ALPINO: Automatisch Ontleden van het Nederlands

Alfa-informatica/Informatiekunde, Rijksuniversiteit Groningen
Gosse Bouma, gosse@let.rug.nl
Gertjan van Noord, vannoord@let.rug.nl

NWO Netherlands Organisation for Scientific Research

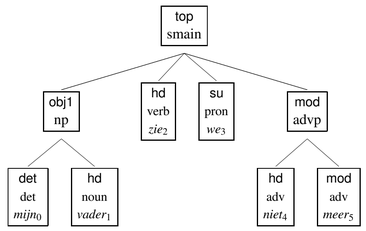
RuG

Alpino

Alpino is een automatische ontleder (parser) voor het Nederlands. Het systeem bestaat uit de volgende onderdelen:

- HPSG grammatica voor het Nederlands
- Uitgebreid woordenboek ($> 100,000$ ingangen)
- POS-tagger voor efficiëntie
- Desambiguatie met Maximum Entropy model
- Construeert CGN dependentiestructuren

Mijn vader zagen we niet meer



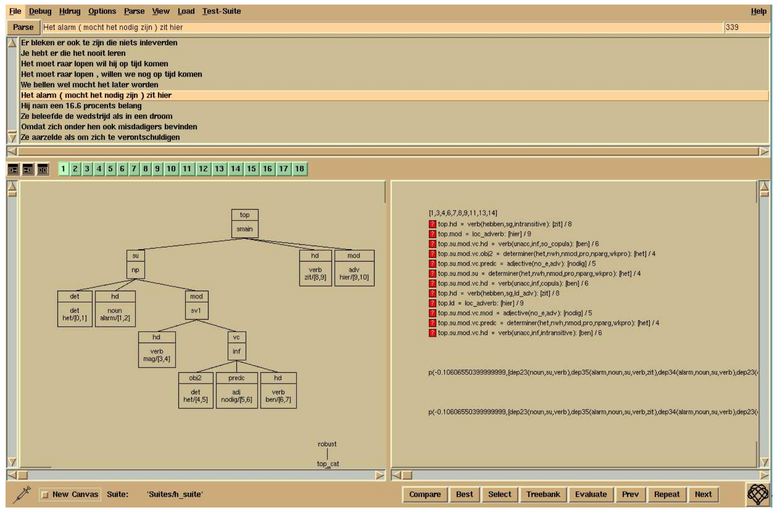
Hoe goed werkt Alpino?

Corpus	Accuratesse (%)
Alpino cdbl eindhoven	85.9
TwNC trouw2001 500	88.0
CLEF03 vragen	92.6
CLEF04 vragen	91.4

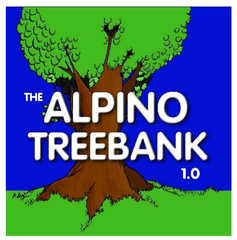
Syntactische Annotatie met Alpino

Alpino kan worden gebruikt voor semi-automatische syntactische annotatie. Hierbij worden verschillende tools beschikbaar gesteld:

- parse selectie tool
- pos-tag selectie tool
- haakjes in de input
- bekijken en bewerken van dependentiestructuren (THISTLE)
- zoeken naar syntactische patronen (XPATH)



De Alpino treebank bestaat uit onder andere de syntactische annotatie van al de 7154 zinnen uit het cdbl deel van het Eindhoven corpus. De Alpino Treebank verscheen op CDROM ter gelegenheid van CLIN 2002 in Groningen. De inmiddels flink verbeterde treebank is raadpleegbaar via `http://www.let.rug.nl/~vannoord/trees/`



Alpino en Question Answering

IMIX

Q Welke wiskundige bewees de Stelling van Fermat?
A [NRC 18/05/1995] Afgelopen maandagavond heeft de wiskundige Andrew Wiles in het McCoshAuditorium van Princeton University in een lezing voor 600 mensen bekend gemaakt dat zijn bewijs van de Laatste Stelling van Fermat, dat door een groot aantal specialisten is beoordeeld , zal worden gepubliceerd in het vooraanstaande tijdschrift Annals of Mathematics.

- Classificatie van vragen op basis van syntactische analyse en Nederlandse deel van EuroWordNet
- Integratie van een Named Entity Classifier in Alpino
- Volledige CLEF corpus syntactisch geanalyseerd
- Extractie van antwoorden op standaardvragen (hoofdstad, inwoneraantal, munteenheid, voorzitter/president/trainer van Organisatie, ...) op basis van syntactische analyse
- Identificatie van Antwoorden op basis van syntactische analyse

Alpino en Corpustaalkunde

De kwaliteit van Alpino is goed genoeg om door Alpino syntactisch geannoteerde corpora (dus niet gecorrigeerd!) te gebruiken voor corpustaalkundig onderzoek.

- Villada: herkennen en classificeren van vaste verbindingen
- Bouma: vooropplaatsing van voorzetselgroepen
- van der Beek: volgorde van NPs bij ditransititieven en AcI-constructies

Voorbeeld: kunnen focuspartikels in een voorzetselgroep optreden? De volgende XPATH query vindt voorkomens in de treebank:

```
'//node[@cat="pp" and ./node[@cat="np"]/node[\
    @root="alleen" or @root="zelfs" or @root="ook"]\
    /@begin = ./node[@pos="prep"]/@end]'
```

..., maar toch met alleen het hoognodige .
Het jacht met alleen vrouwen ,...
...met alleen een bogey op de zesde .
...met ook een scheef oogje naar de taal van Mann
...tegen zelfs de voorzichtigste aanzetten tot vermindering van het aantal parkeerplaatsen
...zou moeten rondkomen van alleen de aow
Je kunt natuurlijk kiezen voor alleen uitgebreide samenvattingen

Het CLEF corpus bestaat uit de volledige jaargangen 1994 en 1995 van het Algemeen Dagblad en het NRC Handelsblad.

Aantal zinnen	4,150,858
Geen parse	0.3%
Fragment parse	8.9%
Volledige parse	90.8%
CPU hours	20,000

Meer weten?

`http://www.let.rug.nl/~vannoord/alp/`