Het Ontleedkundig Laboratorium

Mijnheer de Rector, dames en heren,

in het ontleedkundig laboratorium ontwikkelen we technieken om taal automatisch - dat wil zeggen met een computer - te ontleden.

Taal is overal. Een groot deel van het Internet bestaat uit taal. Onze cultuur en onze kennis is grotendeels in taal gerepresenteerd, bijvoorbeeld in heel veel boeken, die nu via Google Books ook voor iedereen vanuit zijn luie stoel beschikbaar komen. Het schijnt dat Google Books nu zo'n honderdzestig miljard woorden bevat. Op Twitter verschijnen meer dan honderd miljoen tweets. Elke dag! Op de website van de Koninklijke Bibliotheek kun je alle kranten SHEET van de laatste paar honderd jaar raadplegen.

Als je iets met die enorme en steeds verder uitdijende berg informatie wilt kunnen beginnen, heb je technieken nodig die iets van taal begrijpen. Iedereen is bekend met de zoektechnieken van bijvoorbeeld Google. In de toekomst is het mogelijk om direct naar het antwoord op je vraag te kunnen zoeken. Dat is geavanceerder dan het zoeken naar web-pagina's waar bepaalde zoektermen voorkomen die iets met je vraag te maken hebben.

Om zulke toepassingen te kunnen maken moet de computer taal beter kunnen begrijpen dan nu. Wat is daar wel niet voor nodig? Laten we ons even beperken tot de betekenis van Nederlandstalige zinnetjes. Hier is zo'n zinnetje. SHEET

```
met een rollator achtervolgt de bejaarde de inbreker
```

Om dit zinnetje te begrijpen moet je natuurlijk de Nederlandse woorden begrijpen. Maar met de kennis van de woorden alleen ben je er nog niet. Je kunt een zin maken die precies dezelfde woorden

bevat, maar die toch een andere betekenis heeft: SHEET

```
de inbreker met een rollator achtervolgt de bejaarde
```

Nu hoort `met een rollator` bij de inbreker, zo bepalen de regels van het Nederlands, zelfs als zo'n situatie veel minder voor de hand ligt. Om de betekenis van de zin te kunnen begrijpen, moet je dus weten welke woorden bij elkaar horen. Ook de preciese vorm van de woorden is belangrijk: SHEET

```
mannen die de vrouw haten
mannen die de vrouw haat
```

Om het verschil in betekenis van deze twee zinnen te kunnen begrijpen, heb je op de één of andere manier moeten vaststellen dat in het eerste voorbeeld `mannen` het onderwerp bij `haten` is. In het tweede voorbeeld is nu juist `de vrouw` het onderwerp van `haten`. We weten dat dit zo moet zijn, omdat in het Nederlands een onderwerp overeenkomt in getal (enkelvoud of meervoud) met de persoonsvorm.

Deze voorbeelden tonen aan, dat je moet weten welke woorden bij elkaar horen, en wat de rol van de woorden is. Kortom, je moet de ontleding van de zin weten - ook al ben je je daar als taalgebruiker niet erg van bewust.

We nemen aan dat ook de computer de zin moet ontleden als noodzakelijke tussenstap om de zin te begrijpen. Ook de computer moet eerst vaststellen welke woorden in de zin bij elkaar horen, en wat de rol van de verschillende woorden is. Deze automatische ontleding is het onderwerp van mijn oratie.

Om een automatische ontleding te kunnen bewerkstelligen, moet de computer weten hoe in het Nederlands woordgroepen en zinnen kunnen worden opgebouwd, en welke regels er zijn voor enkelvoud en meervoud, eerste-tweede-derde persoon, en naamval (bijvoorbeeld het verschil tussen `ik` en `mij` of `hij` en `hem`). Dat is een heel werk, want er zijn in het Nederlands (net als in andere talen) heel wat

manieren om zinnen en zinsdelen te maken, en de congruentie-eisen zijn soms heel subtiel. Maar als je er even voor gaat zitten kun je een heel eind komen.

Al gauw merk je dan echter, dat sommige zinnen meerdere ontledingen toestaan. Zoals in het volgende voorbeeld, dat U, als U van Scandinavische thrillers houdt, allang had zien aankomen: SHEET

```
mannen die vrouwen haten
```

Zo zonder verdere informatie is eigenlijk niet met zekerheid te zeggen wie hier nu eigenlijk wie haat. En in dit bijzondere geval is deze meerduidigheid vermoedelijk opzettelijk. Maar dit fenomeen doet zich ook voor bij tal van onschuldig ogende zinnetjes. Voorbeeld: SHEET

```
de bejaarde achtervolgt de inbreker met een rollator
```

Is het nu zo, dat de rollator bij de inbreker hoort? Dat kun je niet met zekerheid zeggen, al ligt het niet voor de hand. Maar wat wel of niet voor de hand ligt, is voor een computer natuurlijk niet makkelijk vast te stellen.

In het dagelijks spraakgebruik komt deze vorm van meerduidigheid om de haverklap voor. Mensen hebben daar geen last van. Computers wel. Een voorbeeld: SHEET

```
de Paus heeft gisteren in het Vaticaan
   een sappige biefstuk te eten gehad
```

Ik neem aan dat de betekenis van deze zin voor U geen problemen oplevert. Alpino - een automatisch ontleedsysteem voor het Nederlands - levert hier ook de juiste ontleding op. Maar neem nu de volgende variant: SHEET

```
de Paus heeft gisteren in het Vaticaan
   tweehonderd daklozen te eten gehad
```

Hier kiest Alpino, helaas, voor de lezing waarbij de daklozen hetzelfde lot ondergaan als de sappige biefstuk.

En als je even bij deze zinnen stilstaat, zijn er eigenlijk geen spijkerharde bewijzen aan te dragen waarom de kannibalistische lezing van dit voorbeeld is uitgesloten. Natuurlijk, over het algemeen verorberen pausen geen daklozen, maar dat sluit toch nog niet helemaal uit dat iemand zo'n betekenis zou willen kunnen verwoorden. SHEET Hij kijkt er trouwens gemeen genoeg voor.

Het blad *Onze Taal* bevat op de achterpagina de rubriek *Ruggespraak* met daarin vaak grappige voorbeelden van zulke onbedoelde meerduidigheid. Ik geef er een aantal. Opmerkelijk genoeg plaatst het eerste voorbeeld de paus al weer in een monsterlijk daglicht: SHEET

`Paus ontvangt hoofd Anglicaanse kerk SHEET`

`De Hoop wil verslaafden in cel helpen SHEET`

`Verdachte van liquidatie op A73 vrijgelaten SHEET`

`Man met rijontzegging tot 2016 weer dronken SHEET`

`Weerstand tegen jagen op ministerie genegeerd`

Hoe moeten we dit probleem van meerduidigheid nu aanpakken?

Vroeger was het antwoord van de laboranten van het ontleedkundig laboratorium op deze vraag als volgt: "ja, inderdaad, zinnen kunnen in het algemene geval meerdere ontledingen hebben. Ons programma berekent ze allemaal voor U. Knap hè?" Dit was toch wel een beetje een onbevredigende situatie. Bij langere zinnen stapelden de keuzemogelijkheden zich snel op, zodat soms wel honderden mogelijke ontledingen werden gefabriceerd. En naarmate de

regels van het systeem werden uitgebreid om ook de minder gangbare grammaticale constructies van een taal te behandelen werden deze aantallen alleen maar groter. Voor een zin van twintig woorden (dat is de gemiddelde zinslengte in uw ochtendblad) werden vaak miljoenen ontledingen gemaakt. Als de zin nog langer werd, begon ofwel de computer te haperen, ofwel het geduld van de laborant raakte op. Daar kwam dan nog bij, dat de ambiguïteiten waarmee de automatische ontleder op de proppen kwam, steeds moeilijker uit te leggen waren.

De volgende zin werd in een recent Gronings proefschrift gebruikt als voorbeeld van een niet meerduidige zin: SHEET

```
Door de overboeking vertrok een groep toeristen
   uit het hotel
```

De automatische ontleder wist niettemin dertien ontledingen te fabriceren, waarvan twaalf onzinnige. Als je een spiertje kunt vertrekken, waarom kan `een groep` dan geen `toeristen` vertrekken? Naar analogie met een zin die er erg op lijkt: SHEET

```
Door de overboeking joeg een groep de toeristen
   uit het hotel
```

Tot zover het 'ouderwetse' antwoord op de vraag wat we bij de automatische ontleding met meerduidigheid moeten doen. Het moderne antwoord van de laboranten van het ontleedkundig laboratorium op de vraag over meerduidigheid luidt: "ja, inderdaad, zinnen kunnen in het algemene geval meerdere ontledingen hebben. Maar ons programma berekent alleen de meest waarschijnlijke ontleding voor U." Op basis van grote tekstverzamelingen, wel of niet door studenten handmatig van een ontleding voorzien, worden allerlei voorkeuren verzameld en in een statistisch model gestopt. Dat model zorgt er dan vrij vaak voor, dat inderdaad de beste of in ieder geval een vrij goede ontleding wordt gekozen.

Het voorbeeld van de paus hierboven heeft er toe geleid dat in het

statistische model van Alpino - het hier in Groningen ontwikkelde automatisch ontleedsysteem - ook uitgebreid rekening wordt gehouden met welke zelfstandige naamwoorden typisch voorkomen als lijdend voorwerp, of als onderwerp, bij welke werkwoorden.

Om U een klein beetje een indruk te geven van de werkzaamheden in het laboratorium wil ik hier iets dieper op ingaan. Hoe kan een computer leren welke woorden graag voorkomen met het werkwoord `eten`? Om die informatie met de hand toe te voegen is onbegonnen werk. We doen het daarom automatisch. We laten daartoe de automatische ontleder miljoenen zinnen ontleden. De ontleding, die door de ontleder zelf als beste wordt beoordeeld, wordt bewaard. Vervolgens passen we een statistische maat toe op paren van woorden. We rekenen bijvoorbeeld uit wat de zogenaamde "pointwise mutual information" is van een bepaald werkwoord met een bepaald zelfstandig naamwoord als lijdend voorwerp. Een hoge waarde geeft aan, dat de woorden heel graag samen voorkomen. Dat is dus blijkbaar een goede combinatie. Dit zijn de hoogst scorende paren: SHEET

`onderspit`	`delven`
`knoop`	`doorhakken`
`duimschroef`	`aandraaien`
`halt`	`toeroepen`
`wenkbrouw`	`fronsen`
`steentje`	`bijdragen`
`handje`	`toesteken`

De hoogst scorende paren zijn *vaste uitdrukkingen.* Als je kijkt met welke woorden het werkwoord `eten` hoog scoort, krijg je geen vaste uitdrukkingen, maar wel typisch dingen die je kunt eten: SHEET

`boterham, kaas, druivensuikertablet, mensenvlees, vlees, boterhammetje, hapje, duivensuiker, varkensvlees, ijsje, friet, hamburger, soep, fruit, broodje, spaghetti, frietje, biefstuk, mossel, ongewervelden, pizza, groente, kebab, buikje, vis, geleedpotigen, tussendoortje, steak, insect,`

`pannenkoek, pita, couscous, brood, patatje, junkfood, aas, taartje, pasta, dagschotel, oester, fastfood, nectar, sushi, patat, spruitje, sandwich, voedsel, rijstepap, banaan, chocolade, koolhydraat, cornflakes, snack, boter, haring, kaviaar, macaroni, bes, rundvlees, hap, tapas, paling, aardappel, rijst, wafel, maaltijd, schaaldier, chocola`

Het werkwoord `ontkennen` is een mooi voorbeeld van een werkwoord dat duidelijke voorkeuren heeft, SHEET niet alleen voor lijdend voorwerp, maar ook voor het onderwerp en zelfs voor de bijwoordelijke bepaling. Wie ontkent er? Beklaagden, woordvoerders, betichtend, verdachten, enzovoorts. Wat ontkennen ze? Betrokkenheid, aantijgingen, beschuldigingen, geruchten, berichten en zo verder. En hoe doen ze dat? Glashard, staalhard, categorisch, halsstarrig, stelling of formeel.

Zulke lijstjes zijn niet helemaal perfect, maar U moet daarbij bedenken dat ze volledig automatisch tot stand zijn gekomen. Met behulp van de automatische ontleder zelf. En soms is de informatie erg subtiel. Zo vind ik het veelzeggend dat zowel `Ajax` als `PSV` een betere combinatie vormt als onderwerp van het werkwoord `verliezen` dan `Feyenoord`!

De informatie die in deze lijstjes beschikbaar is, hebben we vervolgens gebruikt om de ontleder een behoorlijk stuk te verbeteren. Dit voorbeeld toont, hoe we in het moderne ontleedkundig laboratorium van alles proberen om de ontleder steeds de juiste ontleding te laten kiezen.

Deze strategie heeft de laatste jaren veel succes gehad. De moderne ontleedsystemen werken over het algemeen erg goed. Om de kwaliteit van een ontleedsysteem te bepalen, worden de resultaten van zo'n systeem vergeleken op een verzameling zinnen waarvoor van te voren al de juiste ontleding (bijvoorbeeld door student-assistenten)

is vastgesteld. Hoe vergelijk je de correcte ontleding met de automatische ontleding? Je kunt zeggen: die is goed of fout. Maar meestal kijken we wat precieser, en we kijken welke woorden op de juiste wijze zijn ontleed - behorend tot de juiste woord-groep, en met de juiste rol. Als je dus spreekt over een accuratesse van - zeg - 90 procent, dan betekent dat, dat de automatische ontleding voor 90 procent van de woorden correct is.

Die 90 procent - want dat is inderdaad de accuratesse die de beste systemen op geschreven taal zo'n beetje halen, is erg goed. Het blijkt bijvoorbeeld dat taalkundigen zelf ook niet veel verder dan 93 of 94 procent komen om redenen waar ik straks nog wat over zal zeggen. De bovengrens, zou je kunnen zeggen, is dus niet 100 procent maar eerder die 94 procent. Bovendien blijkt dat je voor allerlei toepassingen al heel veel baat kunt hebben van die 90 procent. Met andere woorden, je hoeft niet de bovengrens helemaal te halen om iets nuttigs te kunnen doen.

Na deze uitgebreide inleiding kom ik dan nu toe aan de vraag die ik in deze oratie aan de orde wil stellen. Die vraag luidt: kan het automatisch ontleden van het Nederlands eigenlijk nog wel verbeterd worden?

Er is voor beide mogelijke antwoorden op deze vraag wel wat te zeggen.

Degene die van mening is dat er wel degelijk verbetering mogelijk is, kan bijvoorbeeld wijzen op de volgende aspecten.

Op de eerste plaats werkt de automatische ontleding nog helemaal niet zo goed voor gesproken taal, zelfs als we ons beperken tot getranscribeerde spraak, dat wil zeggen, een schriftelijke weergave van

gesproken taal. In gesproken taal komen fenomenen voor waar de automatische ontleders nog niet of slecht tegen kunnen. Al weer jaren geleden werkten we mee aan het automatisch telefoonbeantwoordsysteem voor openbaar-vervoer-informatie. Ter inspiratie bestudeerden we ook wel de vragen die gebruikers aan echte, dat wil zeggen meestal vrouwelijke, telefoonbeantwoorders van OVR stelden. Bijvoorbeeld: SHEET

```
ja en Groningen dat is toch gewoon Amersfoort niet?
```

Het vereist alvast enig inlevingsvermogen van de computer om er achter te komen wat hier de bedoelde betekenis is.

Ook als we in het bekende Corpus Gesproken Nederlands kijken, dan zijn er nogal wat zinnen (nou ja, zinnen) die lastig zijn. Dit is alvast de eerste zin van dat corpus: SHEET

```
uhm moeten langs uhm de Gamma gaan denk ik voor uh
   uh om die afsluiters voor die plafonddozen
```

Dit is ook een typisch voorbeeld: SHEET

Sterretje-a geeft aan dat het een afgebroken woord betreft.

```
ah ja die uh die uh afge*a die afplak*a uh f*a die
   uhm anti*a uhm kras*a uh ...
```

Zo praten mensen echt, en voor dit soort data is de automatische ontleding voor vele verbeteringen vatbaar. Iets vergelijkbaars zien we bij geschreven taal zoals die bijvoorbeeld in Twitter-berichten wordt gebruikt. Hier lijkt een geheel nieuw soort geschreven Nederlands te ontstaan, en ook hier kan de automatische ontleder nog wel wat hulp gebruiken. Een voorbeeldje: SHEET Ik weet ook niet hoe je dit moet voorlezen.

```
tesssssx Normaal zou k me nu aant klaarmaken zyn
   vo scorro, mrgoed geen oog dicht gedaan vanacht
   &ben sick.. Dus f*ck school :-) #trusttteeeeeee
```

Je zou kunnen zeggen: dat is een heel nieuwe taal, geen wonder dat die automatische ontleder voor het Nederlands het niet doet. Maar dat is toch onbevredigend als je ziet hoe makkelijk wij mensen zo'n berichtje begrijpen, ook zonder uitvoerige uitleg van deze nieuwe taal. SHEET

In zekere zin vinden we dezelfde problemen in nette geschreven taal, al spreken we dan eerder van fouten dan van creatief taalgebruik, bijvoorbeeld spelfouten of grammaticale fouten. Die fouten leiden er vaak toe dat de automatische ontleder volledig op het verkeerde been wordt gezet. Het opvallende in zulke gevallen is opnieuw, dat je als menselijke taalgebruiker eigenlijk direct begrijpt wat de schrijver bedoelt, of de fout niet eens opmerkt. SHEET

Tot zover de overwegingen die suggereren dat de automatische ontleding van het Nederlands nog wel verbeterd moet kunnen worden.

Je kunt ook argumenten naar voren brengen die suggereren dat op korte termijn weinig vooruitgang valt te verwachten, bij de automatische ontleding van het Nederlands. In ieder geval voor enigszins zorgvuldig geformuleerde geschreven taal. Ik noem drie van zulke argumenten.

Het eerste argument heb ik zo even eigenlijk al genoemd. De accuratesse - zo'n negentig procent - van de automatische ontleder ligt al vrij dicht bij de mogelijke bovengrens die zo rond de vier-en-negentig procent lijkt te liggen. Die negentig procent is dus al heel hoog. Waarom ligt die bovengrens niet op de honderd procent? Dit komt omdat er zinnen zijn waarbij meerdere ontledingen eigenlijk even goed zijn. Neem de volgende zin: SHEET

```
de teruggang van het aantal juristen in het parlement
```

hierbij moet de automatische ontleder onder andere beslissen of de woordgroep `in het parlement` bij `aantal juristen` hoort, of bij

`teruggang`. In het ene geval gaat het om een teruggang, in het parlement, van het aantal juristen. In het andere geval gaat het ook om een teruggang. En wel van het aantal juristen in het parlement. Wie me nu glazig aankijkt heeft het toch begrepen: beide mogelijkheden komen natuurlijk op hetzelfde neer. Toch is het in het algemeen wel juist dat de ontleder voor deze constructie een specifieke aanhechting moet kiezen. Bekijk hiervoor de volgende twee varianten: SHEET

```
de teruggang van het aantal juristen in toga

de teruggang van het aantal juristen in vergelijking
   met vorig jaar
```

Hier is natuurlijk wel belangrijk dat je de aanhechtingen goed krijgt, want een `teruggang in toga` is onzin, en `juristen in vergelijking met vorig jaar` is ook niet wat bedoeld kan zijn.

Een tweede reden dat ik niet geloof dat de automatische ontleders voor het geschreven Nederlands nog veel beter kunnen worden volgt uit de observatie dat er teveel overgebleven problemen zijn. We zeiden net dat negentig procent al dicht bij de vier-en-negentig procent komt. Die overgebleven vier procent, zou je kunnen zeggen, moet je dan toch nog kunnen overbruggen. Als je een analyse van de overgebleven fouten van het systeem maakt, dan vind je echter dat zo'n beetje elke overgebleven fout uniek is. Je boekt dus nauwelijks vooruitgang door een oplossing voor één fout te verzinnen. Daar komt trouwens het volgende nog bij. Als zo'n fout een tekortkoming van het woordenboek of de grammaticale regels betreft, dan blijkt vaak dat een uitbreiding van dat woordenboek of die regels weer leidt tot nieuwe problemen.

Een voorbeeld. Door zorgvuldige analyse van de resultaten van de automatische ontleder kom je er achter dat het verslag van cricketwedstrijden soms wonderlijke constructies bevat: SHEET

```
Aan bat leek VVV na drie onnodige run outs op 122
   voor 7 ( 43 overs ) verslagen .
```

Zoals u wellicht weet als cricketliefhebber, is het woordje `over`, meervoud `overs`, ook een zelfstandig naamwoord. Als we in de verleiding komen om deze mogelijkheid aan het woordenboek toe te voegen, dus dat `over` ook een zelfstandig naamwoord is, dan leidt dat tot nieuwe problemen, bijvoorbeeld, in de volgende zin. We zijn nu overigens in het bridge-verslag verzeild geraakt: SHEET

```
Hij speelde de vrouw en West ziet geen reden
   deze over te nemen
```

Natuurlijk zul je dan net zien, dat de automatische ontleder vermoedt - niet op de hoogte van het feit of we bridge of cricket spelen - dat West hier `deze over` niet wil nemen, naar analogie met SHEET

```
Hij speelde de vrouw en West ziet geen reden
   deze slag te nemen
```

Dit probleem is dan vervolgens wellicht ook wel weer op te lossen, maar hopelijk voelt U met mij mee dat dat allemaal toch wel een hoop gedoe wordt. En de kans dat er in de test-set een zin uit een cricketverslag voorkomt, zodat je beloond kunt worden voor je inspanningen, is ongeveer nul. Maar goed, de economen onder U hebben het al lang begrepen, we worden hier geconfronteerd met de wet van de afnemende meeropbrengsten. SHEET En in ons eigen vakgebied heeft Hugo Brandt Corstius dit fenomeen al in 1978 beschreven als de Derde Wet van de computer-taalkunde.

Als derde reden dat de kwaliteit van de automatische ontleding voor geschreven taal niet veel meer zal kunnen stijgen noem ik weer even kort de meerduidigheid die ik in de inleiding heb uitgelegd. Het is zo, dat we veel van deze gevallen tegenwoordig goed kunnen krijgen, maar het is toch ook wel zo, dat een computer nooit voldoende kennis kan bevatten om voorbeelden zoals de volgende (eerlijk gevonden

op Internet) altijd en overal goed te doen. SHEET

```
een man met een boodschappenkar die op zijn vrouw
   staat te wachten SHEET

ballen gehakt uit de jus van de keurslager SHEET

Rutte spreekt André Kuipers in ISS
```

En vooruit, nog een paar voorbeelden uit *Ruggespraak*: SHEET

```
vele ouderen eten alleen met de kerst SHEET

wormen van drie kilometer onder de grond SHEET

twee huwelijken op drie stranden
```

Wat volgt hier nu uit voor het onderzoek naar het automatisch ontleden in de komende jaren? Aan de ene kant moeten we het onderzoek meer gaan richten op de automatische ontleding van gesproken taal en van wat wel user-generated-content genoemd wordt - geschreven taal die zich over het algemeen veel minder aantrekt van de geschreven en ongeschreven regels van het Nederlands. Bijvoorbeeld die Twitter-berichten.

Aan de andere kant zie ik allerlei mogelijkheden om de huidige resultaten van de automatische ontleding in te zetten voor allerlei toepassingen, maar ook voor bijvoorbeeld theoretisch taalkundig en psycholinguistisch onderzoek. Zoals U vast weet, is het niet toegestaan om bij het onderzoek je eigen data te verzinnen. Recentelijk is hierover nog enige ophef ontstaan, ik meen in de sociale psychologie. Gelukkig is het nog niet zo erg tot de buitenwereld doorgedrongen dat het in bepaalde taalkundige kringen eigenlijk heel gewoon is, om je eigen data te verzinnen - pardon, ik bedoel om je eigen taal-intuities als data te beschouwen. De automatische ontleding kan, denk ik, deze toch wat duistere praktijken aanvullen met wat objectievere gegevens. SHEET Je hebt dan bovendien ook

de beschikking over frequentiegegevens waar nu vaak nog niet erg naar gekeken wordt.

Maar dat is eigenlijk een heel ander verhaal.

Ik heb gezegd.
SHEET