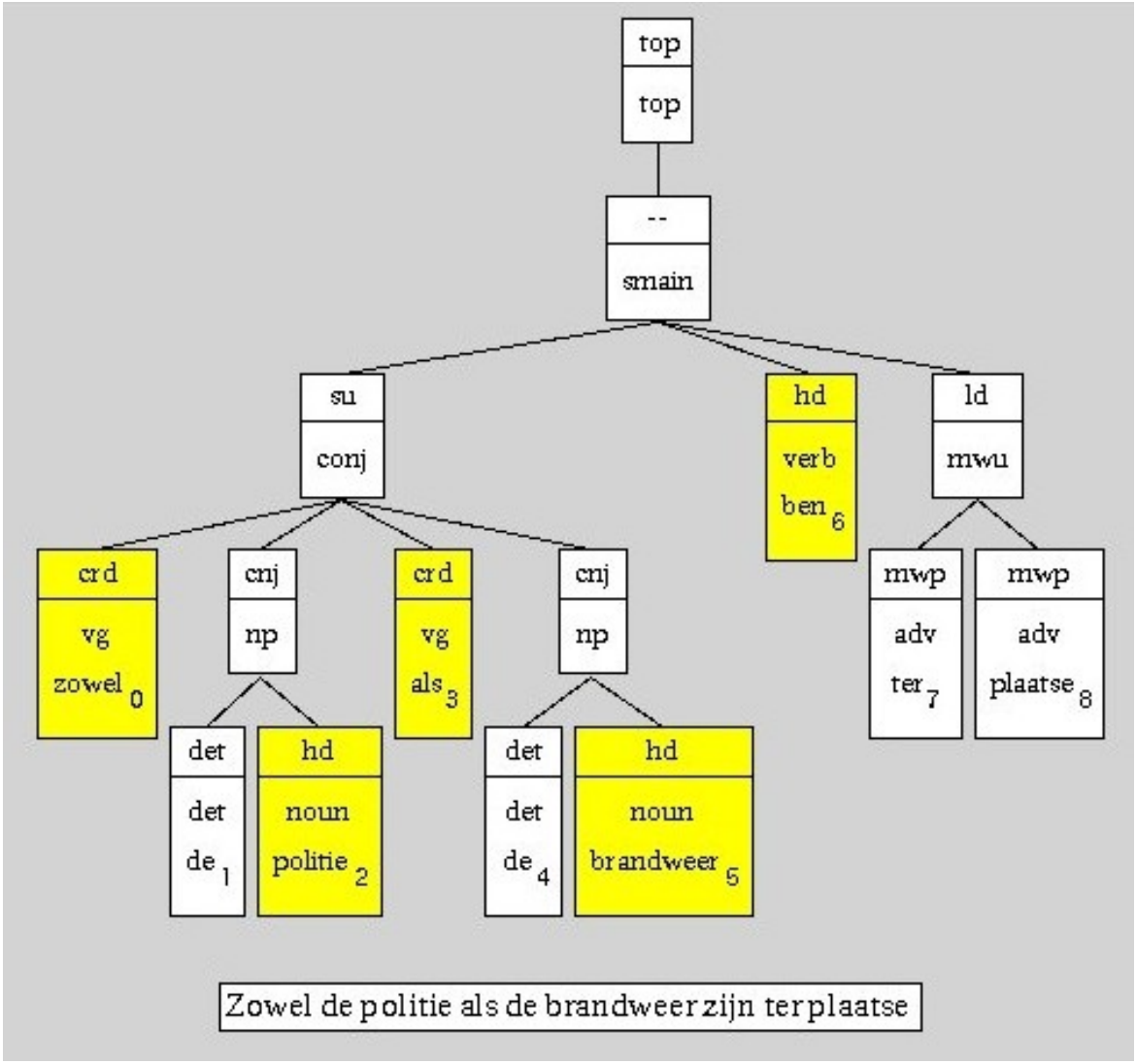
“Both the police and the fire brigade is/are on site.”

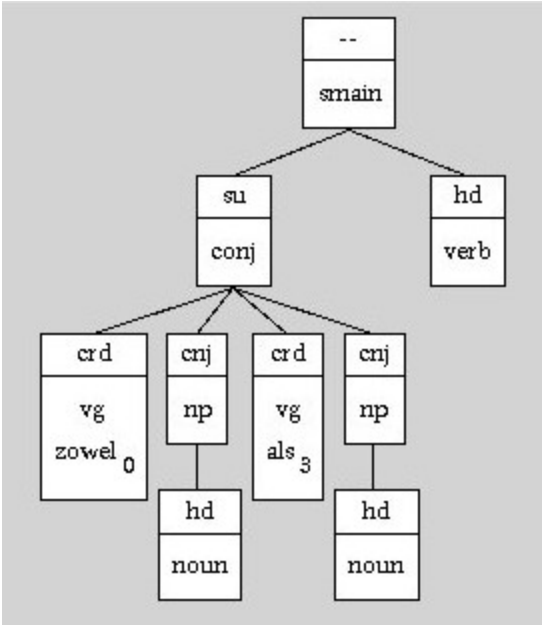
>> parsed with Alpino

GrETEL – annotation

>> add info to Alpino parse

GrETEL – subtree





GrETEL – XPath

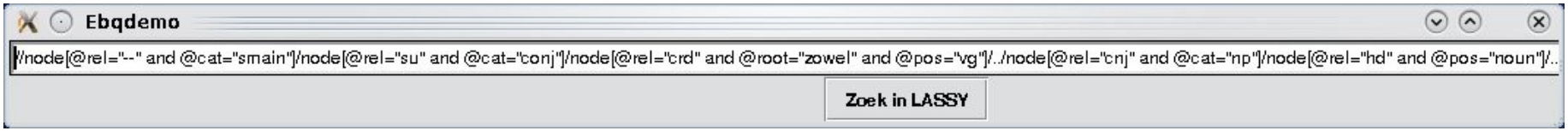
- XPath generated automatically

```
//node[@cat='smain' and ./node[@rel='su' and
@cat='conj']/node[@rel='crd' and @root='zowel' and
@pos='vg']/../node[@rel='cnj' and @cat='np']/node[@rel='hd'
and @pos='noun']/../../node[@rel='crd' and @root='als' and
@pos='vg']/../node[@rel='cnj' and @cat='np']/node[@rel='hd'
and @pos='noun']/../../../node[@rel='hd' and @pos='verb']]
```

= long, specific query

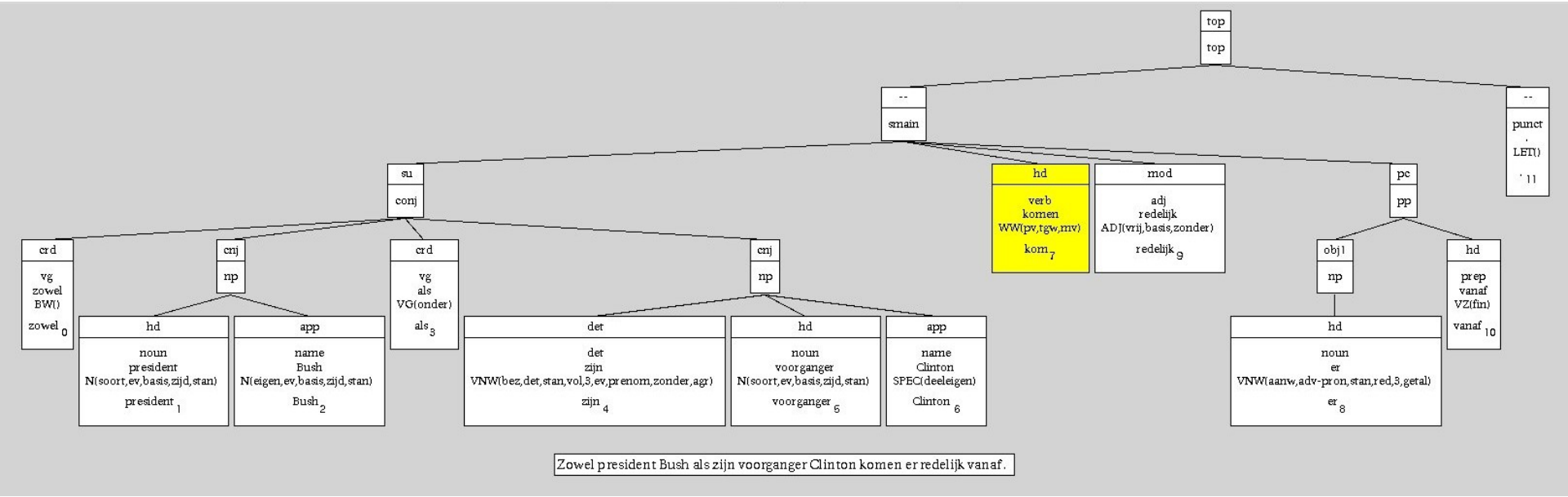
> manually adaptable in advanced search mode

- to generalise (include names, pron)
- to specify (look for singular finite verbs only)



GrETEL – output

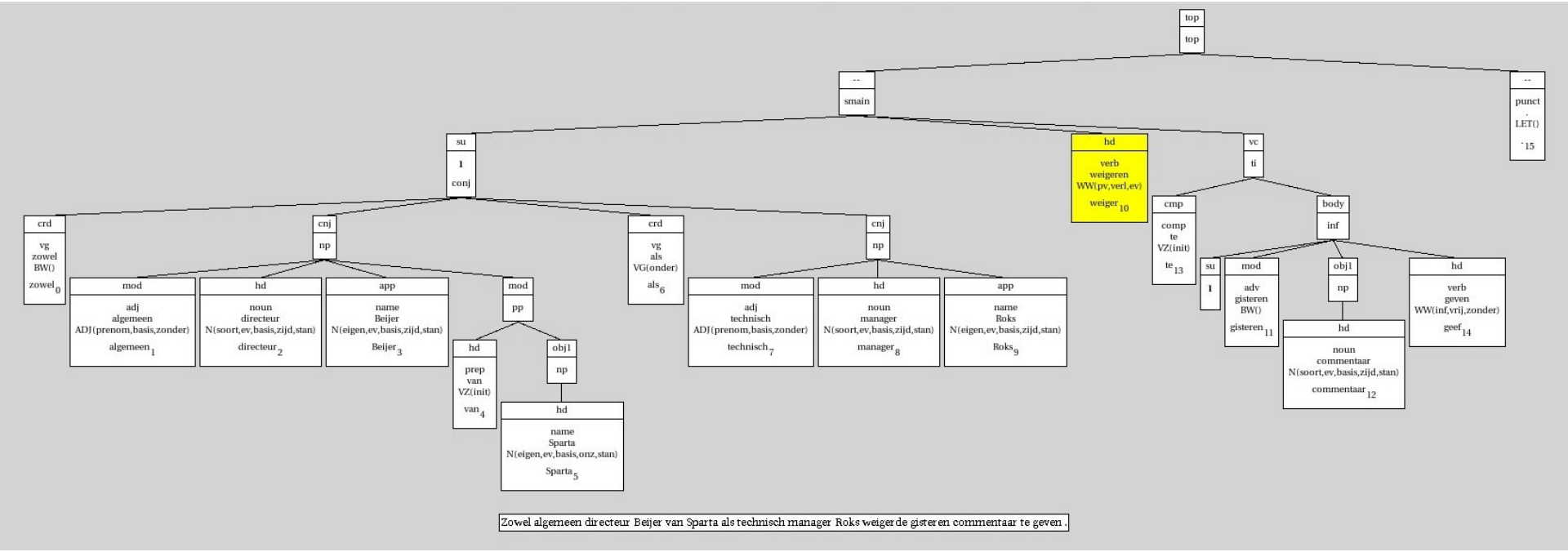
- 22 results > sample of similar sentences
- Singular and plural form of the finite verb



Zowel president Bush als zijn voorganger Clinton **komen** er redelijk vanaf.

“Both president Bush and his predecessor Clinton did reasonably well.”

GrETEL – output



Zowel algemeen directeur Beijer van Sparta als technisch manager Roks **weigerde** gisteren commentaar te geven.

“Both general director Beijer of Sparta and technical manager Roks refused to comment yesterday.”

OUTLINE

- The Nederbooms project
- Querying LASSY
  - Existing tool & querying issues
  - GrETEL
- **Conclusion and future research**

CONCLUSION

- **Nederbooms**: Treebank mining for empirical linguistics

- Search tool for descriptive linguistics: **GrETEL**
  - LASSY Treebank
  - User-friendly:
    - input = natural language
    - XPath generator > knowledge of formal query language not required
    - ordering filter
    - basic and advanced search mode
  - Output = sample of similar sentences

FUTURE RESEARCH

- Speed up search
  - > query LASSY large (1.5 billion words)
  - > breadth-first string representation, Varro (Martens 2010)
- Add more features
  - > e.g. query subcorpora
- Automatically look for closely-related examples (more or less specific)
- Add more quantitative information
- Webservice

Thanks for your attention!

Questions?

liesbeth@ccl.kuleuven.be

REFERENCES

- Augustinus, L., Vandeghinste, V. and Van Eynde, F. *Example-Based Treebank Querying.* Submitted to LREC 2011.
- Jongejan B., Olsen, S. and Fersøe, H. *Validation Report Lassy Corpora Linguistic Validation.* Center for Sprogteknologi, University of Copenhagen, 2011
- Kloosterman, G. *An overview of the Alpino Treebank Tools.* Alfa Informatica, University of Groningen. http://www.let.rug.nl/vannoord/alp/Alpino/TreebankTools.html, 2007.
- Van Belle, W. and Van Langendonck (eds.), *The Dative vol. I*, Amsterdam/Philadelphia: John Benjamins, 1996.
- van der Beek, L., *Topics in Corpus-Based Syntax.* Groningen Dissertations in Linguistics, 2005.
- Van Eynde, F. *Part of Speech Tagging en Lemmatisering van het D-Coi Corpus.* Centrum voor Computerlinguïstiek, KU Leuven, 2005.
- van Noord, G. *At Last Parsing Is Now Operational.* In TALN 2006, pp. 20-42, 2006.
- van Noord Gertjan, Gosse Bouma, Frank Van Eynde, Daniel de Kok, Jelmer van der Linde, Ineke Schuurman, Erik Tjong Kim Sang and Vincent Vandeghinste. Large Scale Syntactic Annotation of Written Dutch: Lassy. In Peter Spyns and Jan Odijk (eds.): Essential Speech and Language Technology for Dutch: resources, tools and applications. Springer, submitted.
- van Noord, G., Schuurman, I. and Bouma, G. *Lassy syntactische annotatie*, revision 19455. http://www.let.rug.nl/vannoord/Lassy, 2011.