Syntactische annotatie voor het Corpus Gesproken Nederlands (CGN)

Ton van der Wouden, Heleen Hoekstra, Michael Moortgat, Bram Renmans en Ineke Schuurman

versie van 9 juli 2002

Abstract

The paper discusses the syntactic annotation for the Spoken Dutch Corpus, a Dutch/Flemish cooperation project to build an annotated corpus of about one thousand hours of continuous speech, which amounts to 10 million words. After a brief introduction to the project, we discuss the kind of syntactic annotations we envisage (dependency structures) and the way they are created (semi-automatically). We mention some peculiarities of spoken language, and we finish with a discussion of some of the kinds of questions the corpus may help answering.

1 Inleiding

Dit artikel besteedt aandacht aan de syntactische annotatie ten behoeve van het Corpus Gesproken Nederlands (in het vervolg meestal CGN). In de tweede paragraaf worden doel en opzet van het CGN besproken, alsmede de plaats van de syntactische annotatie daarin. In de derde paragraaf bespreken we het soort syntactische structuren dat het CGN oplevert. In de tamelijk technische vierde paragraaf gaan we in op de uitgangspunten van de syntactische analyse, en in de vijfde op de praktische implementatie van het proces. In de zesde paragraaf bespreken we een aantal specifieke problemen verbonden aan het ontleden van gesproken taal. In de zevende en laatste paragraaf tenslotte worden enkele voorbeelden behandeld van typen vragen die taalkundigen altijd over het Nederlands hadden willen stellen en die nu met behulp van een syntactisch geannoteerd corpus van het gesproken Nederlands niet alleen gesteld maar misschien ook daadwerkelijk beantwoord kunnen worden.[1]

2 Het Corpus Gesproken Nederlands

Het Corpus Gesproken Nederlands is een samenwerkingsproject van een aantal Nederlandse en Vlaamse universiteiten (Goedertier *et al.* 2000; Oostdijk 2000a; Oostdijk 2000b). Het project, dat wordt gefinancierd door NWO en FWO en beheerd door de Taalunie, is begonnen in juni 1998 en heeft een looptijd van vijf jaar. Het einddoel is een geannoteerd corpus van ongeveer duizend uur lopende spraak, wat neerkomt op zo'n tien miljoen woorden.[2]

Het CGN is bedoeld als een bron, een nieuw soort bron, van informatie voor taalkundig onderzoek en voor taal- en spraaktechnologie. Om deze verschillende doelgroepen optimaal te kunnen bedienen, wordt het corpusmateriaal verzameld in uiteenlopende communicatieve situaties, waaronder spontane dialogen, telefoongesprekken, vraaggesprekken, discussies, debatten, lezingen, nieuwsuitzendingen en voorgelezen literatuur. Tweederde van het materiaal is afkomstig uit Nederland, eenderde uit het Nederlands sprekende gedeelte van

[1]Deze publicatie is tot stand gekomen in het kader van het project "Corpus Gesproken Nederlands" met financiële steun van de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) en de Vlaamse Overheid. Dit artikel is ten dele een vertaling annex samenvatting van gedeelten van Hoekstra *et al.* (2001b) en van Hoekstra *et al.* (2001a).

[2]Meer informatie over het project en over de distributie van het materiaal via de website http://www.lands.let.kun.nl/cgn.

België.[3] Als het corpus voltooid is, zal het de grootste en meest gevarieerde verzameling gesproken Nederlands zijn die tot dusver bijeengebracht is. Tussentijdse versies (die inmiddels al meer dan 100 CD-ROMs vullen) worden gedistribueerd via ELRA.[4]

Het project beoogt verschillende niveaus en typen van annotatie aan te bieden. Het gehele corpus wordt orthografisch getranscribeerd en taalkundig ontleed: ieder woord krijgt een woordsoort toegekend.[5] Een representatieve selectie van zo'n tien procent van de spraakdata – het zogenoemde "kerncorpus" – wordt voorzien van een brede fonetische transcriptie en van een syntactische annotatie. Bovendien ontvangt een kwart van het kerncorpus, in totaal dus zo'n 250.000 woorden, een prosodische annotatie.[6] In deze bijdrage gaan we vooral in op de syntactische annotatie, dus op de redekundige ontleding.

3 De syntactische analyses van het CGN

Uitgangspunt voor de ontleding is de taalkundig ontlede orthografische transcriptie (dus niet het ruwe spraaksignaal - zo ver is de computationele taalkunde nog niet).[7] Het materiaal is opgedeeld in annotatie-eenheden, die niet noodzakelijkerwijze overeenkomen met het klassieke begrip "zin", maar die we toch met die term zullen aanduiden.[8] Een realistisch voorbeeld van zo'n zin is gegeven in (1).[9]

(1) Ik zal u gaan uitleggen hoe we dat zo'n beetje hebben aangepakt dat probleem .

Deze transcriptie van het spraaksignaal is verrijkt met lemma's – dat wil zeggen dat ieder woord is gekoppeld aan een basisvorm in het CGN-lexicon – en de al genoemde taalkundige ontleding:[10]

[3] Er blijft natuurlijk altijd wat te wensen over: voorlopig blijven varianten van het Nederlands als gesproken door kinderen, niet-moedertaalsprekers, inwoners van Suriname en de Antillen enzovoorts buiten beschouwing. Niets verbiedt ons echter om dat soort materiaal te gelegener tijd aan het corpus toe te voegen.

[4] ELRA staat voor European Language Resources Association; informatie via http://www.icp.inpg.fr/ELRA/.

[5] De vakterm binnen de computationele taalkunde is Part Of Speech tagging of POS-tagging, maar in dit artikel zullen we zo veel mogelijk trachten de klassieke Nederlandse terminologie te hanteren.

[6] In de prosodische annotatie worden de belangrijkste grenzen van woordgroepen (frasegrenzen) alsmede één of twee belangrijkste woorden (zinsaccenten) van elke frase aangeduid.

[7] Behalve punten, vraagtekens en beletseltekens (...) wordt er tijdens de orthografische transcriptie van het corpusmateriaal geen interpunctie aangebracht, omdat het onmogelijk blijkt daarin voldoende consistentie tussen transcribenten te bereiken.

[8] Miller & Weinert (1998:30-31) verdedigen zelfs de stelling dat spontane gesproken taal überhaupt niet of nauwelijks zinnen kent die overeenkomen met schrijftaalzinnen, eenheden die beginnen met een hoofdletter en eindigen met een punt. In navolging van Halliday nemen zij aan dat "the language system must be analyzed as having clauses combining into clause complexes" (p. 31). Toch zullen wij in het vervolg die annotatie-eenheden aanduiden met "zin".

[9] Wegens ruimtegebrek hebben we een relatief korte zin gekozen: vele van de zinnen in het corpus bestaan weliswaar uit een enkel woord, maar er zijn er ook van meer dan 150 woorden. De gekozen zin is bovendien naar verhouding tamelijk welgevormd, maar later in dit artikel komen nog voorbeelden met versprekingen, aarzelingen en dergelijke aan de orde.

[10] De taalkundige ontleding geschiedt op een manier die vergelijkbaar is met de redekundige ontleding, namelijk half-automatisch: het resultaat van een team van automatische woordsoorten-toekenningsprogramma's, die gebruik maken van zo'n 360 woordsoorten-labels en gezamenlijk een precisie bereiken van zo'n 95%, wordt met de hand gecontroleerd en zo nodig gecorrigeerd. Voor details aangaande POS-tagging en lemmatizering binnen het CGN verwijzen we naar Van Eynde (2001) en Van Eynde *et al.* (2000).

(2) <au id=1 t=0.000 sp=N00052>

ik	VNW(pers,pron,nomin,vol,1,ev)	ik
zal	WW(pv,tgw,ev)	zullen
u	VNW(pers,pron,nomin,vol,2b,getal)	u
gaan	WW(inf,vrij,zonder)	gaan
uitleggen	WW(inf,vrij,zonder)	uitleggen
hoe	BW()	hoe
we	VNW(pers,pron,nomin,red,1,mv)	we
dat	VNW(aanw,pron,stan,vol,3o,ev)	dat
zo'n	VNW(aanw,det,stan,prenom,zonder,agr)	zo'n
beetje	N(soort,ev,basis,onz,stan)	beetje
hebben	WW(pv,tgw,mv)	hebben
aangepakt	WW(vd,vrij,zonder)	aanpakken
dat	VNW(aanw,det,stan,prenom,zonder,evon)	dat
probleem	N(soort,ev,basis,onz,stan)	probleem
.	LET()	.

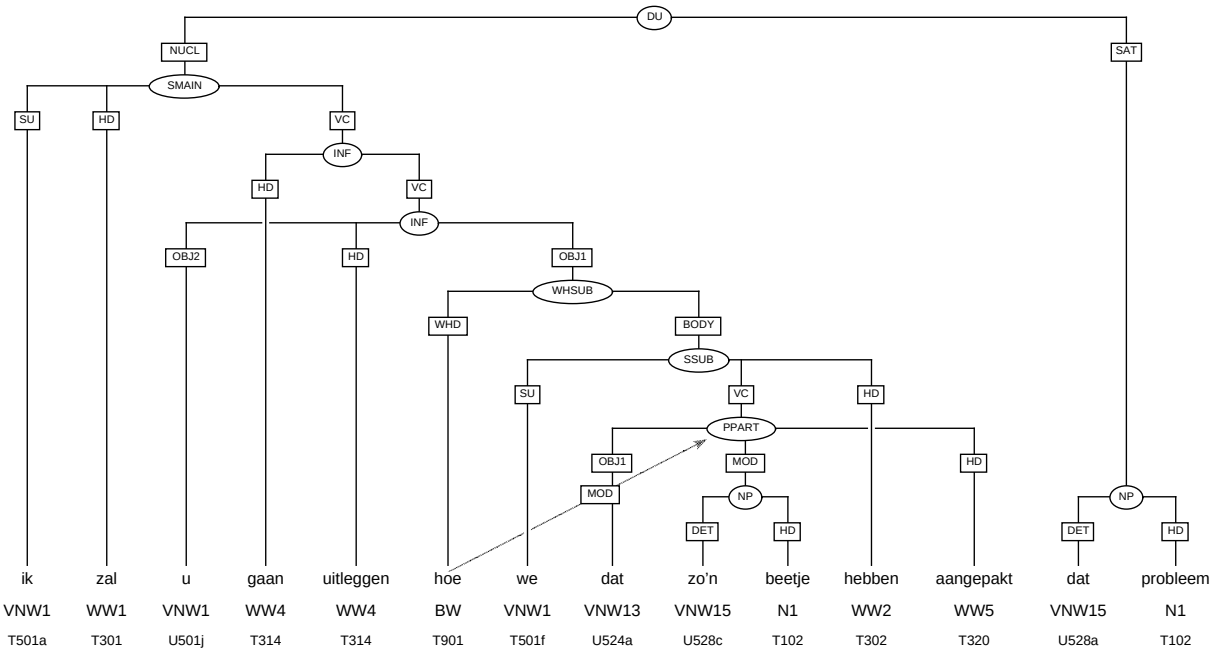
Toelichting: De eerste regel van dit voorbeeld bevat een unieke verwijzing naar de locatie van dit fragment in een spraakbestand. Vervolgens zien we drie kolommen: de eerste geeft het getranscribeerde spraaksignaal, een woord per regel, de tweede kolom biedt woordsoorteninformatie (hoofdwoordsoorten in hoofdletters, kenmerken tussen haakjes), en de derde kolom bevat de lemma's, dus de lexicale basisvormen. Bijvoorbeeld, het eerste woord, *ik*, is een voornaamwoord, en wel een volle vorm (in tegenstelling tot gereduceerde vormen als *we* op regel 8[11]) van een persoonlijk voornaamwoord van de eerste persoon enkelvoud in de eerste naamval, waarvan de basisvorm *ik* luidt. Het tweede woord, *zal*, is een persoonsvorm, tegenwoordige tijd enkelvoud, van het werkwoord *zullen*, enzovoorts.

(3) geeft een beeld van het soort analyse dat van deze zin voor het CGN wordt gegenereerd (hoe dat gebeurt, komt later aan de orde).[12]

(3) **Voorbeeldontleding**

[11] De transcriptie laat ook gereduceerde vormen als *'k* en *'ns* toe, wat laat zien dat de term 'orthografisch' binnen het CGN niet mag worden gelijkgesteld aan 'volgens de (officiële) spellingregels'. Dialectvormen zoals *benne*(*n*) ('zijn') en *effekes* ('eventjes') worden apart gemarkeerd – vergelijk noot 26.

[12] De onderste rij labels (T501a enz.) is (zonder informatieverlies) afgeleid van de taalkundige ontleding (POS-tags) (vergelijk (2)): "T501a" is bijvoorbeeld een afkorting voor "VNW(pers,pron,nomin,vol,1,ev)". De rij labels daarboven is daarvan afgeleid en vormt een reductie van de oorspronkelijke verzameling, die voor de automatische ontleder onhandig groot is. "VNW1" staat voor 'persoonlijk voornaamwoord', "WW1" voor 'werkwoord (persoonsvorm)', "WW4" voor 'werkwoord (onbepaalde wijs)', enz. De namen van de takken in de graaf staan in rechthoekjes: "SU" staat voor 'onderwerp', "HD" voor 'hoofd', "OBJ2" voor 'secundair object', "NUCL" voor 'kern (van een discourse-eenheid DU)', "VC" voor 'verbaal complement', "OBJ1" voor 'primair object', enz.



Het voorbeeld in (3) illustreert een aantal opvallende kenmerken van de CGN-annotatie:

- De annotatie geeft een soort dependentiestructuur (afhankelijkheidsstructuur) en geen constituentenstructuur of functionele structuur. Het resulterende object is dan ook een graaf, en geen klassieke boomstructuur: in zo'n graaf kunnen takken elkaar kruisen, en kunnen dochters meer dan één moeder hebben.
- De afhankelijkheidsrelaties zijn onafhankelijk van de oppervlaktevolgorde en de opbouw van de woordgroepen: het werkwoord *uitleggen* bij voorbeeld selecteert een lijdend voorwerp (in dit geval een afhankelijke vraag) aangeduid met OBJ1, en een meewerkend voorwerp *u*, aangeduid met OBJ2. In de oppervlaktevolgorde staat *u* echter tussen de persoonsvorm *zal* van de hoofdzin (aangeduid met HD(hoofd)) en het hulpwerkwoord *gaan* dat het hoofd is van een hoger werkwoordelijk complement (VC). Aldus ontstaan kruisende afhankelijkheden.
- Het vraagwoord *hoe* vervult de rol van hoofd van de afhankelijke vraag WHSUB, maar tegelijkertijd fungeert het als modificeerder binnen de deelwoordgroep PPART, die zelf weer is ingebed in die afhankelijke vraag. Deze dubbele functie wordt uitgedrukt door de twee dependentielabels WHD (hoofd van een vraagwoordconstructie) en MOD, die *hoe* met de twee moederknopen WHSUB en PPART verbinden.[13]
- Elementen in de uitloop (vergelijk Haeseryn *et al.* (1997:1397 vv.) – in andere kaders spreekt men wel van "rechtsdislocatie") worden niet beschouwd als onderdeel van de eigenlijke zinssyntaxis. Het discourse-verband tussen de "hoofdzin" en de "naar rechts verplaatste constituent" (in dit geval de naamwoordgroep *dat probleem*) wordt uitgedrukt door beide constituenten samen te nemen in een DU (voor Discourse Unit) waarin ze respectievelijk de rol van NUCL (kern) en SAT (satelliet) vervullen. Als in een latere fase van het project anaforische relaties ook geannoteerd zullen worden, kan er een verband gelegd worden tussen het cataforische (vooruitwijzende) voornaamwoord *dat* in de kern-component en de satelliet-naamwoordgroep *dat probleem*.
- Leestekens, zoals de (in de orthografische transcriptie ingevoegde) punt op de laatste regel van het voorbeeld in (2), worden in de ontleding buiten beschouwing gelaten.[14]

In de volgende paragrafen gaan we nader in op de uitgangspunten en de implementatie van het ontleedproces.

[13] In principe is het zelfs mogelijk dat een element of een constituent meer dan twee rollen vervult.

[14] Waarmee we overigens niet willen beweren dat leestekens vanuit taalkundig oogpunt oninteressant zijn – vergelijk Nunberg (1990).

4 Uitgangspunten syntactische annotatie

De syntactische structuren die samen met de andere vormen van verrijking en de CGN-geluidsfiles het Corpus Gesproken Nederlands vormen, worden halfautomatisch afgeleid. Later komen we nog terug op enige details van de praktische implementatie, maar duidelijk zij, dat het belangrijkste doel van de onderneming niet de parser (ontleedautomaat) is, maar de verzameling ontlede 'zinnen'. Dat impliceert dus een cruciaal verschil met de thema's van sommige andere artikelen in deze aflevering van *Nederlandse Taalkunde*.[15]

Om 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 theorieneutraal primair annotatieniveau in termen van afhankelijkheidsstructuren (vergelijk ook Skut *et al.* (1997)), waarbij in het algemeen nauw aansluiting gezocht wordt bij de traditionele Nederlandse zinsontleding, in casu de ANS (Haeseryn *et al.* 1997).[16]

4.1 Output: Annotatiegrafen

De taalkundige analyse van het CGN verrijkt het materiaal op twee manieren: met categoriale informatie en met informatie over afhankelijkheden. Een voorbeeld: in (3) heeft de woordgroep *zo'n beetje* de categorie NP (zelfstandignaamwoordgroep) gekregen, en bovendien is aangeduid dat die de functie van MOD (bepaling) vervult in de PPART (voltooiddeelwoordgroep) *dat zo'n beetje hebben aangepakt*. De resulterende structuren noemen we afhankelijkheidsstructuren of dependentiestructuren; het zijn, zoals we al zagen, in elk geval geen boomstructuren of constituentenstructuren in de klassieke zin.[17]

4.2 Formeel

Formeel is een CGN-dependentiestructuur $D = \langle K, T \rangle$ een gelabelde gerichte, acyclische graaf (DAG). We beschikken over disjuncte verzamelingen CAT en DEP voor de labeling van respectievelijk knopen K en takken T.[18]

- Knopen: $\mathsf{CAT} = \mathsf{POSCAT} \cup \mathsf{PHCAT}$: categorielabels ($c$-labels), de vereniging van lexicale (part-of-speech) en frasale labels.
- Takken: DEP: dependentielabels (d-labels).

We onderscheiden *gelede* en *ongelede* dependentiestructuren. Een ongelede dependentiestructuur is een knoop met een c-label uit POSCAT, met andere woorden, een subgraaf die enkel een woord bevat. De elementaire bouwstenen van gelede dependentiestructuren noemen we lokale dependentie-*domeinen*. De moederknoop van een dependentiedomein is gelabeld met een frasaal label uit PHCAT. De dochters hebben c-labels uit CAT. De d-labels voor de moeder-dochtertakken worden gevormd door een *hoofd*, samen met de *complementen* en de *modificeerders* van dat hoofd.[19]

Hoofd Het hoofd van een dependentiedomein projecteert het c-label van de moederknoop.

Complementen Het complementatiepatroon bepaalt de interpretatie van het hoofd in termen van thematische structuur. Een complement-label komt per domein hoogstens één keer voor.[20]

[15] Uit de bijdrage van Van der Beek et al. elders in dit themanummer blijkt echter dat een ontleed corpus een van de nevendoelen is van de Groningse collega's.

[16] De afhankelijkheidsstructuren van het CGN hebben zich inmiddels de facto ontwikkeld tot de standaard voor de computationele syntactische analyse van het Nederlands: vergelijk bijvoorbeeld Bouma *et al.* (2001).

[17] Puristen zouden kunnen opmerken dat in een zuivere dependentie-ontleding geen categoriale informatie thuishoort, maar dit ter zijde.

[18] In (3) staan de knooplabels in ovale hokjes en de taklabels in rechthoekige.

[19] In speciale gevallen kan een structuur toch verschillende complementen van hetzelfde type hebben, of zelfs twee hoofden. We komen daarop terug.

[20] Maar vergelijk de vorige noot.

Modificeerders Modificerende elementen laten het c-label van de moederknoop ongemoeid; ze kunnen weggelaten worden zonder effect op de thematische structuur. Een en hetzelfde modificeerder-label kan binnen een domein dan ook meerdere keren voorkomen.

4.3 Consequenties

Het samennemen van complementatie en modificatie binnen één dependentiedomein leidt tot 'ondiepe' annotatiestructuren. Enkele gevolgen:

- een nieuw domein (hiërarchisch niveau) wordt pas geopend als een nieuw hoofd daar aanleiding toe geeft;

  VOORBEELD: Ondiepe verbale projecties. We onderscheiden *finiete* en *niet-finiete* verbale projecties. De persoonsvorm is hoofd van de finiete zinstypen, de infinitief of het deelwoord van de niet-finiete. Er is dus in de finiete zin geen behoefte aan een tussenliggend VP-niveau.

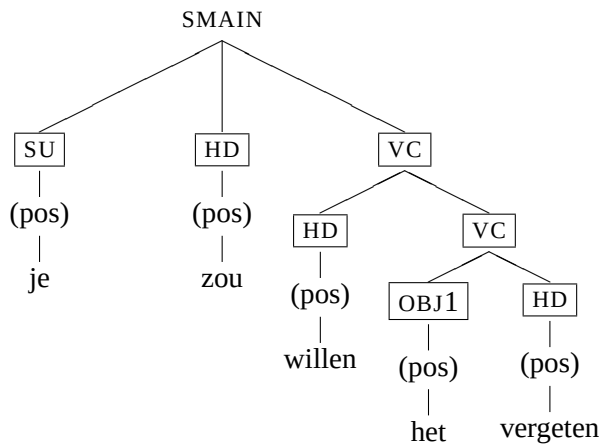
- complementatie en modificatie zijn *relaties* tussen woordgroepen en een hoofd; als er geen complementen of modificeerders zijn, is er ook geen aanleiding tot niet-vertakkende projecties;

- dependentiedomeinen zijn, in het standaardgeval, *lexicaal verankerd*: het c-label van het hoofd valt samen met de POS-tag;

  VOORBEELD: In de CGN-annotatie hebben werkwoordsgroepen op het dependentieniveau een geneste structuur, gemotiveerd door de onderscheiden subcategorisatie-eisen van de samenstellende hoofden. Bij wijze van voorbeeld: de zin *je zou het willen vergeten* krijgt de volgende dependentiestructuur (we zien even af van de oppervlakte-woordvolgorde):

  (4)

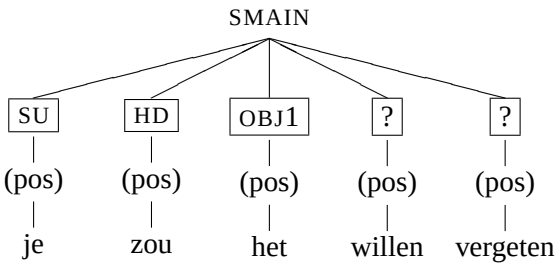
  Mogelijke uitvoer: verbaal cluster



  Op het niveau van (oppervlakte-)constituentenstructuur kan die geneste structuur evenwel ook als een 'platte' reeks werkwoorden worden uitgevoerd, bijvoorbeeld ten behoeve van een HPSG-gebruiker.

(5)

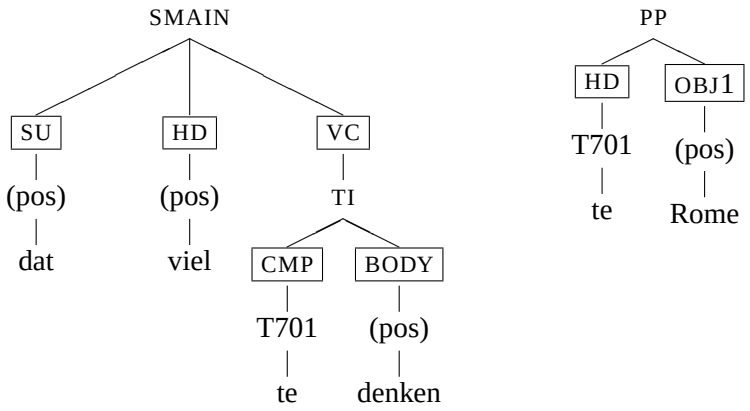
Mogelijke uitvoer: plat



- de eis dat het hoofd het c-label van de moederknoop moet projecteren, houdt in dat we het d-label van het hoofd kunnen laten disambigueren als de POS-informatie voor dat doel niet specifiek genoeg is;

  VOORBEELD: In de woordsoortentoekenning wordt geen onderscheid gemaakt tussen *te* als hoofd van een voorzetselgroep (PP) en *te* als hoofd van een niet-finiete verbale projectie, de *te*-infinitief (TI): beide worden benoemd als voorzetsel (T701). We disambigueren door middel van het d-label voor *te*: het hoofd van de TI krijgt het label CMP.[21]

(6)



- afhankelijkheidsstructuur is onafhankelijk van opervlaktevolgorde;

  VOORBEELD: Kruisende en meervoudige afhankelijkheden. We benadrukten al eerder dat de CGN-annotatie grafen oplevert en geen bomen. Grafen met kruisende takken worden gebruikt om afhankelijkheidsrelaties aan te geven die niet stroken met de oppervlaktevolgorde of met de constituentenstructuur. Constituenten kunnen bovendien meer dan één afhankelijkheidsrol krijgen, en dat is de methode die we gebruiken om niet-lokale afhankelijkheden zoals in bijvoeglijke bijzinnen en constituentvragen te representeren.[22]

  – Enerzijds bepalen de elementen die dit soort configuraties introduceren (constituenten met een vragend voornaamwoord of een betrekkelijk voornaamwoord) het c-label van hun moederknoop, dus zijn het hoofden van afhankelijkheidsstructuren.

[21] De annotatievoorbeelden hebben in dit artikel de c-labels in klein kapitaal; de d-labels zijn ingelijst. De d-labels versieren de *takken* van de annotatiegraaf: het zijn dus geen *knopen*. De POS-informatie wordt ongewijzigd van de POS-annotatie overgenomen, en hier niet altijd verder uitgespeld. Dat *te* het label CMP krijgt, net zoals bijvoorbeeld het voegwoord *dat* van finiete ingebedde zinnen, wil overigens niet zeggen dat we menen dat *dat* en *te* precies dezelfde status hebben – dat is iets voor de theoretici om te beslissen.

[22] Specificatie van andere niet-lokale afhankelijkheden, zoals de verwijzing van pronomina en de interpretatie van begrepen subjecten, wordt uitgesteld tot een latere fase van het project.

– Anderzijds willen we ook in staat zijn aan te geven wat de rol is die deze elementen vervullen in de rest van de zin; het relevante lokale afhankelijkheidsdomein kan immers willekeurig diep ingebed zijn.

Voorbeelden van kruisende en meervoudige afhankelijkheden hebben we al ontmoet in de graaf van (3).

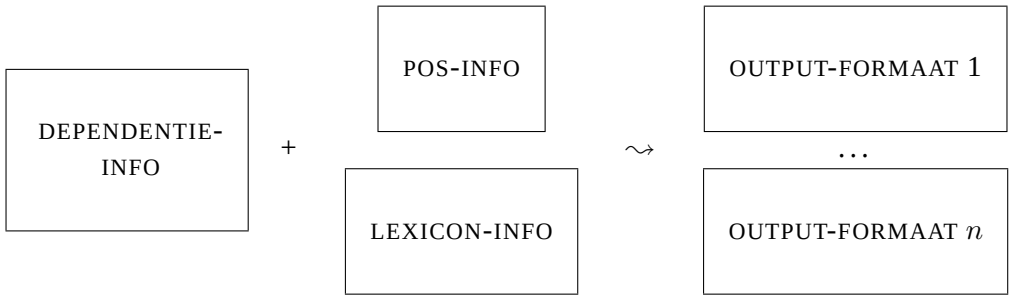
De dependentie-annotatie valt dus niet te identificeren met een constituentenstructuur – niet met een (klassieke) dieptestructuur, noch met een oppervlaktestructuur. In de filosofie van het CGN is die laatste hoogstens een *afgeleide* van de dependentie-annotatie.

De dependentiestructuur dient, eventueel na koppeling met het CGN-lexicon, wel voldoende informatie te leveren om automatisch een (oppervlakte-)constituentenstructuur als exportformaat af te leiden. Welke vorm die constituentenstructuur dan aanneemt, kan afhangen van de beoogde gebruikersgroep en van de rol die aan een constituentenstructuur wordt toebedeeld in het geheel van de CGN-annotatieniveaus (zie hieronder).

4.4 Afgeleide structuren

Deze primaire annotatiestructuren van het CGN kunnen worden verrijkt met informatie uit de taalkundige ontleding en uit het CGN-lexicon. De combinatie van deze informatiebronnen kan verschillende uitvoerformaten opleveren die meer of minder toegesneden zijn op de wensen van verschillende gebruikersgroepen.[23]

(7)



Wat betreft afgeleide output-formaten valt te denken aan:

- verrijking van de c-labels met morfosyntactische kenmerken: de verkorte labels kunnen uitgevouwen worden, zodat bijvoorbeeld ook op zulke kenmerken gezocht kan worden;
- verrijking van de d-labels met 'diepe' afhankelijkheden (zoals informatie over semantische controle: begrepen subjecten en dergelijke);
- oppervlakte-constituentenstructuren in een gebruikersvriendelijke notatie (met of zonder 'lege elementen', etc.);
- presentatie-kwesties: keuzemogelijkheden voor de 'taal' van de labels (Nederlands, Engels, ...)

Moortgat & Moot (2001) gaan nader in op de automatische conversie van de syntactische structuren van het CGN. Het zal duidelijk zijn dat sommige vormen van verrijking automatisch kunnen gebeuren, terwijl andere (soms veel) extra werk vragen.

5 Methodologie

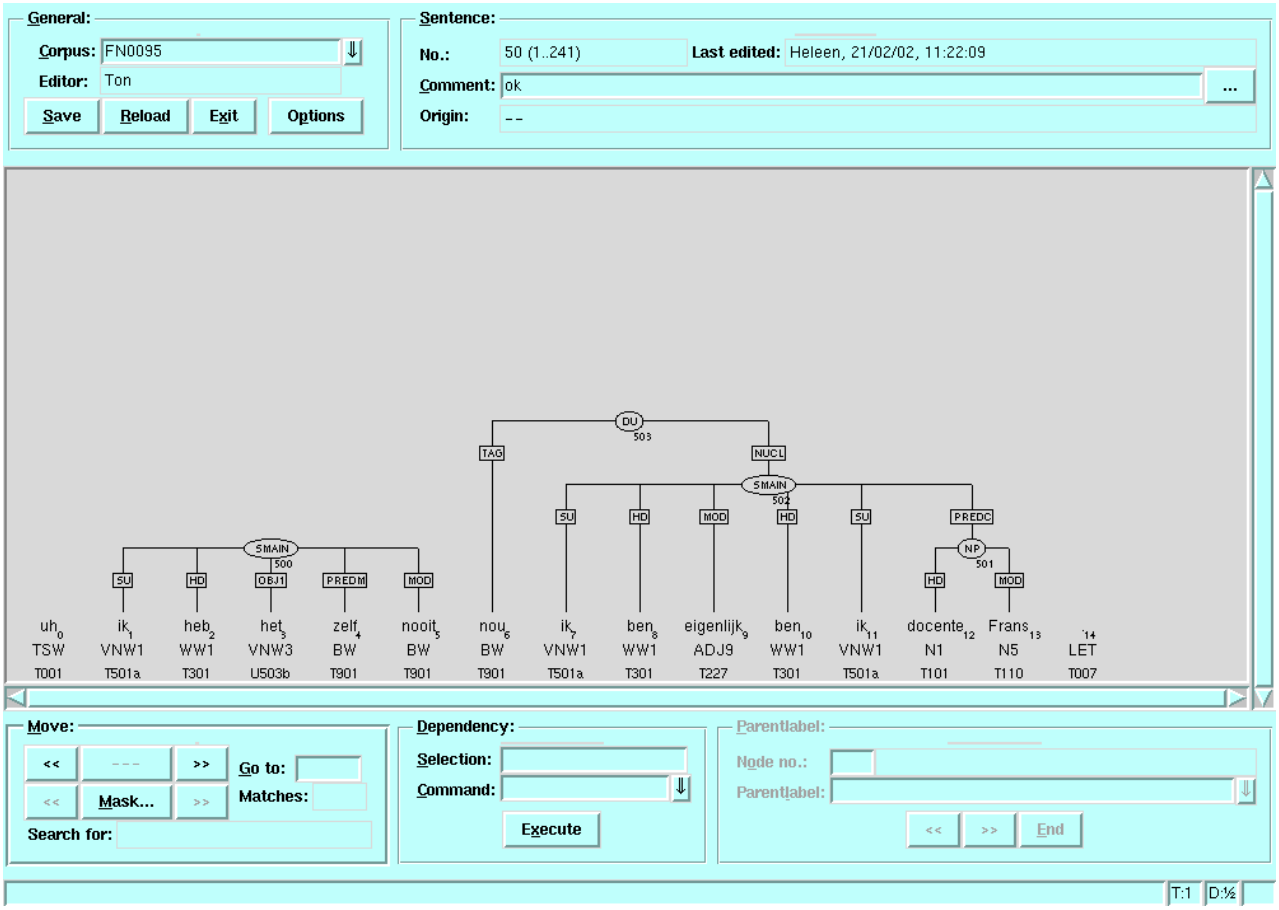
Onderdeel van de taken van het CGN is dus, als vermeld, de taalkundige ontleding van een (gebalanceerd, gestratificeerd, zo representatief mogelijk) subcorpus van een miljoen woorden. Uitgangspunt voor de ontleding is de orthografische transcriptie, verrijkt met een (grove) indeling in "annotatie-eenheden" en een fijnmazig

[23] Details van de syntactische annotatie worden expliciet gemaakt in (Moortgat *et al.* 2001); de laatste versie daarvan wordt steeds meegeleverd op de CGN-CD's.

systeem van woordsoorten. Met prosodische informatie wordt ook rekening gehouden, maar voorlopig niet in het automatische gedeelte van het proces.

Het proces geschiedt omwille van tijd en consistentie semi-automatisch. We maken gebruik van het programma ANNOTATE, dat in Saarbrücken ontwikkeld is als onderdeel van de zogenaamde NEGRA-tools.[24] De afbeelding in (8) laat zien hoe ANNOTATE zich aan de gebruiker presenteert.

(8) **Annotate**



ANNOTATE is de buitenkant van een annotatie-omgeving: annotatoren kunnen het programma gebruiken om boompjes, of in ons geval grafen, te construeren voor zinnen. ANNOTATE kan echter ook "samenwerken" met parsers (ontleedautomaten) van verschillende typen om automatisch, of eventueel na menselijke controle, zinnen te ontleden. Er is voorlopig gekozen voor een ontleder die standaard bij ANNOTATE geleverd wordt, te weten een zelflerend systeem dat ontwikkeld is door Thorsten Brants. Deze ontleedautomaat ontwikkelt op basis van een corpus van ontlede zinnen (in de literatuur staat zoiets bekend als een Tree Bank) en statistiek een theorie over de grammatica die aan die zinnen ten grondslag moet (of zou kunnen) liggen.[25] Die theorie op basis van al ontlede zinnen – in de literatuur bekend als *taalmodel* – kan dan gebruikt worden om het systeem voorstellen te laten doen over mogelijke ontledingen van nieuwe zinnen. Die ontlede zinnen kunnen dan, na controle en eventuele correctie, weer toegevoegd worden aan het corpus, op basis waarvan dan weer een nieuw taalmodel kan worden gegenereerd.

Naarmate er meer werk verzet is, is zo'n statistische grammatica natuurlijk betrouwbaarder – dat is tenminste wat je zou verwachten en wat gerapporteerd wordt voor het Duits. In de praktijk blijkt dat overigens niet eens mee te vallen. Toegegeven, de ontleder heeft nog maar zelden problemen met bijvoorbeeld de correcte ontleding van (eenvoudige) zelfstandignaamwoordgroepen en voorzetselgroepen, maar met hogere structuren

[24] Vergelijk Plaehn (1998) en http://www.coli.uni-sb.de/sfb378/negra-corpus/annotate.html.

[25] Het onderliggend mechanisme van de ontleedautomaat maakt gebruik van gestapelde Markov-modellen (Cascaded Markov Models of CMM's (Brants 1999)).

gaat het nog steeds teleurstellend vaak mis. Dat dient vermoedelijk ten dele te worden toegeschreven aan het feit dat de ontleder van 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 gedeelte van die 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 dialogen en multilogen die bij informanten thuis in de huiselijke kring zijn opgenomen, en de monologen die in de Tweede Kamer zijn opgenomen. Van elk van deze drie teksttypen geven we hieronder een klein fragment.[26]

(9) **interview met leerkracht Nederlands**

A. uhm lees*a leest u zelf veel uh.

B. mm-hu ja. ja. kranten lezen tijdschriften lezen. ja. ja dat is toch effe voor ’t ook een beetje voor ’t werk en ja.

(10) **informanten spelen Scrabble**

A. oh d’r zijn d’r nog iets van vier of zo.

B. drie of vier. nou. ga toch gewoon effe klein uh verder of zo.

A. uhm. oké maar dan zijn d’r dus ook niet erg veel klinkers meer en je hebt nu alleen maar medeklinkers. dus.

B. dan moet ’k die eerst kwijtraken.

(11) **tweedekamerlid in een commissievergadering**

de vraag of de ge*a of de minister de garantie wil geven dat de inzet en de aanpak van de sluitende aanpak dat ie niet ten koste gaat van de uh inzet en aandacht voor mensen die al nu al langdurig aan de aan de kant aan ’t werk*x die garantie is mijn vraag.

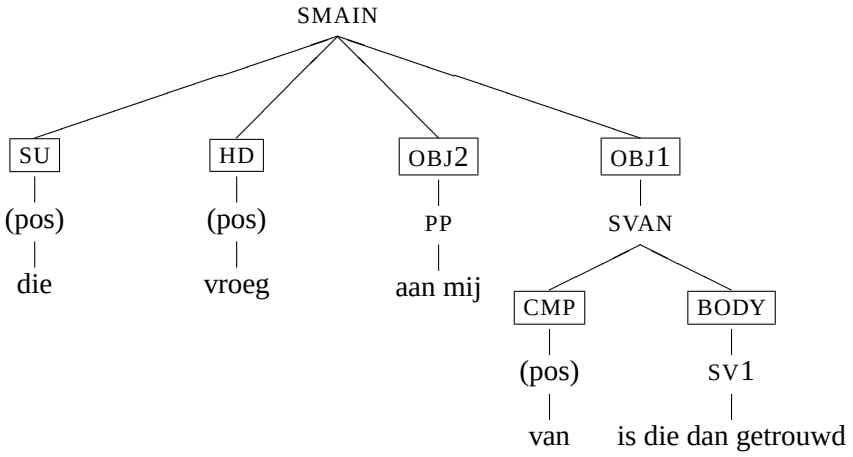
Er bestaan grote taalkundige verschillen tussen de diverse subcorpora van het CGN (Hoekstra *et al.* 2001a; van der Wouden *et al.* 2002): de mate van zinsinbedding bijvoorbeeld is in de parlementaire teksten veel groter dan in de interviews met de leraren of de spontane gesprekken. Omgekeerd vinden we in de parlementaire teksten nauwelijks het gebruik van *van* als voegwoord:[27]

(12) die vroeg aan mij van: is die dan getrouwd?

[26] In de orthografische transcriptie worden de volgende labels gebruikt (Goedertier & Goddijn 2000): *v voor woorden uit een vreemde taal, *d voor dialectwoorden, *z voor dialectisch uitgesproken woorden uit de standaardtaal, *n voor nieuwe woorden (woorden die nog niet in het CGN-lexicon voorkomen), *t voor tussenwerpsels, *a voor afgebroken woorden, *u voor afwijkende uitspraak en versprekingen, *x voor woorden waarvan de transcribeur niet zeker is.

[27] Het is ook mogelijk om dit *van* – dat volgens van den Toorn *et al.* (1997:529) al in de negentiende eeuw voorkwam – op te vatten als de hoorbare tegenhanger van de dubbele punt (Romijn 1999), maar nogmaals (cf. noot 21), dat is niet aan ons om te beslissen.

(13)



Gegeven nu dat er zulke grote verschillen zijn tussen de verschillende subcorpora, is het misschien niet zo verrassend dat het automatisch ontleden niet zo goed gaat. Het ligt immers in de rede dat een statistische ontleedautomaat die getraind is op parlementaire zinnen niet zo heel veel raad weet met een dialoog-zin als (12). Omgekeerd kan training op spontane dialogen er gemakkelijk toe leiden dat bijvoorbeeld een *dat* dat een bijzin inleidt niet als eerste als zodanig zal worden herkend, omdat andere gebruiken van *dat* veel vaker in het trainingscorpus voorkwamen.

6 Spreektaalfenomenen

Gesproken taal kent een aantal constructies die in verzorgde geschreven taal in het algemeen als onwelgevormd worden beschouwd en daarom zelden in de literatuur besproken worden (maar vergelijk de Vries (1911); Jansen (1981); de Vries (2001) enzovoorts). Een zo'n fenomeen, het "expletief" of "performatief" gebruik van *van*, kwam hierboven al aan de orde in de bespreking van zinnen als (12). Maar er is nog wel meer. Zinsdelen kunnen bijvoorbeeld om discourse- of performance-redenen weggelaten of verdubbeld worden:

(14) a. Is goed! ('topic-drop': zin zonder onderwerp)

b. Doen we ('topic-drop': zin zonder lijdend voorwerp)

c. Zijn vader beschuldigt hem dat hij zijn moeder vermoord heeft (zin zonder voorlopig voorzetselvoorwerp)[28]

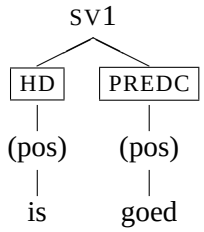
d. ik ben eigenlijk ben ik docente Frans ('spiegelconstructie')[29]

Het is niet aan het CGN om te bepalen wat correct Nederlands is en wat niet. Alles wat in het corpus opgenomen is, wordt in principe als grammaticaal beschouwd en moet dus een ontleding krijgen. De ongewone zinstypen in (14) worden dan ook 'gewoon' ontleed: als het zo uitkomt, dan maar zonder onderwerp of lijdend voorwerp, of juist met twee onderwerpen, verbale hoofden enzovoorts. Voor de zin in (14a) wordt dat geïllustreerd in (15), voor de analyse van (14d) verwijzen we naar (8).

[28]Een standaardvoorbeeld van dit type constructie is *bananen ben ik dol op*, reden waarom van der Horst & van der Horst (1999:268–272) haar aanduiden met de term 'bananenzin'. Zij gaan ook in op de geschiedenis van dit soort constructies.

[29]De term 'spiegelconstructie' is van Huesken (2001); Jansen (1981:h.7) spreekt in navolging van de Vries (1911) van 'herhalingskonstrukties', terwijl de ANS de term 'overloopconstructies' gebruikt (Haeseryn *et al.* 1997:21.2.5, 1259 vv.).

(15)



Dit soort spreektaalconstructies dient overigens scherp te worden onderscheiden van een ander spreektaalfenomeen, 'performance-errors' en de daarop volgende 'reparaties':

(16) a. Bij een huwelijk was het vroeger gemakkelijk gezegd: tot de dood hon*u ons scheidt hè.

b. En het is waarschijnlijk het uh misschien het eten van sushi of ik weet niet wat

In (16a) zegt de spreker *hon* waar hij of zij kennelijk *on*(*s*) bedoelt; de fout wordt in elk geval onmiddellijk hersteld. In (16b) begint hij of zij met *het is waarschijnlijk*, bedenkt vervolgens dat dat tot een tè sterke uitspraak gaat voeren, en vervangt het bijwoord door een iets zwakker woord, te weten *misschien*.

In deze en gelijksoortige gevallen nemen we aan dat de zinsfragmenten die gecorrigeerd worden daarmee automatisch niet tot de zin behoren; in elk geval worden ze in de syntactische annotatie ook buiten beschouwing gelaten.[30]

Wordt een verspreking niet gecorrigeerd, dan wordt het 'foute' woord gewoon in de annotatie betrokken: in zin (17a) fungeert *Maal* als zelfstandig naamwoord, en *impliciepe* in (17b) is kennelijk bedoeld als een bijvoeglijk naamwoord dat *vooronderstellingen* modificeert, en zo is het dat die woorden worden geanalyseerd.[31]

(17) a. maar ik zou d'r een enorm pleidooi voor willen houden om die Linge en die Maal*u en die Rijn een absoluut eigen ge*a uh uh eigen leven te geven.

b. het verhelderen van begrippen het opsporen van je impliciepe*u vooronderstellingen of het aangeven van de redenen die die vooronderstellingen ondersteunen

7 Toepassingen

In deze paragraaf geven we in het kort een paar voorbeelden van het soort vragen dat het corpus kan helpen beantwoorden. Deze lijst voorbeelden is natuurlijk naar believen uit te breiden.

- ik zoek alle zinnen met een vorm van het werkwoord *geven*, opgesplitst naar
  - intransitief gebruik (*Jan geeft* - bij een kaartspelletje bijvoorbeeld);
  - transitief gebruik (*Jan geeft een feestje*);
  - ditransitief gebruik (*Jan geeft de hond een koekje, Jan geeft een koekje aan de hond*);
  - onpersoonlijk gebruik (*het geeft niks*);
  - andere mogelijkheden?

[30] Een reviewer van *Nederlandse Taalkunde* is bang dat dit soort zelfcorrecties nu niet meer automatisch terug te vinden is, maar die angst lijkt ongegrond: als de zoektaal maar sterk genoeg is (zie hieronder), dan moet het bijvoorbeeld mogelijk zijn om te zoeken naar woorden die wel in de zin maar niet in de syntactische boom voorkomen.

[31] Ook pauzevullers als *uh* vallen in onze visie buiten de syntaxis – waarmee niet gezegd wil zijn dat een goedgeplaatst, goedgetimed *uh* geen (pragmatische of andere) functie zou kunnen vervullen.

- ik zoek echte voorbeelden van lange Wh-verplaatsing (*met wie zei je dat je dacht dat de kroonprins gaat trouwen?*).
- ik zoek zinnen met inbeddingen in inbeddingen (*Jan zei dat Piet klaagde dat Henk snurkte; ken jij iemand die iets geschreven heeft dat over meervoudige inbedding gaat? de uit de met een rieten dak getooide villa ontvreemde kunstschatten*).
- ik ben geïnteresseerd in het intonatiepatroon van zinnen met een kale NP in de rol van direct object op de eerste plaats (*boterhammen lust ik graag*).
- er wordt beweerd dat modale partikels (woordjes als *maar, eens, even*, cf. van der Wouden (2002)) maar op één plaats in de zin voorkomen (Krivonosov 1963; de Vriendt & Van de Craen 1986). Zijn er zinnen die daartegen pleiten, bijvoorbeeld zinnen van de structuur

  X [$_{\text{ADVP}}$ p1 p2] Y [$_{\text{ADVP}}$ p3 p4] Z

  met Y niet leeg, en de beide partikelgroepen (aangeduid met ADVP) en Y dochters van dezelfde moederknoop? Een (onacceptabel) geconstrueerd voorbeeld van zo'n zin is:

  ga nu [$_{\text{ADVP}}$ eerst maar] in die stoel [$_{\text{ADVP}}$ eens even] zitten

  waarbij het (welgevormde) cluster *eerst maar eens even* doorbroken wordt door een voorzetselconstituent *in die stoel*.
- is er een voorkeursvolgorde voor de complementen bij ditransitieve werkwoorden?
- is /n/-deletie bij nomina gevoelig voor het onderscheid onderwerp-lijdend voorwerp?
- wat voor intonatiepatronen vind je zoal bij de balansschikking?

Het nog niet altijd triviaal is om een antwoord te krijgen op zulke vragen. Nogmaals, het corpus is nog in opbouw, en de exploratiesoftware, de hulpmiddelen om het corpus daadwerkelijk te raadplegen, is nog niet af, zodat het op dit moment soms behoorlijk lastig kan zijn, verschillende informatielagen met elkaar in verband te brengen. Met name de syntactische annotatie is op dit moment nog niet toegankelijk via het exploratieprogramma COREX (Kilpatrick & Hellwig 2002).

8 Besluit

In deze bijdrage zijn we ingegaan op doelstellingen, uitgangspunten en praktische details van het onderdeel syntactische annotatie van het Corpus Gesproken Nederlands. We hebben besproken hoe dat corpus tot stand komt, we zijn kort ingegaan op het soort problemen dat men bij het analyseren van echte spreektaal tegenkomt, en we hebben getracht de bruikbaarheid van het corpus te illustreren met voorbeelden van het soort vragen dat taalkundigen, ongeacht hun theoretische voorkeuren, met dit nieuwe instrument kunnen gaan stellen. De praktijk zal ongetwijfeld uitwijzen dat die taalkundigen nog veel creatiever zullen zijn in het gebruik van het corpus dan bij de samenstelling en verrijking ervan bedacht kon worden.

References

BOUMA, GOSSE, GERTJAN VAN NOORD, & ROBERT MALOUF, 2001. Alpino: Wide-coverage computational analysis of Dutch. via odur.let.rug.nl/alfa/papers/papers/.

BRANTS, THORSTEN, 1999. Cascaded Markov Models. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics EACL-99*, Bergen, Norway, 1999.

DE VRIENDT, SERA, & PIET VAN DE CRAEN. 1986. Over plaatsingsmogelijkheden van schakeringspartikels. *Interdisciplinair Tijdschrift voor Taal- en Tekstwetenschap* 6, 101–116.

DE VRIES, JELLE. 2001. *Onze Nederlandse Spreektaal*. Den Haag: SDU Uitgevers.

DE VRIES, WOBBE, 1911. Dysmelie. Opmerkingen over syntaxis (vervolg). Verhandeling behoorende bij het programma van het gymnasium der gemeente Groningen voor het jaar 1911–1912.

GOEDERTIER, WIM, & SIMO GODDIJN, 2000. Protocol voor orthografische transcriptie. Interne publicatie CGN-project, beschikbaar via http://lands.let.kun.nl.

——, ——, & JEAN-PIERRE MARTENS, 2000. Orthographic transcription of the Spoken Dutch Corpus. Proceedings LREC 2000.

HAESERYN, WALTER, & OTHERS (eds.). 1997. *Algemene Nederlandse Spraakkunst*. Groningen and Deurne: Martinus Nijhoff and Wolters Plantijn. 2e, geheel herz. dr.

HOEKSTRA, HELEEN, MICHAEL MOORTGAT, BRAM RENMANS, INEKE SCHUURMAN, & TON VAN DER WOUDEN. 2001a. On certain syntactic properties of spoken Dutch. Paper delivered at Computational Linguistics in the Netherlands, Enschede, November 2001.

——, MICHAEL MOORTGAT, INEKE SCHUURMAN, & TON VAN DER WOUDEN. 2001b. Syntactic Annotation for the Spoken Dutch Corpus Project (CGN). In *Computational Linguistics in the Netherlands 2000*, ed. by W. Daelemans, K. Sima'an, J. Veenstra, & J. Zavrel, 73–87. Amsterdam/New York: Rodopi.

HUESKEN, NICOLE. 2001. Mirrorsentences. Repetition of inflected verb and subject in Spoken Dutch. Master's thesis, Utrecht University, General Linguistics. www.let.uu.nl/~Nicole.Huesken/personal/scriptie/scriptie.pdf.

JANSEN, FRANK. 1981. *Syntaktische konstrukties in gesproken taal*. Leiden dissertation.

KILPATRICK, PAUL, & BIRGIT HELLWIG, 2002. Corpus Gesproken Nederlands (COREX) version 1.4 Manual. CGN-CD.

KRIVONOSOV, ALEKSEJ T. 1963. *Die modalen Partikeln in der deutschen Gegenwartssprache*. Humboldt Universität Berlin dissertation. published Göppingen (1977): Kümmerle.

MILLER, JIM, & REGINA WEINERT. 1998. *Spontaneous spoken speech. Syntax and Discourse*. Oxford: Clarendon.

MOORTGAT, MICHAEL, & RICHARD MOOT. 2001. CGN to Grail. extracting a type-logical lexicon from the CGN annotation. In *Computational Linguistics in the Netherlands 2000*, ed. by W. Daelemans, K. Sima'an, J. Veenstra, & J. Zavrel, 126–143. Amsterdam/New York: Rodopi.

——, INEKE SCHUURMAN, & TON VAN DER WOUDEN, 2001. Syntactische annotatie. Internal working document CGN, Utrecht, May 2001.

NUNBERG, GEOFFREY. 1990. *The linguistics of punctuation*. Menlo Park, Calif. [etc.]: Center for the Study of Language and Information. (CSLI lecture notes 18).

OOSTDIJK, NELLEKE. 2000a. Building a corpus of spoken Dutch. In *Computational Linguistics in the Netherlands 1999. Selected Papers from the Tenth CLIN Meeting*, ed. by Paola Monachesi, 147–157. Utrecht: Utrecht University, Utrecht Institute of Linguistics OTS.

——. 2000b. Het Corpus Gesproken Nederlands. *Nederlandse Taalkunde* 5, 280–284.

PLAEHN, OLIVER, 1998. Annotate: Bedienungsanleitung. Document Projekt C3 Nebenläufige Grammatische Verarbeitung. Universität des Saarlandes, FR 8.7 Computerlinguistik.

ROMIJN, KIRSTEN. 1999. Ik schrijf van niet, maar ik zeg van wel. *TABU* 29, 173–178.

SKUT, W., B. KRENN, & H. UZKOREIT. 1997. An annotation scheme for free word order languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*. Washington, D.C.: . available via arxiv.org/format/cmp-lg/9702004.

VAN DEN TOORN, M.C., & OTHERS (eds.). 1997. *Geschiedenis van de Nederlandse taal*. Amsterdam: Amsterdam University Press.

VAN DER HORST, JOOP, & KEES VAN DER HORST. 1999. *Geschiedenis van het Nederlands in de twintigste eeuw*. Den Haag/Antwerpen: Sdu/Standaard.

VAN DER WOUDEN, TON. 2002. Partikels: naar een partikelwoordenboek voor het Nederlands. *Nederlandse Taalkunde* 7, 20–43.

——, HELEEN HOEKSTRA, MICHAEL MOORTGAT, BRAM RENMANS, & INEKE SCHUURMAN. 2002. Syntactic Analysis in the Spoken Dutch Corpus (CGN). In *Proceedings of the third International Conference on Language Resources and Evaluation*, ed. by Manuel González Rodríguez & Carmen Paz Suárez Araujo, 768–773. Paris: ELRA.

VAN EYNDE, FRANK, 2001. Part of speech tagging en lemmatisering. Technical report, Centrum voor Computerlinguïstiek K.U. Leuven, http://lands.let.kun.nl/cgn/publicat.htm.

——, JAKUB ZAVREL, & WALTER DAELEMANS. 2000. Lemmatisation and morphosyntactic annotation for the Spoken Dutch Corpus. In *Computational Linguistics in the Netherlands 1999. Selected Papers from the Tenth CLIN Meeting*, ed. by Paola Monachesi, 53–62. Utrecht: Utrecht University, Utrecht Institute of Linguistics OTS.