Alpino and Corpus Linguistics

Gertjan van Noord
University of Groningen

April 11, 2005

Alpino and Corpus Linguistics

- Context: Wide-coverage Parsing with *Alpino*
  - ⋆ sophisticated linguistic analysis
  - ⋆ care about disambiguation and efficiency
  - ⋆ corpus-based evaluation methodology
- Applications in Corpus Linguistics

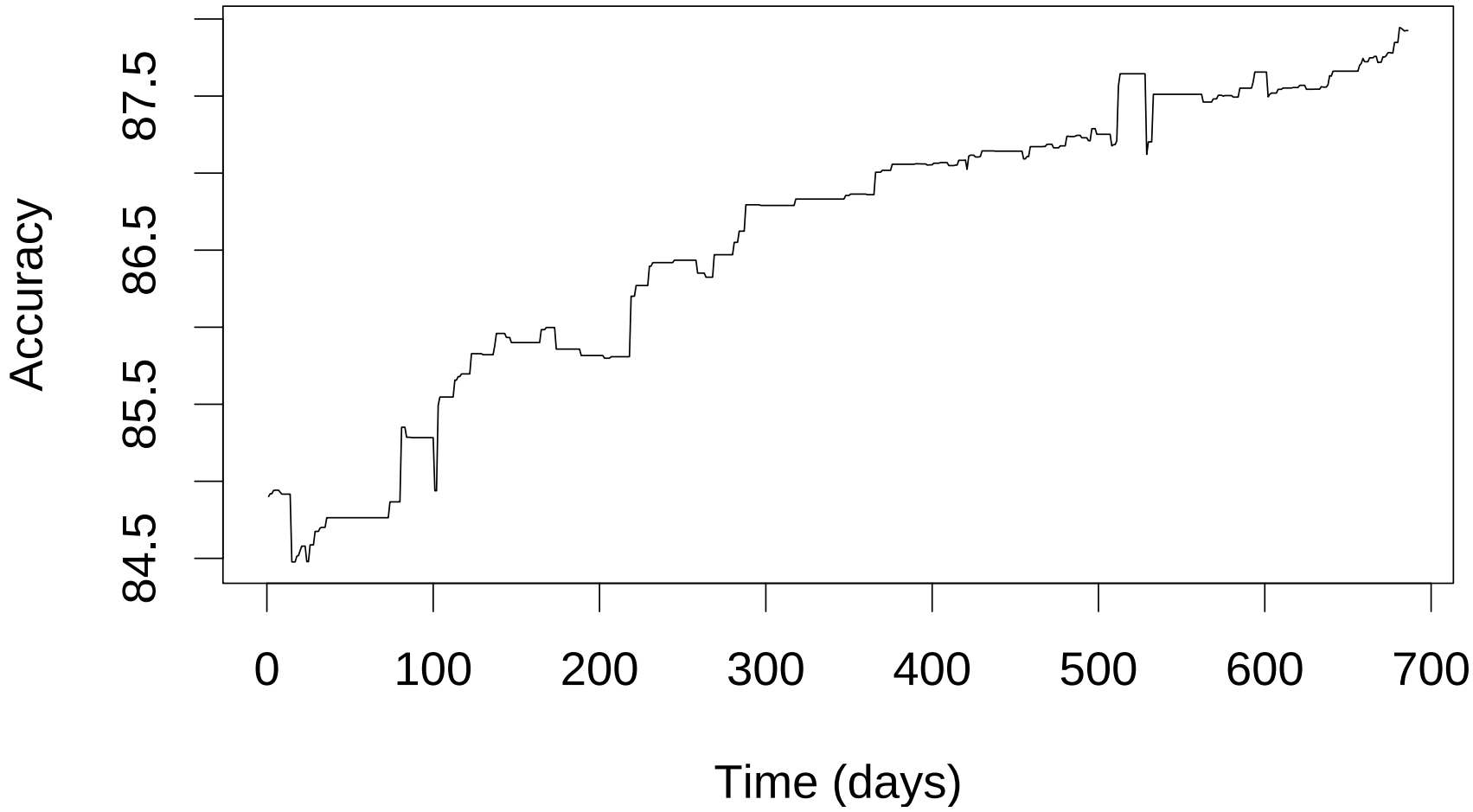
Background: Alpino

- Wide Coverage parsing of Dutch
- Large lexicon
- Constructionalist HPSG
- *Disambiguation:* Parse Selection with log-linear model
- *Efficiency:* HMM lexical analysis filter
- Constructs CGN Dependency Structures

Results

- Evaluation: named dependency triples

corpus	sents	length	CA%
Alpino treebank	7136	19.7	87.9
Trouw newspaper	500	17.0	88.8


Accuracy
84.5
85.5
86.5
87.5
0
100
200
300
400
500
600
700
Time (days)

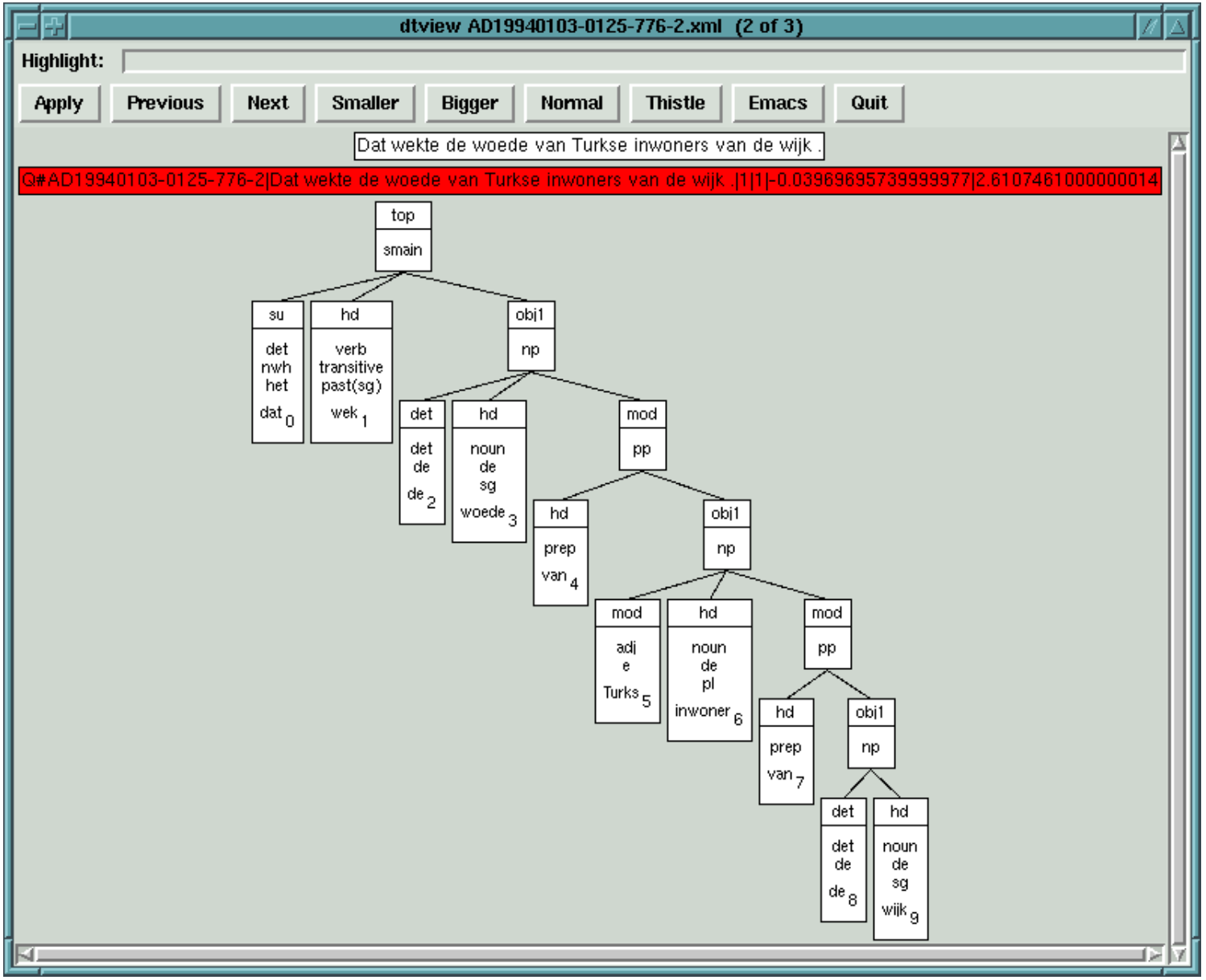
Machine-annotated Corpus Material

Four years of news-paper text (NRC and AD 1994, 1995)

Number of sentences	4,150,858
No parse	0.3%
Fragment parse	8.9%
Full parse	90.8%
CPU hours	20,000

Browse and Search Corpus Material

- Dependency structures coded in XML
- DtView for browsing dependency structures
- DtEdit for editing dependency structures
- DtSearch allows full XPATH queries
- Search patterns include hierarchical relations, grammatical relations, precedence relations, syntactic category, word (stem)


dtview AD19940103-0125-776-2.xml (2 of 3)
Highlight:
Apply
Previous
Next
Smaller
Bigger
Normal
Thistle
Emacs
Quit
Dat wekte de woede van Turkse inwoners van de wijk .
Q#AD19940103-0125-776-2|Dat wekte de woede van Turkse inwoners van de wijk .|1|1|-0.03969695739999977|2.6107461000000014
top
smain
su
det
nwh
het
dat 0
hd
verb
transitive
past(sg)
wek 1
obj1
np
det
det
de
de 2
hd
noun
de
sg
woede 3
mod
pp
hd
prep
van 4
obj1
np
mod
adj
e
Turks 5
hd
noun
de
pl
inwoner 6
mod
pp
hd
prep
van 7
obj1
np
det
det
de
de 8
hd
noun
de
sg
wijk 9

Example

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<top>
  <node rel="top" cat="smain" begin="0" end="10">
    <node rel="su" frame="determiner(het,nwh,nmod,pro,nparg)" pos="det" begin="0" end="1"
    <node rel="hd" frame="verb(hebben,past(sg),transitive)" pos="verb" begin="1" end="2"
    <node rel="obj1" cat="np" begin="2" end="10">
      <node rel="det" frame="determiner(de)" pos="det" begin="2" end="3" root="de" word="
      <node rel="hd" frame="noun(de,both,sg)" pos="noun" begin="3" end="4" root="woede" w
      <node rel="mod" cat="pp" begin="4" end="10">
        <node rel="hd" frame="preposition(van,[af,uit,vandaan,[af,aan]])" pos="prep" begi
        <node rel="obj1" cat="np" begin="5" end="10">
          <node rel="mod" frame="adjective(e,nonadv)" pos="adj" begin="5" end="6" root="I
          <node rel="hd" frame="noun(de,count,pl)" pos="noun" begin="6" end="7" root="inw
          <node rel="mod" cat="pp" begin="7" end="10">
            <node rel="hd" frame="preposition(van,[af,uit,vandaan,[af,aan]])" pos="prep"
            <node rel="obj1" cat="np" begin="8" end="10">
              <node rel="det" frame="determiner(de)" pos="det" begin="8" end="9" root="de
              <node rel="hd" frame="noun(de,count,sg)" pos="noun" begin="9" end="10" root
            </node>
```

```
          </node>
        </node>
      </node>
    </node>
  </node>
  <sentence>Dat wekte de woede van Turkse inwoners van de wijk .</sentence>
  <comments>
    <comment>Q#AD19940103-0125-776-2|Dat wekte de woede van Turkse inwoners van de wijk .
  </comments>
</top>
```

Applications in Corpus Linguistics

Impersonal pronouns

- *dat*, *die*, *deze*, *dit*, *het* do not occur after a preposition
- **met dat/die*
  daarmee
- **met deze/dit*
  hiermee
- **met het*
  ermee

DtSearch: impersonal pronouns

- `//node[@cat="pp" and ./node[@word="dat"]]`
- Identifies many false parses

```
apr1723.xml        [In dat] geval zou op z'n minst moeten ..
apr1749.xml        Daar staat [tegenover dat] het ..
...
```

- 2310 het, 61 dit, 304 dat, 310 die, 345 deze

Counter examples

- 'deze' is sometimes possible:

```
jan16909.xml     [Bij deze] wens ik mijn familie , vrienden en ...
jan111135.xml    De schaatsers kennen mijn gestoordheid [in deze] . ’
jan113787.xml    Met passages [als deze] bewijst Boskma dat ...
```

- in addition: 558 examples where pronoun is modified:

```
apr1411.xml      De visie van de advocaat-generaal staat daarmee
                 lijnrecht [tegenover die van de Hoge Raad] .
apr13294.xml     Zijn tijden op de 10.000 meter lagen net
                 [boven die van het toenmalige wereldrecord] .
apr18677.xml     Het niveau van de ruwe olie komt net niet
                 [boven dat van het zeewater] buiten uit
...
```

Impersonal Pronouns

- Alpino now distinguishes impersonal pronouns
- Lexical impersonal pronouns cannot combine with a preposition
- ⋆ prevents many false parses
  ⋆ reduces ambiguity in other cases

Weak Pronouns

- Weak/minor pronouns: *we*, *me*, *je*, *ze* . . . weak variant of *wij*, *mij*, *jij/jou*, *zij* . . . (van Eynde 1996)
- Alpino did not distinguish weak pronouns
- Investigate various properties of weak pronouns:
  - ⋆ Coordination (**hem en me*)
  - ⋆ Modification (**je met je gezeur*)
  - ⋆ Comparatives (**groter dan je*)
  - ⋆ Topicalization (**je heb ik gezien*)

Weak Pronouns (2)

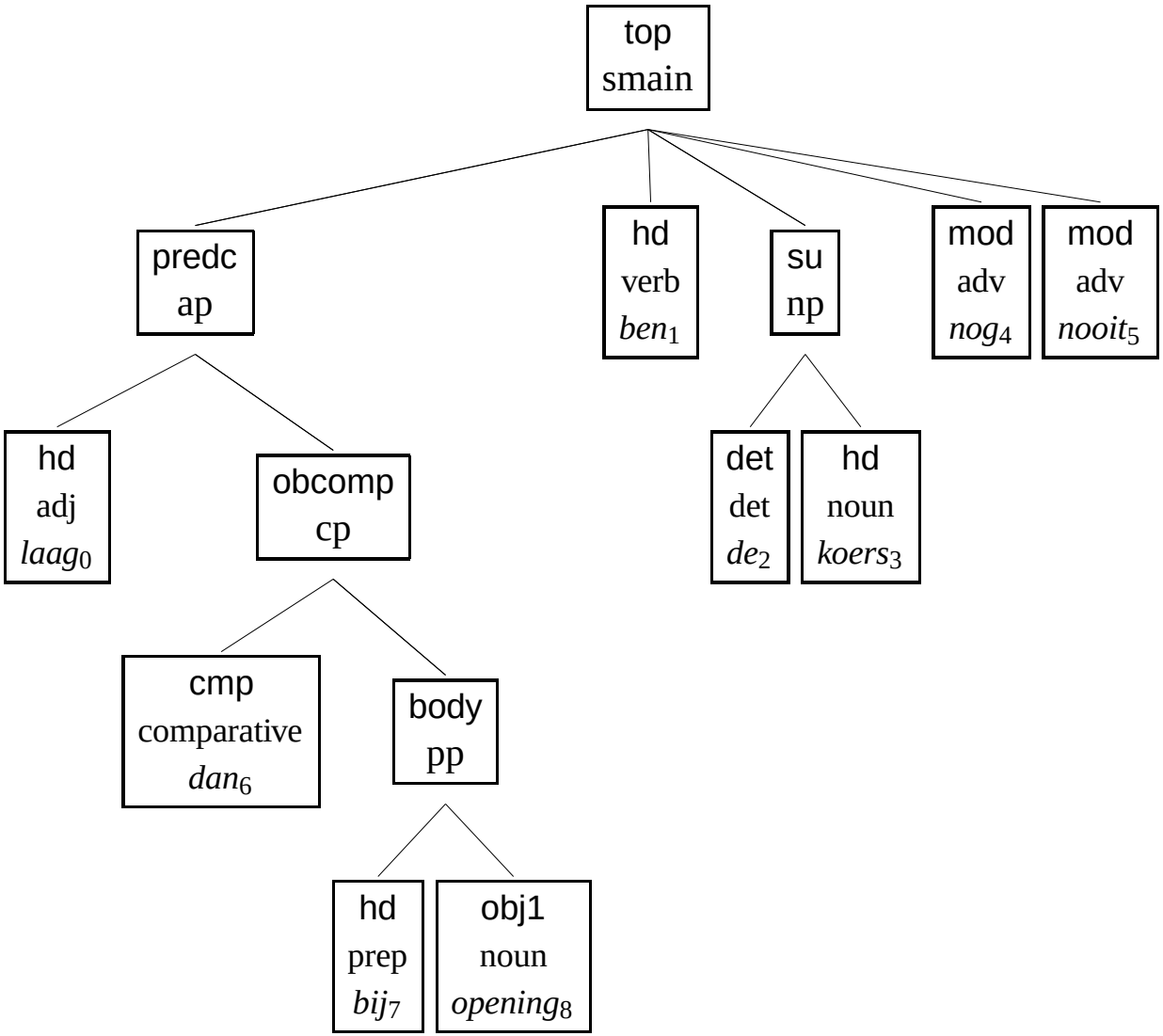
- Coordination: 1029 occurrences, only one correct:
- `Merk op dat met ’ [we] ’ en ’ man ’ , niet de islamitische man kan worden bedoeld`
- Similar results for modification (4447), comparatives (208) and topicalization (600)
- 3 exceptions:

```
apr35790.xml     [Me] dunkt dat het niet onredelijk is dat ...
feb4326.xml      [Me] dunkt , hoe minder men overkomt zoals ...
maa41420.xml     [Me] dunkt dat het voor de UEFA geen gekke gedachte ...
```

Extraposition of comparatives out of topic

- Reviewer: extraposition of comparative out of topic is impossible: **Lager was de koers nog nooit dan bij opening*
- Alpino grammar allows this
- We can search for the relevant pattern

Dependency Structure



DtSearch queries

```
//node[@cat="smain" and
       ./node[./node[@rel="obcomp"]]/@begin = @begin]'

//node[@cat="smain" and
       ./node[./node[@rel="obcomp"]
            /@end > ../node[@rel="hd"]/@begin
                                  ]/@begin = @begin]
```

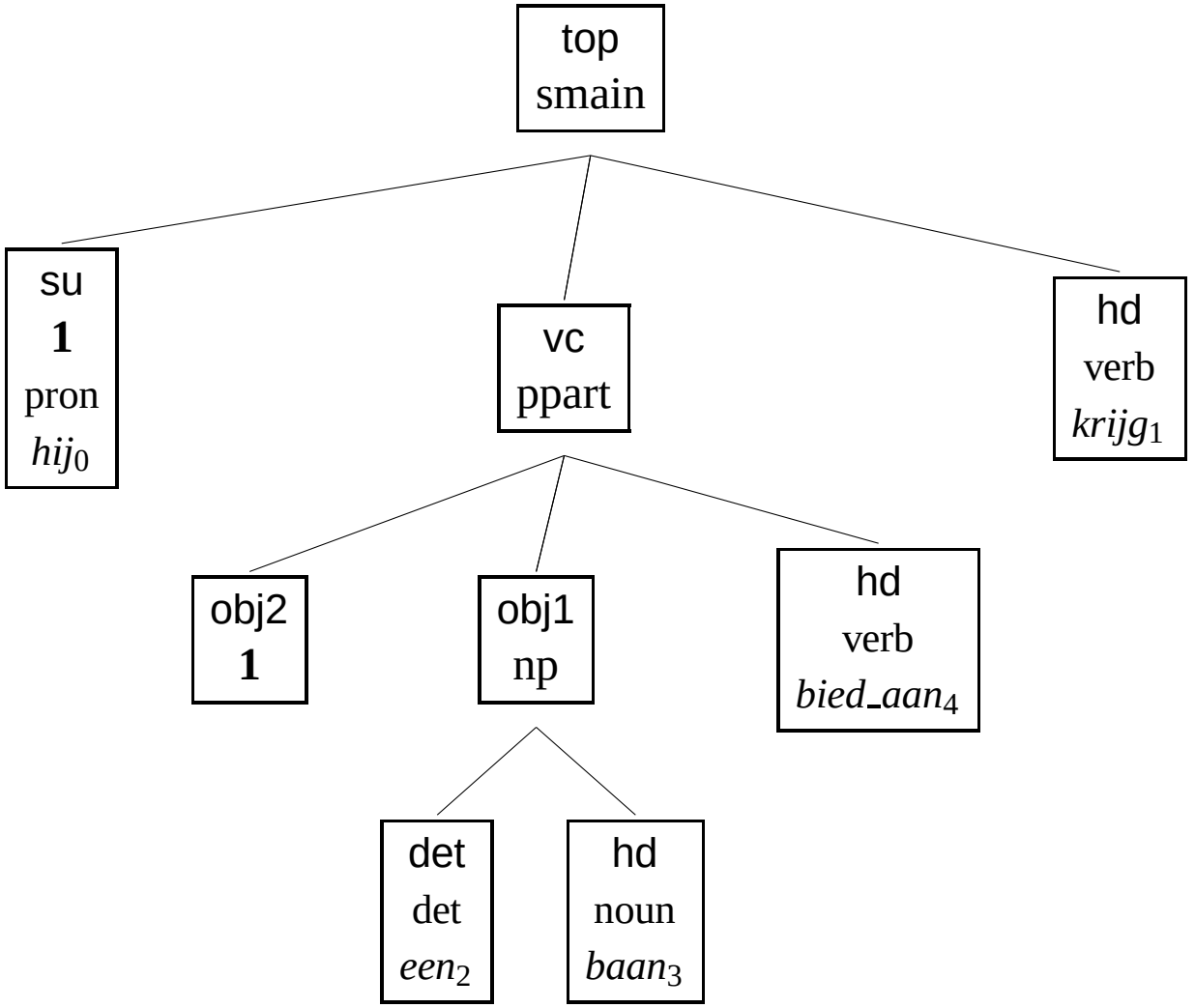
Extraposed obcomp out of topic

Liever benadrukt hij die tegenstellingen dan de bedriegelijke harmonie
Nog eerder zal de machtige Mekong droogvallen dan dat de co-premier
zijn macht uit handen geeft
Zo intens lelijk zijn mijn voeten in de loop van een decennium
geworden dat ik de mensenmassa's op het strand er in de zomer
niet mee wil lastigvallen
Eerder brengt men een hemel vol wolken in kaart dan dit oeuvre
Veel eerder vindt er een herschikking in het midden plaats dan dat er
werkelijk massaal uit dat midden wordt gevlucht
Eerder is er sprake van het kabinet ' ondanks Kok ' dan het '
kabinet-Kok
Liever sluis ik honderden en honderden guldens door aan loodgieter ,
fietsenmaker en elektricien dan dat ik zelf ook maar n vinger
uitsteek naar het fonteintje bij het toilet , een kapot achterlicht of
een weigerende stofzuiger
liever waren ze onafhankelijk dan dat ze zich aan iemand bonden
Liever is Jim schuldig aan een sprong , dan de prooi van een
aanvechting

eerder gaat zoo’n kameel door het oog van een naald, dan dat een rijke
in zou gaan in het koninkrijk der hemelen

krijgen-passive

- Wij bieden hem een baan aan
- Hij krijgt een baan aangeboden


top
smain
su
1
pron
hij0
vc
ppart
hd
verb
krijg1
obj2
1
obj1
np
hd
verb
bied_aan4
det
det
een2
hd
noun
baan3

First ten examples

Symor kreeg zaterdag van Van Thijn de Zilveren Erepenning voor menslievend hulpbetoon uitgereikt , een onderscheiding die bij Koninklijk Besluit wordt toegekend aan mensen die zich hebben onderscheiden door moed , beleid en zelfopoffering .

Hij kreeg ook maar gedeeltelijk wachtgeld toegekend .

Goossens zegt eraan gewend te zijn dat hij de zwartepiet krijgt toegespeeld .

Marco van Basten krijgt ondanks langdurige arbeidsongeschiktheid zijn volle salaris uitbetaald .

Zelf hoor ik ze soms spreken over incompetent lesgeven , als ze vlak voor het schoolonderzoek een videofilm krijgen voorgeschoteld .

Het kabinet heeft gisteren besloten dat de slachtoffers van de watersnood in Limburg een groot deel van hun schade vergoed krijgen .

Slachtoffers van de overstroming in Limburg krijgen alleen van de schade die niet te verzekeren was een gedeelte vergoed .

Ze krijgen toch een bepaald stempel opgedrukt , waardoor de kans op een baan of woning niet groot is .

Het schoolonderzoek moet zodanig zijn dat iedere eindexamenkandidaat hetzelfde werk krijgt voorgeschoteld , dat er niet met cijfers valt te sjoemelen en dat de docent min of meer door de mand valt als het gros van zijn leerlingen slechte resultaten behaalt .

Volgens de kustwacht kreeg de schipper van de Gunnar Langva al in een vroeg stadium een standaard-bergingscontract van Smit Lloyd aangeboden , maar de man weigerde dat .

More: Agreement

- *het hemd* and *iemand* **de** *hemd van het lijf vragen*
- *een groot/beroemd/gevierd man/gitarist/. . .* but not *een groot vrouw*
- *de mens* and *dat mens*
- *de keer* and *dit keer* but not *dat keer* (idem for *maal*)
- . . . . . .

More: Active sentences without subject

- topic-drop: *kan wel*
- surprise: *aldus/zo geschiedde*

More: Relative clauses with *die* without explicit antecedent

- er zijn er die dit afkeuren
- je hebt er ook die dit afkeuren