De ontleding van het Corpus Gesproken Nederlands

Ton van der Wouden en Heleen Hoekstra

Een corpus, in de zin van een verzameling tekst, bij elkaar brengen is tegenwoordig geen kunst meer. Met een computer en een internetaansluiting kun je het klassieke moeizame proces van intikken van alle data overslaan en heb je gratis of bijna gratis toegang tot eindeloze verzamelingen literatuur, kranten, pornografie en noem het soort geschreven taal dat je wilt maar op. Zeg maar hoeveel je moet hebben: 1 miljoen woorden (zoals het Corpus Eindhoven uit de jaren '70), 38 miljoen (zoals het grootste direct raadpleegbare corpus van het Instituut voor Nederlandse Lexicologie), 100 miljoen (zoals het British National Corpus), of misschien nog wel meer?[1] Geen probleem.

Veel moeilijker is het om aan gesproken materiaal te komen. Natuurlijk, er wordt heel wat afgekletst, er wordt zelfs veel meer gepraat dan geschreven, maar het resultaat van al dat gepraat is niet zomaar beschikbaar. En met alleen maar op band of CD opnemen van al die spraakklanken ben je er ook nog niet, want je wilt er ook in kunnen zoeken en tellen. En daarom word nu dus het Corpus Gesproken Nederlands opgebouwd.

Dat corpus bestaat niet alleen uit het spraakgeluid, al vult dat wel het leeuwendeel van de inmiddels honderden CD's waarop het gedistribueerd wordt, nee, de honderden uren spraak worden allemaal uitgetikt, min of

[1] Eén manier om de grootte van het corpus internet te schatten (Van Oostendorp en Van der Wouden 1998) is uit te gaan van een woord met een min of meer bekende frequentie. Het woordje *eens* komt gemiddeld eens op de 1000 woorden voor, en vrijwel uitsluitend in Nederlandstalige teksten. De zoekmachine Google vond op 19 november ongeveer 300000 vindplaatsen van die string, wat betekent dat Google toegang biedt tot een corpus van naar schatting tenminste 300 miljoen woorden Nederlandse tekst.

meer conform de Nederlandse spellingsregels. Spraakherkening gaat nog niet goed genoeg, zeker niet als de software niet aan de spreker gewend is, dus automatisch kan die omzetting van spraak nog niet plaatsvinden.

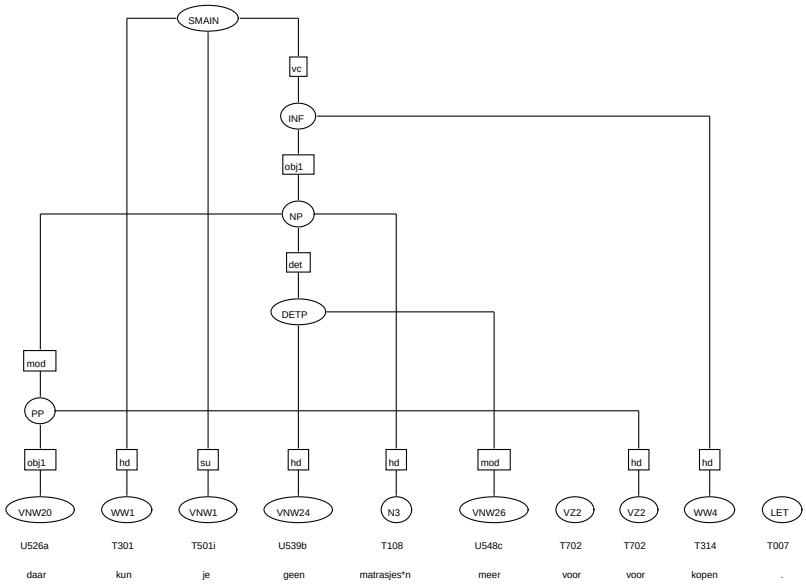
Een corpus van alleen maar tekst (we spreken van “ongeanalyseerd”) heeft voor taalkundig onderzoek echter maar beperkte waarde, tenminste als je vraagstelling verder gaat dan woordjes en lettertjes tellen. Een corpus wordt dan ook een stuk nuttiger als je het verrijkt met extra informatie. Dat toevoegen van informatie is echter heel veel werk, werk dat vaak niet (helemaal) geautomatiseerd kan worden en dus (soms heel veel) geld kost. Per soort verrijking moet dus iedere keer weer een kosten-baten-afweging gemaakt worden.

Daarom ook wordt niet het hele Corpus Gesproken Nederlands, maar slechts een representatief deel van de spraakdata - zo’n tien procent van het totaal van tien miljoen woorden - voorzien van een syntactische annotatie; in gewone-mensentaal, de zinnen van het corpus worden ontleed. In het navolgende gaan we kort in op eigenschappen van die annotatie en op de manier waarop die tot stand komt. Er valt meer over te lezen in de laatste aflevering van 2002 van het tijdschrift Nederlandse Taalkunde.

De ontleding begint bij de taalkundig ontlede orthografische transcriptie, en niet bij het ruwe spraaksignaal - de computertaalkunde is namelijk nog lang niet ver genoeg om spraak uit het wild betrouwbaar te kunnen ontleden. De genoemde orthografische transcriptie wordt gecontroleerd en dient vervolgens als invoer voor een taalkundige ontleedautomaat (parser), die alle woorden van een woordsoort voorziet (lidwoord, zelfstandig naamwoord, persoonsvorm enzovoort) (in het vak spreken we van P(art)o(f)S(peech)-tagging). Dit soort ontleders is tegenwoordig heel goed: het foutenpercentage ligt in de buurt van de 2. Toch is ook dat nog niet goed genoeg voor opname in het corpus. Ook dit resultaat wordt daarom gecontroleerd en waar nodig gecorrigeerd, en pas daarna is het goed genoeg om in het corpus opgenomen te worden. Al met al is het hele proces echter tamelijk efficiënt, vandaar dat vrijwel het hele corpus die taalkundige ontleding krijgt. Die gecorrigeerde taalkundig ontlede orthografische transcriptie fungeert vervolgens als input voor de syntactische annotatie, oftewel de redekundige ontleding.

Het primaire doel van deze redekundige ontleding is het herkennen en

benoemen van woordgroepen en zinsdelen, alsmede van hun relaties en afhankelijkheden. Om het proces van annotatie en correctie werkbaar te houden, moet de annotatie zo eenvoudig mogelijk zijn. Ook is het zaak zoveel mogelijk gebruikers van dienst te kunnen zijn, zodat adoptie van (één versie van) één theoretisch kader ongewenst is. Anderzijds zijn de CGN–gebruikers gebaat bij een zo rijk mogelijke output. Er is daarom gekozen voor een type ontleding dat maximaal onafhankelijk is van theoretische modes; in het algemeen wordt nauw aansluiting gezocht bij de traditionele Nederlandse zinsontleding, in casu de ANS. Het gaat daarbij dan ook om uit de traditie bekende entiteiten als zelfstandignaamwoordgroepen en voorzetselgroepen, en functies als onderwerp, gezegde en voorzetselvoorwerp. Het resultaat is een soort boompjes, maar soms wel met kruisende takjes. Hieronder staat een voorbeeld van zo'n boompje.



Zoals het plaatje laat zien, wordt in de ontleding de oorspronkelijke volgorde van de zin gerespecteerd en werken we zonder functionele projecties.

Het gevolg is dat vele relaties en afhankelijkheden niet lokaal zijn, wat weer leidt tot veel kruisende afhankelijkheden.

De Nederlandse en de Belgische zinnen worden volgens dezelfde principes ontleed om de mogelijkheden tot vergelijking van verschillende soorten Nederlands te optimaliseren.

De syntactische structuren die samen met de geluidsfiles en de andere vormen van verrijking het Corpus Gesproken Nederlands vormen, worden halfautomatisch afgeleid. Een computerprogramma wordt gevoed met een aantal ontlede zinnen en stelt op basis daarvan een statistische grammatica op. Op basis van die grammatica doet het programma voorstellen voor het ontleden van nieuwe zinnen; studentassistenten fiatteren of corrigeren deze voorstel-ontledingen, hun werk wordt ook weer gecontroleerd en zo nodig gecorrigeerd, en dan worden de ontledingen geacht goed genoeg te zijn voor opname op de CDs van het CGN. Bovendien worden ze toegevoegd aan het corpus op basis waarvan het programma zijn grammatica opstelt, zodat de voorstellen van het programma steeds beter worden, in elk geval in theorie.

In de praktijk blijkt dat overigens niet mee te vallen. De ontleder heeft weliswaar nog maar zelden problemen met bijvoorbeeld de correcte ontleding van (eenvoudige) zelfstandignaamwoordgroepen en voorzetselgroepen, maar met hogere structuren gaat het nog steeds teleurstellend vaak mis. Dat dient vermoedelijk ten dele te worden toegeschreven aan het feit dat de gebruikte ontleder (van Thorsten Brants) bedoeld is voor krantentekst, tekstmateriaal dat in principe als welgevormd (dat wil zeggen, in overeenstemming met de regels van de (schrijftaal)grammatica) kan worden beschouwd en in elk geval veel minder aarzelingen, correcties en versprekingen bevat dan spreektaal. Een ander probleem voor de automatische ontleder wordt veroorzaakt door het feit dat die strikt lokaal werkt en geen raad weet met discontinue structuren, of ze nu het gevolg zijn van de eigenaardigheden van de Nederlandse syntaxis (*daar* en *voor* horen bij elkaar in *DAAR kun je geen matrasjes meer VOOR kopen*) of door eigenaardigheden (zoals versprekingen en aarzelingen) van de spreker (*daar kun je geen matrasjes meer VOOR voor kopen*). Een gedeelte van de tegenvallende progressie is bovendien waarschijnlijk toe te schrijven aan het gebrek aan homogeniteit van het corpus. We hebben namelijk de indruk dat er grote

verschillen zijn tussen, bijvoorbeeld, de interviews met leraren Nederlands, de spontane gesprekken die bij informanten thuis in de huiselijke kring zijn opgenomen en de monologen uit de Tweede Kamer.

En waar gebruik je zo'n geannoteerd Corpus Gesproken Nederlands nou voor? Dat moeten de gebruikers natuurlijk zelf weten, maar we kunnen wel een paar voorbeeldjes geven.

- Nog niet zo lang is het mogelijk het zelfstandig naamwoord *richting* te gebruiken als voorzetsel: *de expeditie trok richting Noordpool, een stukje service richting de klant*. De ANS constateert dat dit gebruik niet voor iedereen aanvaardbaar is, het corpus kan de vraag helpen beantwoorden in welke kringen en in welke situaties dit gebruik daadwerkelijk voorkomt.

- Theoretici uit de school van Chomsky hebben altijd de mond vol over lange Wh-verplaatsingen (*wie denk je dat ik gisteren gesproken heb?*): het corpus kan helpen na te gaan of die dingen ook echt bestaan.

- Is /n/-deletie (*moete manne met baarde zijn*) gevoelig voor het onderscheid naamwoord-werkwoord enerzijds en onderwerp-lijdend voorwerp anderzijds?

Het is helemaal niet gezegd dat dit interessante onderzoeksvragen zijn, laat staan dat er interessante antwoorden op zullen komen, maar een geannoteerd corpus als het Corpus Gesproken Nederlands stelt ons wel eindelijk in staat die vragen te stellen.