Natuurlijke-taalverwerking 1

Daniël de Kok

Natuurlijke-Taalverwerking

Het college Natuurlijke-taalverwerking is een inleiding in de computationele taalkunde en maakt deel uit van het curriculum van Informatiekunde en Kunstmatige Intelligentie. Aan de hand van enige toepassingen leggen we uit wat de belangrijkste onderzoeksvragen van de computationele taalkunde zijn. Vervolgens richten we ons het ontwikkelen van computationele grammatica's en op het automatisch verwerken (ontleden) van natuurlijke taal met behulp van zulke grammatica's.

Vandaag

- Praktische zaken
- College-overzicht
- Wat is natuurlijke-taalverwerking?
- Context-vrije grammatica's

Overzicht

Praktisch

Natuurlijke-taalverwerking

Context-vrije grammatica's

Studiehandleiding

- Zie `http://www.let.rug.nl/~dekok/ntv/`
- Links naar de syllabus, aanvullende literatuur, college-aantekeningen, practicumopdrachten
- Practicum start komende week
- Beoordeling
  - Practicum (5 opdrachten, 50%)
  - Tentamen (50%)
  - Beide onderdelen moeten voldoende zijn
  - Te laat ingeleverd is niet ingeleverd
- Eregalerij

Praktisch

- Hoorcolleges: Daniël de Kok (d.j.a.de.kok@rug.nl)
- Practica: Jelmer van der Linde
- Mail gerust voor praktische vragen, vragen over de opdrachten gelieve stellen tijdens de practica
- Practica vinden plaats in 12 0119 op maandag 11-13.

Literatuur

- Syllabus Natuurlijke-taalverwerking 1 (pdf, pdf, 2 blz/A4).
- Learn Prolog Now! (Blackburn, Bos, Striegnitz), Chapters 7 and 8 (Definite Clause Grammars)
- Chart Generation, Martin Kay, Proceedings of the 34th annual meeting on Association for Computational Linguistics
- Statistical Methods and Linguistics, Steven Abney, In: Judith Klavans and Philip Resnik (eds.), The Balancing Act: Combining Symbolic and Statistical Approaches to Language. The MIT Press, Cambridge, MA. 1996

Optionele literatuur

- Prolog and Natural Language Processing, Pereira en Shieber, 1988/2002
- An Introduction to Unification-Based Approaches to Grammar, Shieber, 1986/1988/2003

Plagiaat

- Tijdens de werkcolleges van dit vak is overleg met medestudenten toegestaan.
- **Maar:** ingeleverde opdrachten moeten altijd en volledig eigen werk zijn.
- Eigen inspanning is leerzamer.
- Uiteraard wordt gebruik van standaardpredikaten in Prolog aangemoedigd!

Wat leer je dit college?

- Hoe je een computergrammatica schrijft.
- Hoe je met behulp van een grammatica zinnen kunt ontleden.
- Hoe je met een grammaticale analyse feiten uit zinnen kunt extraheren.
- Hoe je met een grammaticale analyse vragen kunt beantwoorden.
- Hoe je met een grammatica zinnen kunt bouwen (genereren).
- Hoe je een grammaticale analyse of gegenereerde zin kunt beoordelen.

College-overzicht

- Week 1: Context-vrije grammatica's
- Week 2: Definite Clause Grammar (DCG)
- Week 3: Unificatie-grammatica I
- Week 4: Unificatie-grammatica II
- Week 5: Automatisch Ontleden
- Week 6: Parse selectie
- Week 7: Generatie

Overzicht

Praktisch

Natuurlijke-taalverwerking

Context-vrije grammatica's

Wat is natuurlijke-taalverwerking?

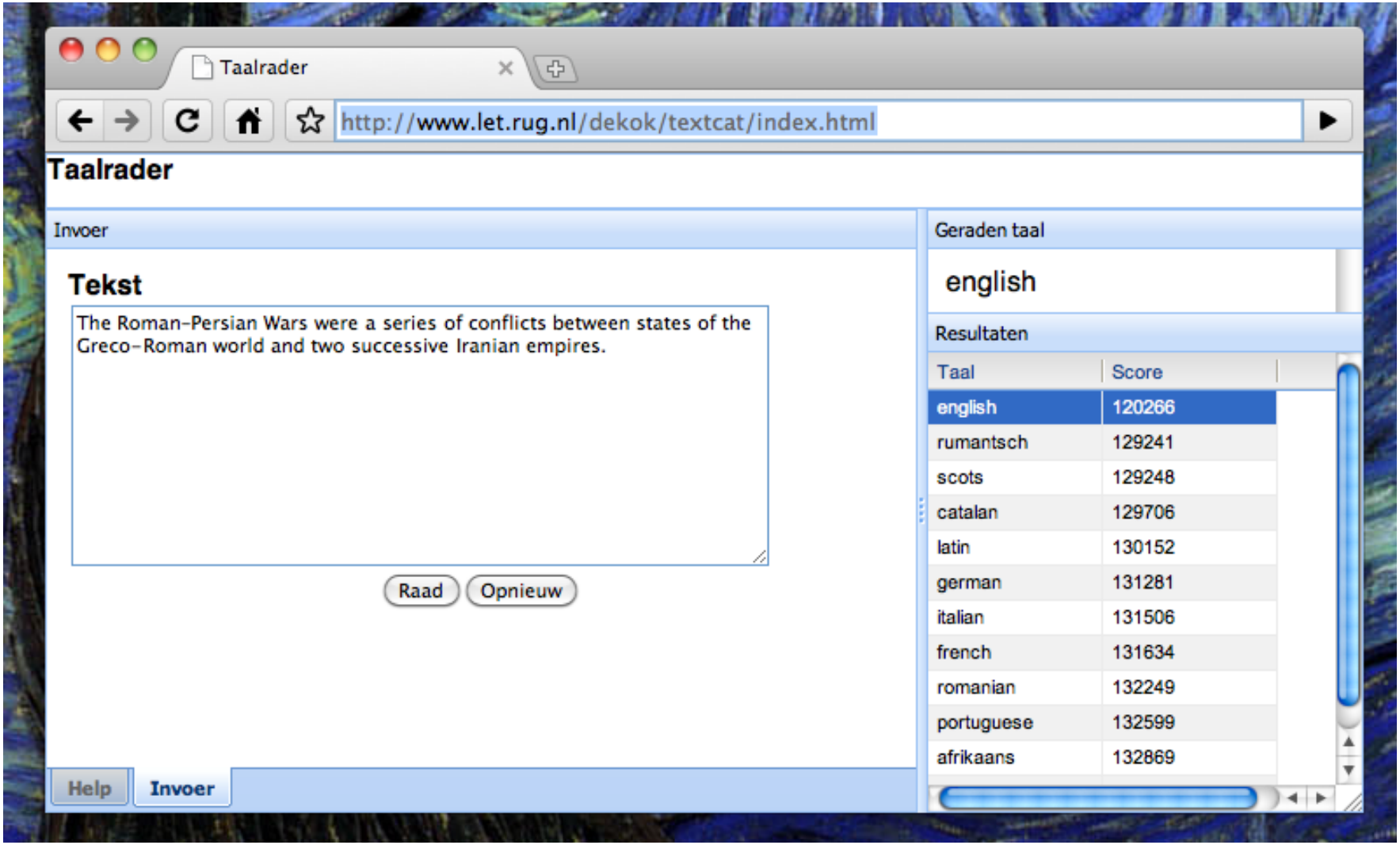
Het ontwikkelen van programma's en toepassingen waarbij *kennis van de structuur en de betekenis van natuurlijke taal* een rol speelt.

Voorbeelden

- spellingscontrole;
- grammatica-correctie (identificeren en corrigeren van grammaticale fouten in tekst);
- automatisch vertalen;
- automatisch vragen beantwoorden;
- automatisch e-mail beantwoorden;
- informatie extractie;
- zoeken

Spellingscontrole

Taalrader


Taalrader
http://www.let.rug.nl/dekok/textcat/index.html
Taalrader
Invoer
Tekst
The Roman-Persian Wars were a series of conflicts between states of the Greco-Roman world and two successive Iranian empires.
Raad
Opnieuw
Help
Invoer
Geraden taal
english
Resultaten
Taal
Score
english 120266
rumantsch 129241
scots 129248
catalan 129706
latin 130152
german 131281
italian 131506
french 131634
romanian 132249
portuguese 132599
afrikaans 132869

Automatisch vragen beantwoorden

Automatisch vragen beantwoorden (2 voor 12)

- `Wie is de voorzitter van het Europese Parlement?`
- **Klaus Hänsch** , voorzitter van het Europese Parlement , drukte het iets sterker uit...
- Ook de voorzitter van de CDA-delegatie in het Europese Parlement , oud-minister **Maij** ....

Automatisch vragen beantwoorden

- `Wanneer vond de Duitse hereniging plaats?`
- Sinds de Duitse hereniging **in oktober 1990** is de sterfte in Oost-Duitsland sterk toegenomen.
- Al **in 1962** voorspelde hij de Duitse hereniging en het uiteenvallen van de Sovjet-Unie.

Zoeken

- Tekst tokiniseren.
- Opsporen lemma's, andere schrijfwijzen.
- Herkennen van bijvoorbeeld namen en vaste uitdrukkingen.
- Afhandelen spelfouten in de tekst of query.

Makkelijk en moeilijk

Makkelijk?	Moeilijk?
Spellingcontrole	Grammaticale controle
Voice Response systemen	Volledige spraakherkenning
Rapporten genereren uit tabellen	Samenvatten van artikelen
Vertaalhulp	Automatisch vertalen
Domein-specifieke dialoogsystemen	Turing-test
Web-search	Vragen beantwoorden

Zelfs makkelijke toepassingen zijn moeilijk

- Spellingcorrectie:
  - **Lijkt gemakkelijk**: markeer alle woorden die niet in het woordenboek staan,
  - Maar **is moeilijk**: geen woordenboek is volledig, iedere dag worden nieuwe woorden geïntroduceerd.

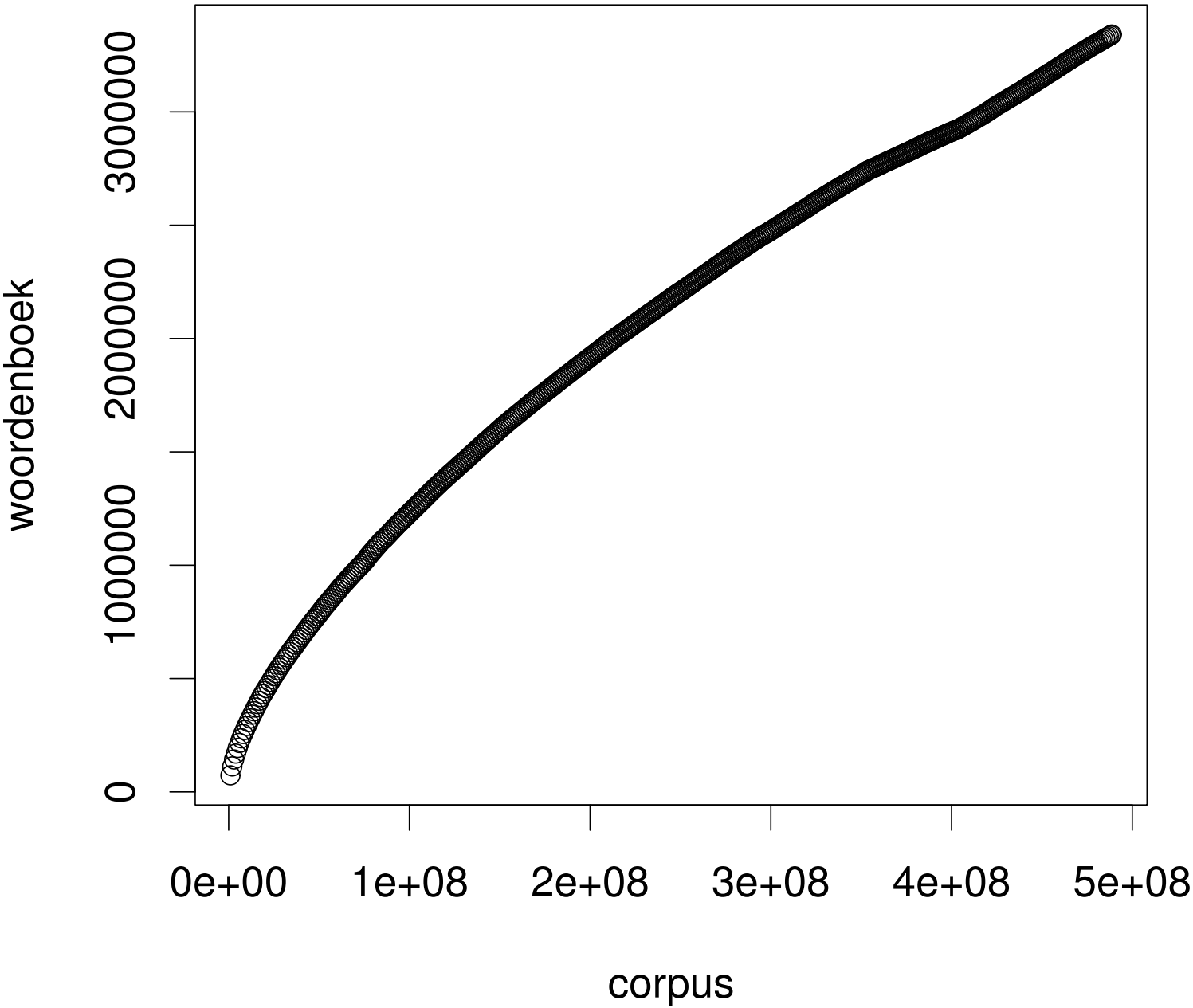
Omvang van een woordenboek

- 125K (*Groene Boekje*)
- 500K+ (*van Dale*).
- Soms ontbreekt 40% van de **woordtypes** in een tekst in het woordenboek.
  - **Tokens:** aantal **woorden** in een tekst,
  - **Types:** aantal **verschillende woorden** in een tekst.

Meer cijfers

- Kun je een goede woordenlijst afleiden uit een corpus (verzameling tekst):

	tokens	types
1 krant	36K	8K
1 maand	1556K	96K
1 jaar	18M	410K
TwNC	488M	3340K
DCOI	54M	1220K


woordenboek
0
1000000
2000000
3000000
0e+00
1e+08
2e+08
3e+08
4e+08
5e+08
corpus

Spellingcorrectie $\neq$ opzoeken

- Syntactische context:
  - Wat gebeurd/t er?
  - Wat is er gebeurd/t?

Spellingcorrectie $\neq$ opzoeken

- Deze jongen **vind(t)** je aardig.
  - (Daarom wil hij een date.)
  - (Daarom wil jij een date.)
- Wel/geen spelfout hangt af van *betekenis*.

Brandt Corstius

- De derde wet van de computer-taalkunde:
  - Na een bepaalde tijd, bv 1 jaar, werken, krijg je 80% goede resultaten. Elke halvering van de *gap* tussen 80 en 100% betekent een vermenigvuldiging van de aanvankelijk bestede tijd met een vaste factor die groter is dan 1.
- Wat je ook doet, de semantiek gooit roet.

Bijdrage vakgebieden

- *Computationele taalkunde:* formalismes voor computationele grammatica's, corpustaalkunde, computationele semantiek.
- *Informatica:* ontleedalgoritmen, finite state technieken, (logische) programmeertalen.
- *Kunstmatige intelligentie:* toepassing in robots, machine learning technieken.
- *Informatiekunde:* slim coderen van data, integratie in toepassingen, interfaces.

Overzicht

Praktisch

Natuurlijke-taalverwerking

Context-vrije grammatica's

Grammatica

- Veel toepassingen vereisen een kennis van de structuur van zinnen (zinsbouw, grammatica):
  - Grammatica-correctie (*jan word ziek*);
  - automatisch vertalen;
  - generatie van zinnen;
  - ....

Grammatica

- Een taal is een (oneindige) verzameling zinnen;
- zinnen zijn reeksen woorden;
- niet alle reeksen woorden zijn zinnen;
- een grammatica beschrijft:
  - Welke reeksen woorden goede zinnen vormen;
  - en wat de structuur van die reeksen is.

Herschijfregels

a $\rightarrow$ b

a kan herschreven worden als *b*

Bijvoorbeeld:

- NP $\rightarrow$ Det N
- N $\rightarrow$ eendje
- Det $\rightarrow \epsilon$

Conventies:

- *Terminals* met kleine letters;
- *non-terminals* met hoofdletters;
- ϵ voor de lege string;

Herschrijfregels (2)

- Een grammatica bestaat uit een verzameling van herschrijfregels.
- Een reeks van terminals (een zin) behoort tot de taal van een grammatica als het S symbool te herschrijven is tot die reeks van terminals.

Context-vrije grammatica's

Verzameling van regels met de volgende vorm:

$V \rightarrow w$

- V is één non-terminal symbool;
- w is reeks van terminal, non-terminal of epsilon symbolen.

Zoek de CFG regels

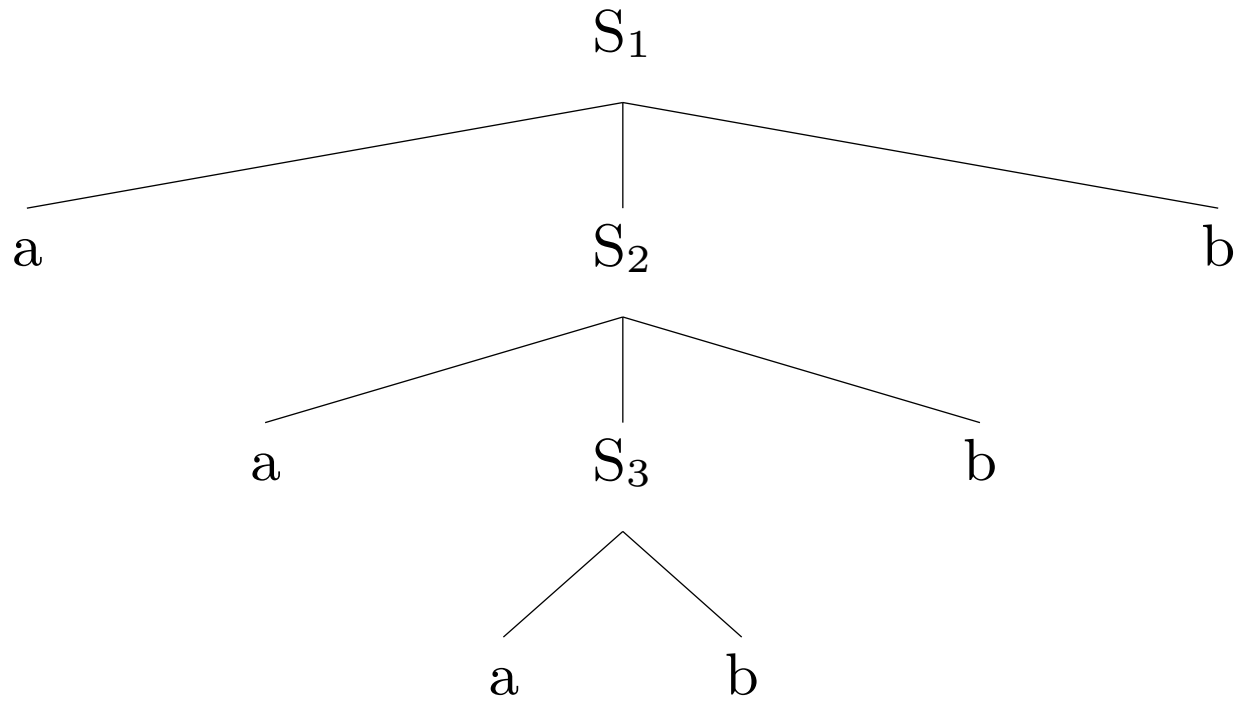
1. A $\rightarrow \epsilon$
2. a B $\rightarrow$ c
3. A B $\rightarrow$ C
4. C $\rightarrow$ D E
5. D $\rightarrow$ d

Een CFG voor $a^n b^n$, $n > 0$

$S \rightarrow a\ S\ b$
$S \rightarrow a\ b$

Ontleding *aaabbb*

S
a S b
a a S b b
a a a b b b



Bedenk een grammatica voor $a^n b^n$, $n \geq 0$

Ter inspiratie, de grammatica voor $a^n b^n$, $n > 0$:

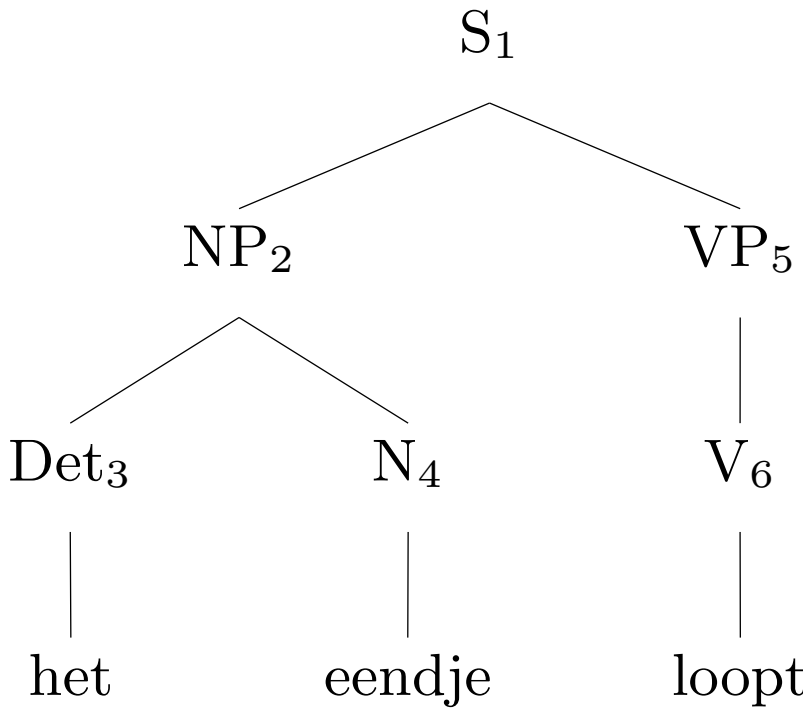
S $\rightarrow$ a S b
S $\rightarrow$ a b

Een kleine CFG voor het Nederlands

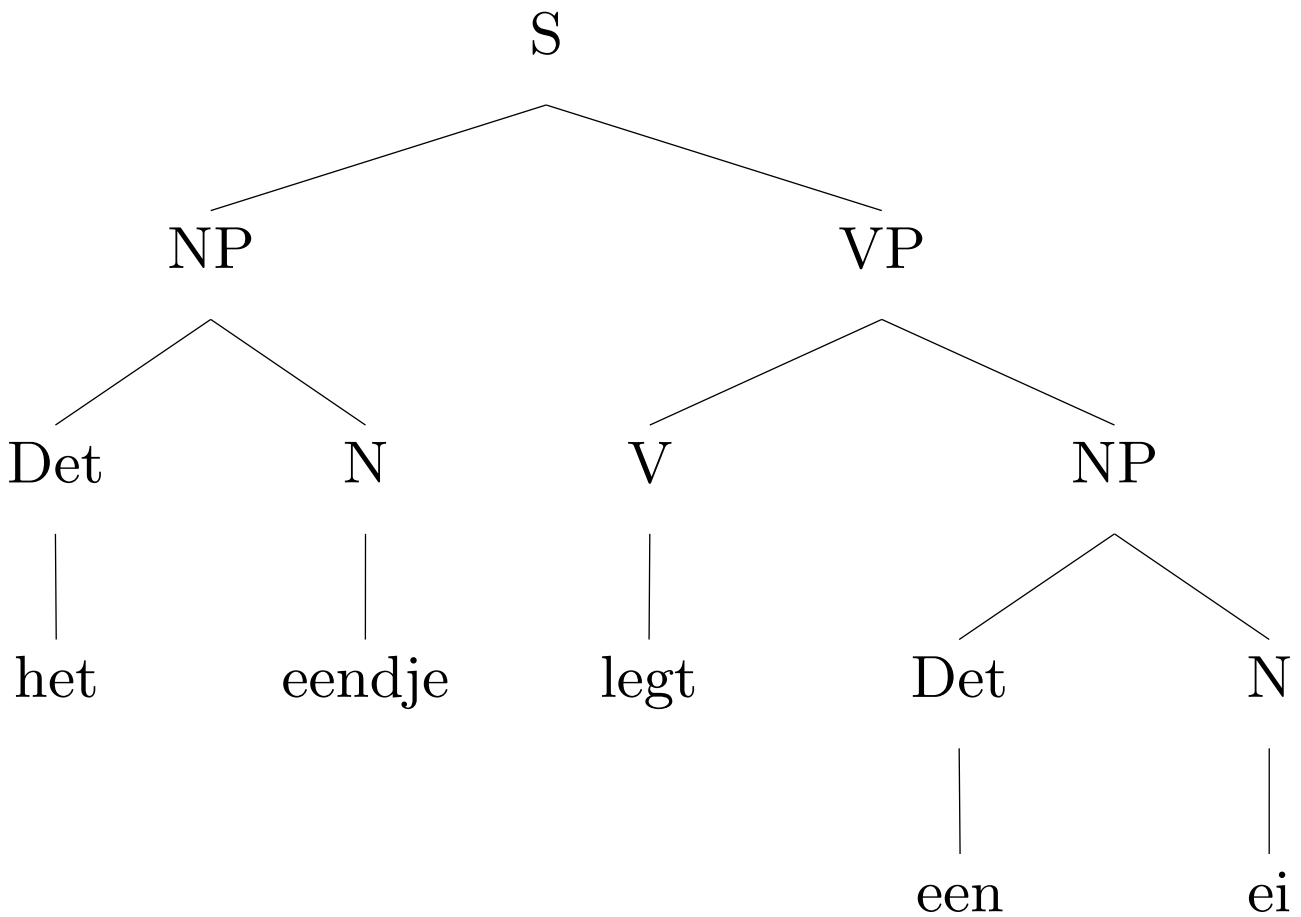
S	$\rightarrow$	NP VP	Det	$\rightarrow$	*een*
NP	$\rightarrow$	Det N	Det	$\rightarrow$	*het*
N	$\rightarrow$	A N	N	$\rightarrow$	*eendje*
VP	$\rightarrow$	V	N	$\rightarrow$	*ei*
VP	$\rightarrow$	V NP	V	$\rightarrow$	*legt*
			V	$\rightarrow$	*loopt*

Ontleding *het eendje loopt*

S
NP VP
Det N VP
het N VP
het eendje VP
het eendje V
het eendje loopt



Welke grammaticaregels zijn gebruikt?



Taal en grammatica

- Een *reeks* woorden W wordt *herkend* door grammatica G, wanneer je, door S te herschrijven, W kunt genereren;
- alle reeksen die door G worden herkend, vormen de *taal* van G.

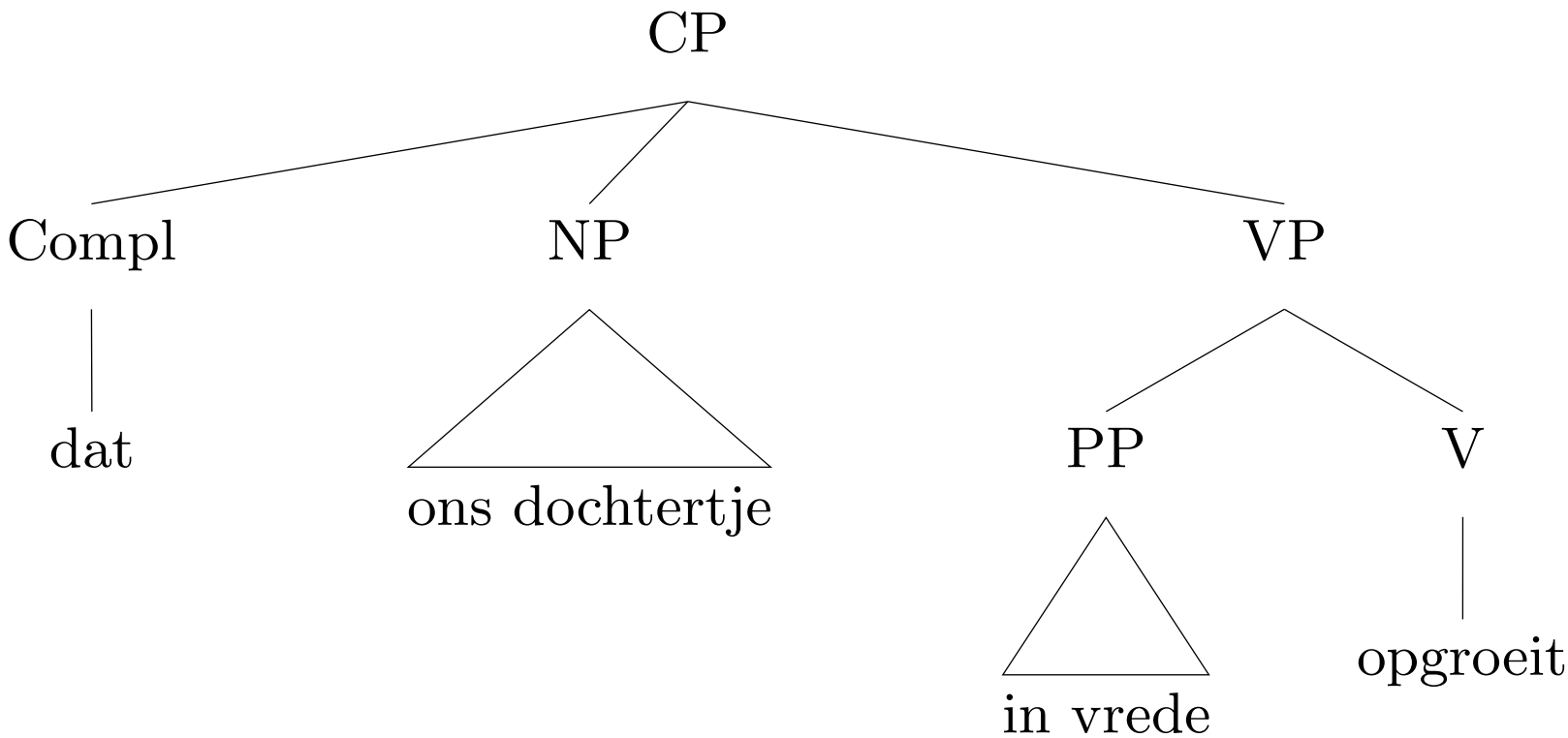
All-and-only principe

- **All:** Een grammatica moet alle zinnen van een taal kunnen herkennen,
- **Only:** Een grammatica mag geen ongrammaticale zinnen herkennen.
- Bijna alle grammatica's voldoen niet aan **All**;
- veel grammatica's voldoen niet aan **Only**.

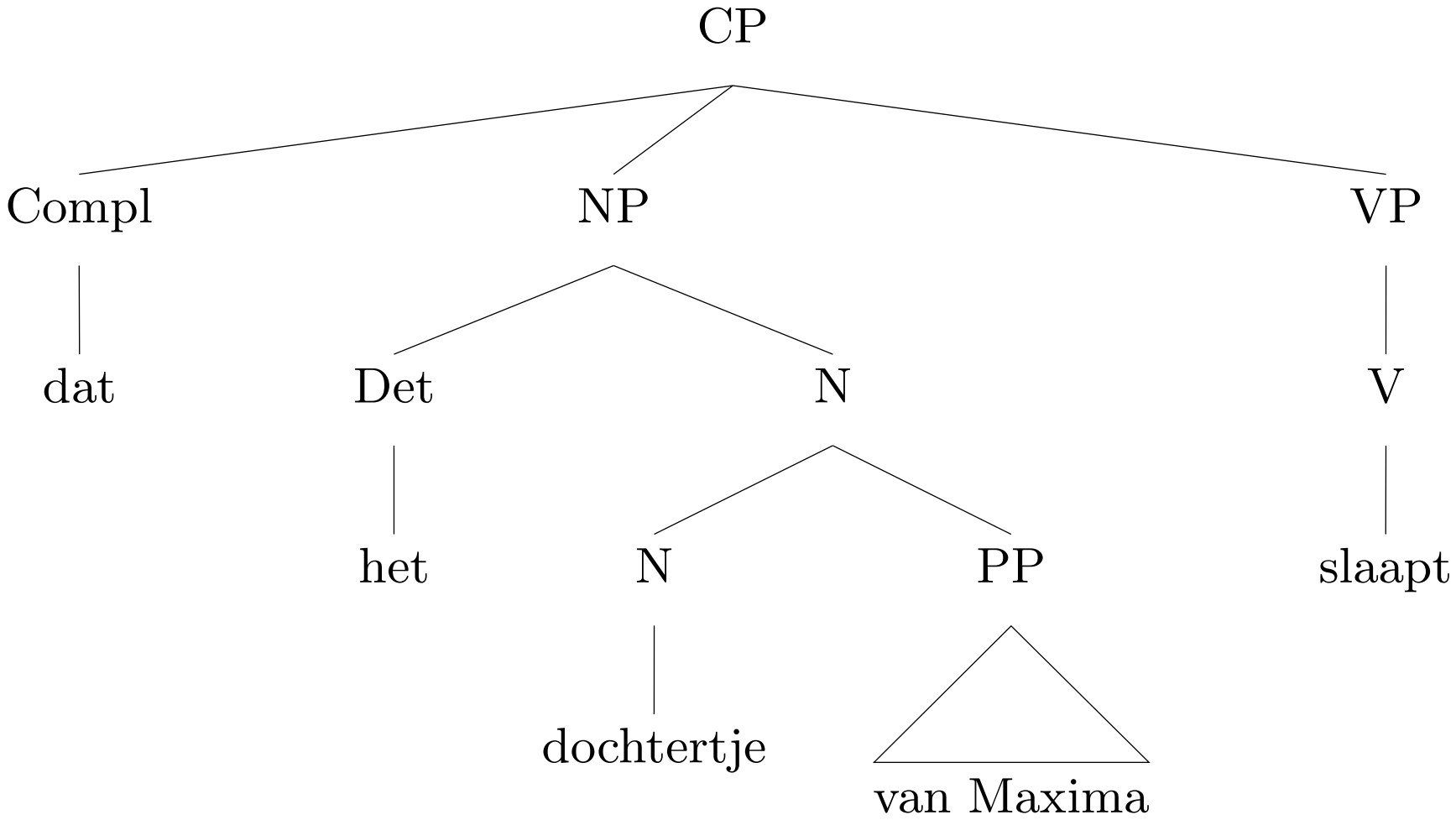
Ambiguïteit

- Wij willen dat ons dochtertje **in vrede opgroeit**
- Wij hopen dat het **dochtertje van Maxima** slaapt
- VP $\rightarrow$ VP PP
- N $\rightarrow$ N PP
- Bij ambiguïteit is er meer dan één analyse van een zin.

Ambiguïteit



Ambiguïteit



Ambiguïteit groeit exponentieel

- Wanneer deel 1 van een zin 5 mogelijke analyses heeft, en deel 2 3, heeft de hele zin 3×5 analyses
- Grammatica's die duizenden analyses aan een zin van 20 woorden toekennen zijn niet ongewoon.