

Rich Statistical Parsing and Literary Language

Andreas van Cranenburgh

De dag die je wist dat
zou komen. . .

229 enerverende
pagina's

Niet bekend van TV.
Geen pageturner.

Overzicht

- ▶ Ontleden: zinsbouw automatisch analyseren
- ▶ Literatuur: tekstuele correlaten van literariteit?

ONTLEDEN

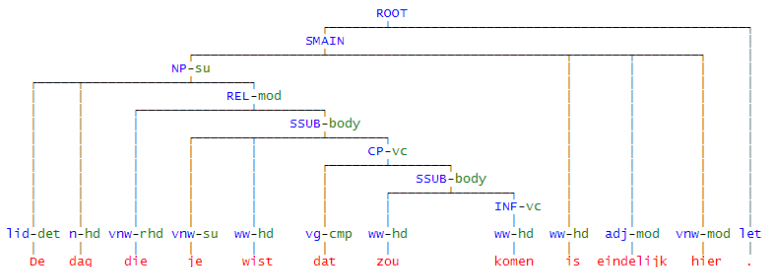
Ontleden

Analyseren van zinsbouw

- ▶ Grammatica automatisch afgeleid uit teksten
- ▶ Frazestructuren, desambiguatie
- ▶ Niet-lokale afhankelijkheden, grammaticale relaties

Data-Oriented Parsing demo

Enter a sentence in Dutch, English, German, or French (auto-detected). The sentence will be parsed and the most probable parse tree will be shown ([show technical details](#)).



([show fragments](#); [show alternative analyses](#); [show info](#); [link](#))

Sentence:

detect language ▼

MPP ▼

RFE ▼

n-best ▼

CKY ▼

Parse

L I T E R A T U U R

De onderzoeksvraag

Kan een computermodeel literatuur herkennen?

Het Nationale Lezersonderzoek

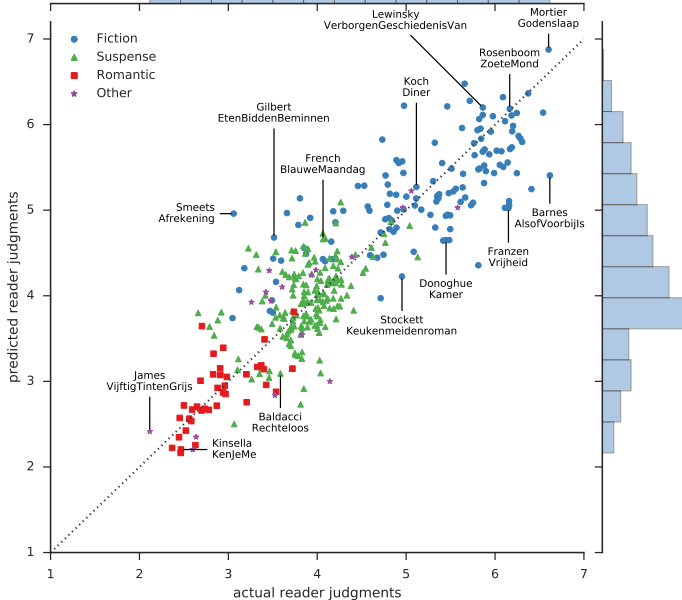
- ▶ Vragenlijst, 401 recente romans (Nederlands/vertaald)
- ▶ Beoordeling: gegeven titels van gelezen boeken, hoe literair/goed op een schaal van 1-7?
- ▶ 13,000 deelnemers

Tekstkenmerken

- ▶ Woordenschat, zinslengte, &c.
- ▶ Cliché's: Wie denk je wel dat je bent!, Hoe moeilijk kan het zijn?
- ▶ Thema's: familie, politieonderzoek, uiterlijk/uitgaan
- ▶ Woordgebruik: de oorlog, een boek, mobiele telefoon
- ▶ Zinssneden, constructies

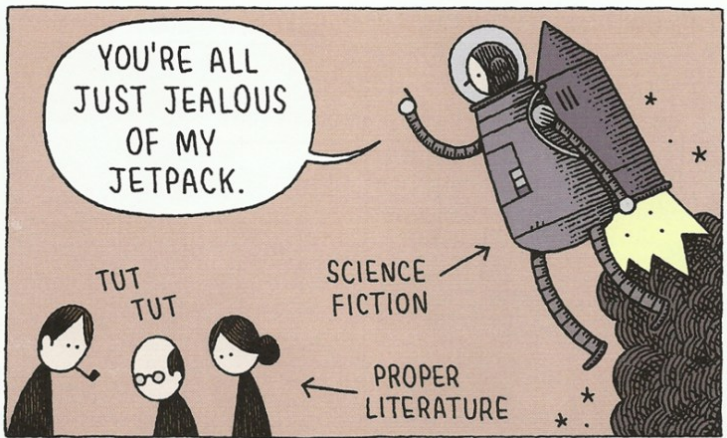
Resultaten

	R^2
mean sent. len.	16.4
+ % direct speech sentences	22.9
+ top3000vocab	23.5
+ bzip2_ratio	24.4
+ cliches	29.9
+ topics	52.2
+ bigrams	58.2
+ % modifying PPS	58.6
+ avg. dependency length	57.1
+ fragments	59.7
+ Genre	74.0
+ Translated	73.8
+ Author gender	76.1



Bevindingen

- ▶ **Ja**, computer herkent literatuur op basis van tekst
- ▶ **Hoe**: optelsom groot aantal factoren
- ▶ **Hypothese**: Rijk, gestileerd taalgebruik



Credit: Tom Gauld