KATHOLIEKE UNIVERSITEIT
LEUVEN

SEEING THE WOOD FOR THE TREES

Liesbeth Augustinus
Vincent Vandeghinste
Frank Van Eynde

KU LEUVEN

digiTAAL - December 19, 2011

DEDUCTION and INDUCTION

KATHOLIEKE UNIVERSITEIT LEUVEN

Descriptive and theoretical linguistics

- Focus on knowledge
- Introspection
- Theory > data
- Verify theory through data with corpus or treebank examples

Computational corpus linguistics

- Focus on data
- Inductive generalisation
- Data > theory
- Provide tools for linguistic research

How to make interaction between both approaches possible?

TREEBANKS

- Corpus with syntactic annotations
- Dependency structures and/or constituent structures
- Examples:
  - English (Penn Treebank)
  - German (VerbMobil, Tiger)
  - French (French Treebank)
  - Czech (Prague Dependency Treebank)
  - Swedish (Swedish Treebank)
  - Dutch (CGN, LASSY, SoNaR)

- Annual conference on Treebanks and Linguistic Theories since 2002

NEDERBOOMS

KATHOLIEKE UNIVERSITEIT LEUVEN

CLARIN
Common Language Resources and Technology Infrastructure

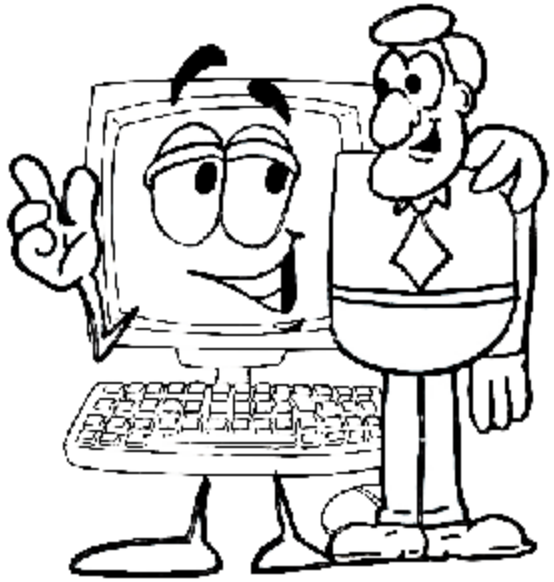
NEDERBOOMS

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

NEDERBOOMS

KATHOLIEKE UNIVERSITEIT LEUVEN

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
- LASSY Treebank
- Querying LASSY
- Conclusion and future research

OUTLINE

- The Nederbooms project
- **LASSY Treebank**
- Querying LASSY
- Conclusion and future research

LASSY

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

LASSY

- Validation report
  - Jongejan et al. 2011
  - Center for Sprogteknologi

- LASSY small
  - Lemmatisation — error rate 0,365%
  - POS-tagging — error rate 1,37%
  - Dependency relations — 99,8%
  - Sentence level accuracy — 97,8%

LASSY

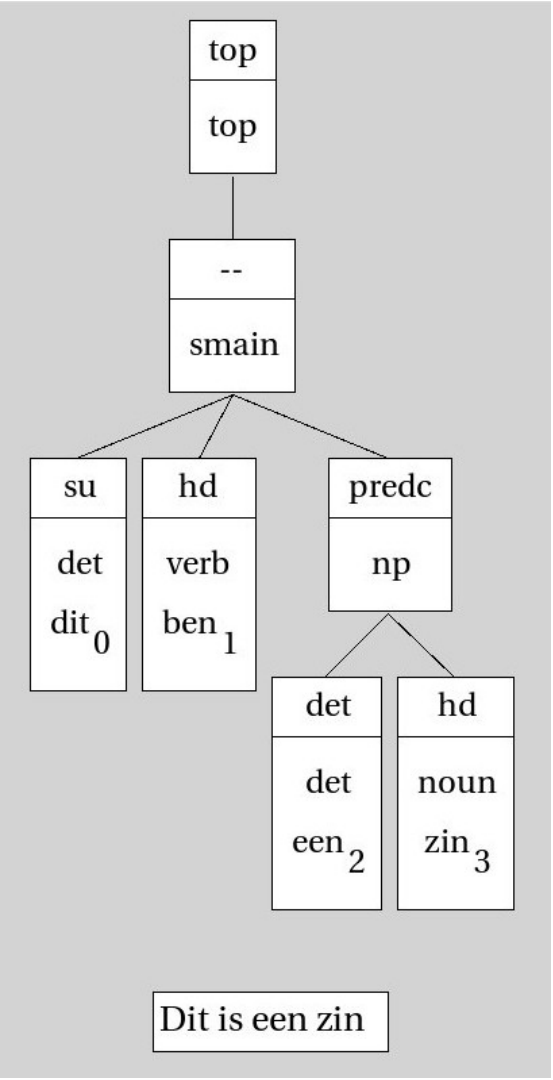
KATHOLIEKE UNIVERSITEIT LEUVEN

"Dit is een zin" >> ALPINO >>

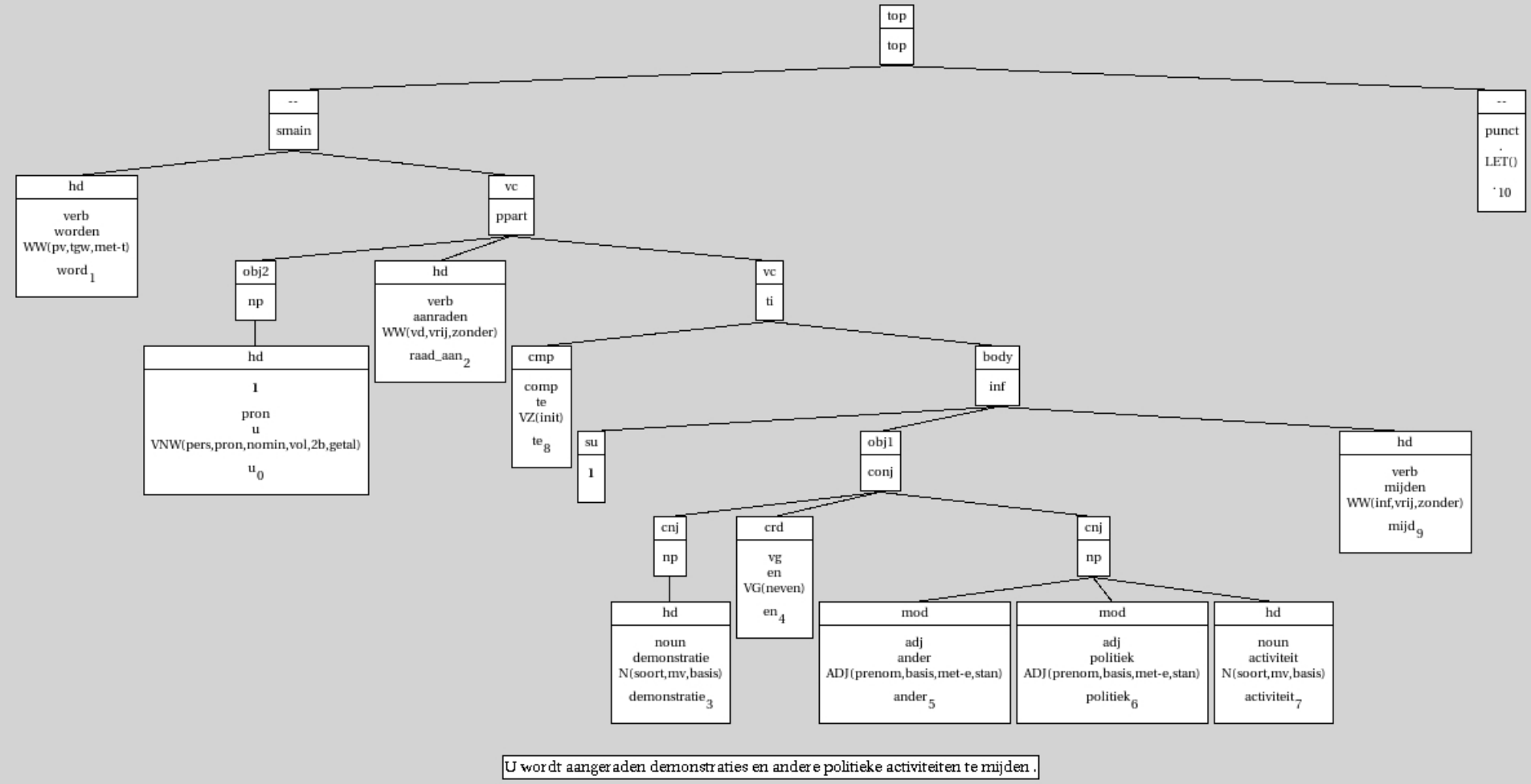
```
<alpino_ds version="1.3">
  <node begin="0" cat="top" end="4" id="0" rel="top">
    <node begin="0" cat="smain" end="4" id="1" rel="--">
      <node begin="0" end="1" frame="determiner(het,nwh,nmod,pro,nparg)"
      id="2" infl="het" lcat="np" pos="det" rel="su" root="dit" sense="dit"
      wh="nwh" word="Dit"/>
      <node begin="1" end="2" frame="verb(unacc,sg_heeft,copula)" id="3"
      infl="sg_heeft" lcat="smain" pos="verb" rel="hd" root="ben" sc="copula"
      sense="ben" tense="present" word="is"/>
      <node begin="2" cat="np" end="4" id="4" rel="predc">
        <node begin="2" end="3" frame="determiner(een)" id="5" infl="een"
        lcat="detp" pos="det" rel="det" root="een" sense="een" word="een"/>
        <node begin="3" end="4" frame="noun(de,count,sg)" gen="de" id="6"
        lcat="np" num="sg" pos="noun" rel="hd" root="zin" sense="zin"
        word="zin"/>
      </node>
    </node>
  </node>
  <sentence>Dit is een zin</sentence>
</alpino_ds>
```

LASSY

"Dit is een zin" >> ALPINO parser >>



LASSY



U wordt aangeraden demonstraties en andere politieke activiteiten te mijden. (WR-P-E-H-0000000049.p.33.s.2)

“You are adviced to avoid demonstrations and other political activities.”

OUTLINE

- The Nederbooms project
- LASSY Treebank
- **Querying LASSY**
- Conclusion and future research

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 occurrences of 'politiek' ”

```
//node[@word='politiek']
```

QUERYING LASSY

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

    = standard query language for xml trees

- Some examples
  - “look for all occurrences of 'politiek' ”

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

QUERYING LASSY

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

    = standard query language for xml trees

- Some examples
  - “look for all occurrences of 'politiek' ”

    ```
    //node[@word='politiek'] vs.//node[@root='politiek']
    ```

    ```
    //node[@root='politiek' and @pos='noun']
    ```
  - “look for all nodes in which a noun is modified by the adjective 'politiek'”

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

QUERYING LASSY

- “look for all verbs that occur as a head of a *te*-infinitive.”

```
//node[@rel='vc' and @cat='ti']/node[@cat='inf']
/node[@rel='hd' and @pos='verb']
```

XPath

- Not user-friendly
- Knowledge of Alpino grammar necessary

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

Zowel ... als ... + singular or plural?

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

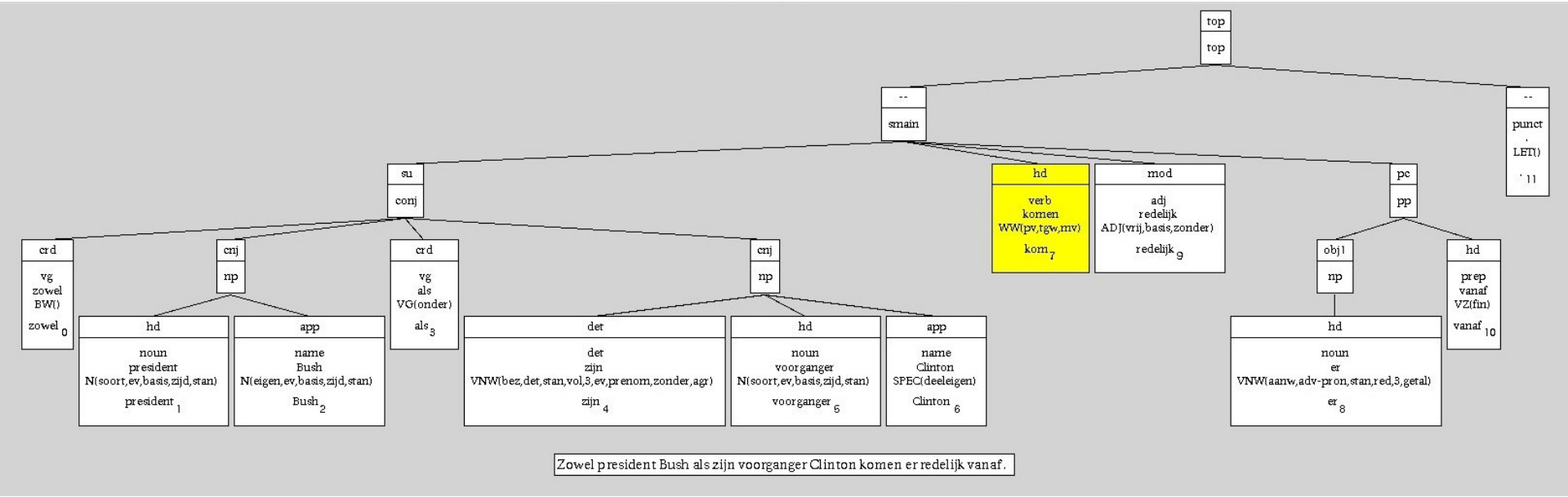
“Both the police and the fire brigade is/are on site.”

GrETEL


Ebqdemo
Markeer de interessante delen
VOORBEELDZIN
POS
Lemma
Woord
Niet relevant
Zowel
de
politie
als
de
brandweer
zijn
ter
plaatse
Volgorde is belangrijk
Eenvoudig zoeken
Zoeken met XPath
Verzenden
Terug

GrETEL

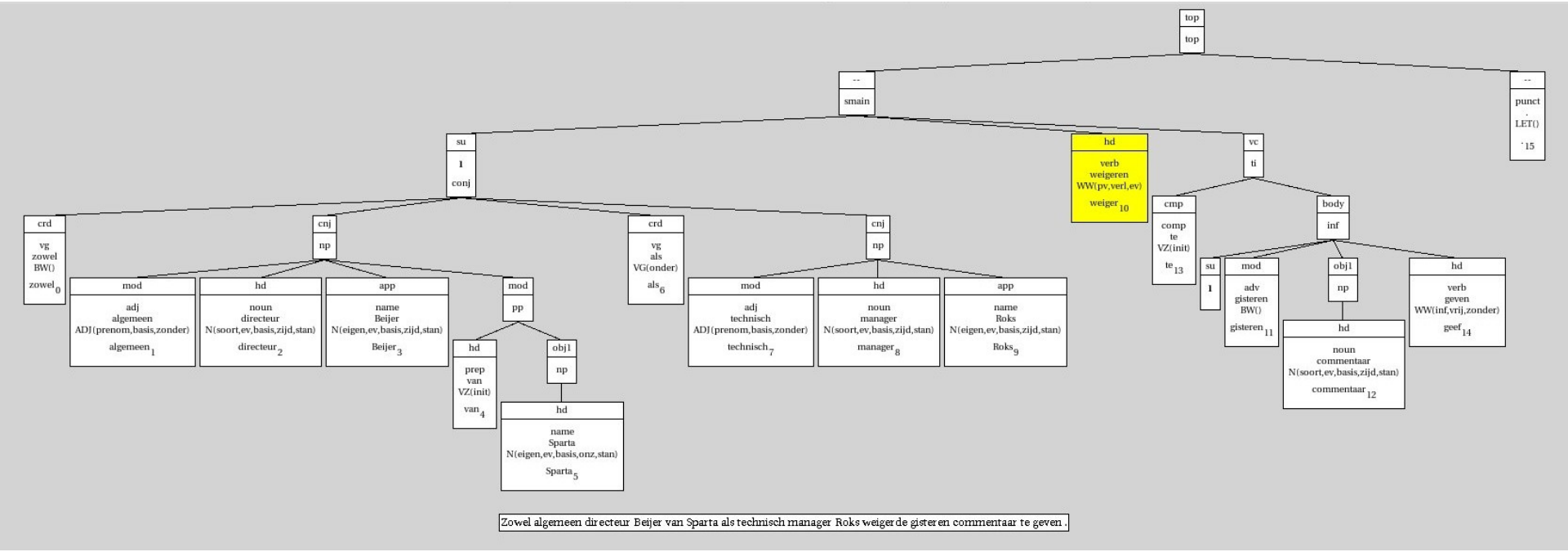
- 22 results
- Singular and plural form of the finite verb



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

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

GrETEL



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.”

GrETEL

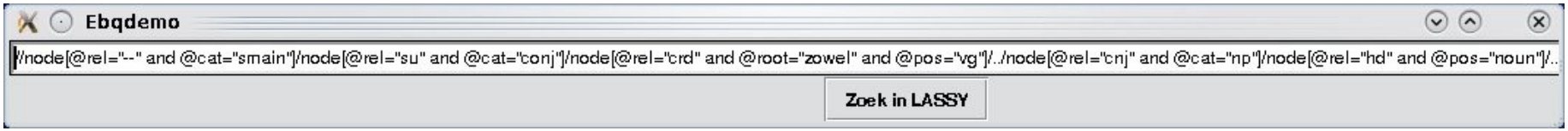
- XPath generated automatically

```
//node[@rel="--" and @cat="smain"]/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

KATHOLIEKE UNIVERSITEIT LEUVEN

- Treebank mining without knowledge of the internal structure of the treebank
- XPath automatically generated
- Basic and advanced search mode
- Ordering filter

OUTLINE

- The Nederbooms project
- LASSY Treebank
- Querying LASSY
- **Conclusion and future research**

CONCLUSION

NEDERBOOMS

- Treebank mining for empirical linguistics
- Search tool for descriptive linguistics: **GrETEL**
  - LASSY Treebank
  - User-friendly:
    - input = natural language
    - knowledge of formal query language not required
  - Sample of similar sentences

FUTURE RESEARCH

- Speed up search

  > query LASSY large (1.5 billion words)

- Add more features

  > e.g. query subcorpora

- Automatically look for close-related examples (more or less specific)
- Add more quantitative information
- Webservice

Questions?

Suggestions?

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.