Van de redactie

Sinds het moment dat het Corpus Gesproken Nederlands (CGN) voor onderzoeksdoeleinden beschikbaar was – de definitieve versie kwam gereed in 2004, maar daarvoor waren al voorlopige versies in omloop – is het corpus veelvuldig gebruikt voor het beantwoorden van heel uiteenlopende onderzoeksvragen. Daarvan getuigen ook verschillende artikelen in de afgelopen jaargangen van *Nederlandse Taalkunde*. Het leek de redactie daarom nuttig om in een themanummer expliciet aandacht te besteden aan het CGN, met name aan de mogelijkheden en beperkingen ervan voor taalkundig onderzoek. Verder bevat dit nummer in de rubriek DigiTaal een bijdrage van Evie Coussé over digitale bronnen voor taalhistorisch onderzoek van het Nederlands, en de gebruikelijke, andere rubrieken.

In een uitgebreid, inleidend artikel schetst Laura van Eerten van de centrale voor Taal- en Spraaktechnologie (TST-centrale), die tegenwoordig het CGN beheert, de achtergronden en ontstaansgeschiedenis van het corpus. Daarnaast biedt haar bijdrage, aan de hand van een specifieke onderzoeksvraag, een gedetailleerde introductie tot het zoekprogramma Corex. De drie overige artikelen doen verslag van onderzoek dat met behulp van het CGN is uitgevoerd. Albert Oosterhof en Evie Coussé bespreken de rol van corpusonderzoek in de formele semantiek en demonstreren de voor- en nadelen van het CGN aan de hand van een aantal *case studies*. Ton van der Wouden gebruikt het CGN om te bepalen welke woordcombinaties in het gesproken Nederlands fungeren als 'extended lexical units' – wat nieuw licht moet werpen op de taakverdeling tussen grammatica en lexicon. Ronny Boogaart e.a. doen verslag van hun onderzoek in het CGN naar verschillen tussen Belgisch Nederlands en Nederlands Nederlands in het gebruik van modale uitdrukkingen.

We realiseren ons dat het CGN gebruikt wordt in meer, ook andersoortig, onderzoek dan in dit themanummer aan bod komt. Dit zal ongetwijfeld duidelijk worden in de komende jaargangen van *Nederlandse Taalkunde*.

Over het Corpus Gesproken Nederlands

LAURA VAN EERTEN*

Abstract

Resulting in the largest speech corpus of Dutch (+/- 9 million words), the construction of the *Spoken Dutch Corpus* (CGN) can be considered a unique project. Financed by the Dutch and the Flemish Government the CGN was successfully completed in 2004. This paper presents an overview of the CGN, from the beginning of the project in 1998 until now. First the project background – motivation and project organisation – will be described. Then a detailed description of the corpus contents will be presented, followed by an introduction to the corpus exploitation software Corex. A small-scale Corex experiment is carried out for further illustration. Besides, an overview of divergent corpus-based research using the CGN is given. The paper concludes with practical information about distribution by the *TST-centrale,* and future perspectives such as an online web version of the corpus and the additional JASMIN-CGN project.

1 Inleiding

Sinds jaar en dag worden grote verzamelingen data gebruikt voor wetenschappelijk onderzoek. Deze verzamelingen, veelal omvangrijke tekstbestanden, worden corpora genoemd. De ontwikkeling van het Corpus Gesproken Nederlands (CGN) was daarom niet direct een vernieuwend project, maar wel uniek: nog niet eerder was er zo een grote hoeveelheid gesproken Nederlandse taaldata bij elkaar gebracht. Enkele Nederlandse en Vlaamse universiteiten hebben vijf jaar lang gewerkt aan het verzamelen, catalogiseren en annoteren van de ongeveer negenhonderd uur spraak die het corpus bevat. Het CGN-project, dat gefinancierd werd door NWO en de overheid van Nederland en Vlaanderen, is in 2004 succesvol afgerond. Een representatieve steekproef van het levende Nederlands rond de laatste eeuwwisseling was hiermee een feit.

In dit artikel wordt een overzicht gegeven van het Corpus Gesproken Nederlands vanaf de start van het project tot nu. Eerst wordt in paragraaf 2 de achtergrond – motivatie en

* De auteur is als taalkundige verbonden aan het Instituut voor Nederlandse Lexicologie afdeling TST-centrale, Matthias de Vrieshof 2-3, 2311 BZ Leiden. E-mailadres: eerten@inl.nl. Met dank aan Griet Depoorter voor het kritisch lezen en becommentariëren van de concepttekst, het uitvoeren van zoekacties in Corex en het maken van schermafbeeldingen. Tevens dank aan de redactie van Nederlandse Taalkunde voor suggesties en commentaar bij een eerdere versie van dit artikel. Uitgebreide informatie en documentatie over het Corpus Gesproken Nederlands is te vinden op de website van de TST-centrale: www.tst.inl.nl, onder Producten. Vragen of opmerkingen over het CGN kunnen ge-e-maild worden naar: servicedesk@inl.nl.

projectorganisatie – van het project uiteengezet. Paragraaf 3 behandelt de opbouw en de inhoud van het corpus. In paragraaf 4 wordt mede aan de hand van een uitgewerkte onderzoeksvraag een introductie gegeven in het zoekprogramma Corex. Vervolgens wordt in paragraaf 5 een overzicht gegeven van een aantal onderzoeken waarin gebruikgemaakt is van het CGN. In paragraaf 6 wordt ten slotte praktische informatie verstrekt over onderhoud en distributie bij de TST-centrale, gevolgd door toekomstige plannen zoals een online webversie van het CGN en het aanvullende project JASMIN-CGN.

2 Achtergrond

2.1 Motivatie

De voornaamste reden voor de aanvang van het project in 1998, was het versterken van de positie van het Nederlands in de taal- en spraaktechnologie. Voor het Engels waren er al veel taalbronnen beschikbaar zoals het *British National Corpus* en het *Wall Street Journal text corpus.* De beschikbaarheid van deze bronnen heeft de Engelse taal- en spraaktechnologie in een leidende positie geplaatst. Voor de ontwikkeling van de Nederlandse taal- en spraaktechnologie ontbrak dergelijk materiaal. Het project Corpus Gesproken Nederlands werd opgezet om dit hiaat op te vullen.

Naast de taaltechnologische motivatie was er ook vanuit het linguïstische onderzoeksveld vraag naar een Nederlands spraakcorpus. Tot voor de start van het project waren er voor taalkundig onderzoek alleen tekstcorpora beschikbaar. Zoals iedereen die zich met taal bezighoudt weet, is gesproken taal in bepaalde opzichten veel complexer dan geschreven taal: onvolledige zinnen, versprekingen, aarzelingen, interjecties en wederzijdse beïnvloeding van klanken zijn veelvoorkomende fenomenen. Er was nog geen corpus beschikbaar waarin dit soort eigenschappen van gesproken Nederlands onderzocht kon worden.

Een derde argument voor de ontwikkeling van het corpus had betrekking op het educatieve belang. Toepassing van gesproken taalmateriaal in het onderwijs biedt nieuwe perspectieven. In het onderwijs van het Nederlands wordt vaak alleen gebruik gemaakt van geschreven Nederlands, terwijl in het onderwijs van bijvoorbeeld Frans of Duits luistervaardigheid een prominent onderdeel vormt van het programma. De ontwikkeling van het CGN zou voor spraaktoepassingen binnen het Nederlandse taalonderwijs een eerste stap in de goede richting zijn.

2.2 Projectorganisatie

Het project ging in 1998 van start onder leiding van een bestuur bestaande uit zes leden, met een evenredige vertegenwoordiging uit Vlaanderen en Nederland. Het bestuur werd benoemd door de Vlaamse en Nederlandse financiers en bestond uit vertegenwoordigers van de twee overheden, Nederlandse en Vlaamse onderzoeksinstituten en de Landelijke Onderzoeksschool Taalwetenschap (LOT). Daarnaast nam een afgevaardigde van de Nederlandse Taalunie als waarnemer deel aan de bestuursvergaderingen. De voorzitter van het bestuur was Professor W. Levelt van het Max Planck Instituut te Nijmegen. Het bestuur stelde een stuurgroep in die verantwoordelijk was voor de daadwerkelijke uitvoering van het project.

De dagelijkse coördinatie was in handen van twee projectleiders: W. Goedertier (Rijksuniversiteit Gent) voor Vlaanderen en N. Oostdijk (Radboud Universiteit Nijmegen) voor Nederland. Door de projectleiders werden drie werkgroepen aangestuurd: corpusopbouw, signaalanalyse en corpusannotatie. De werkgroep corpusopbouw was verantwoordelijk voor het ontwerp en de samenstelling van het corpus, de werving van sprekers en de acquisitie van opnames. De werkgroep signaalanalyse ontwikkelde de protocollen en de procedures voor orthografische transcriptie, woordsegmentatie, fonetische transcriptie en prosodische annotatie. De verantwoording voor de daadwerkelijke uitvoering van de annotaties lag bij de werkgroep corpusannotatie (zie ook de website van het CGN http://www.tst.inl.nl/cgndocs/doc_Dutch/start.htm).

3 De inhoud van het corpus

3.1 Spraakfragmenten en annotaties

Het Corpus Gesproken Nederlands bevat bijna dertienduizend spraakfragmenten van Nederlandse en Vlaamse sprekers in verschillende taalgebruiksituaties. Naast voorgelezen teksten en fragmenten van radio en televisie, zoals nieuwsuitzendingen, (sport-) commentaren actualiteitenrubrieken en reportages, zijn er ook spontane conversaties, telefoondialogen, interviews, debatten, lezingen en Nederlandse lessen opgenomen. Deze categorieën zijn in het CGN nauwkeurig opgedeeld in vijftien verschillende componenten:

(1)
- a spontane conversaties (*face-to-face*)
- b interviews met leraren Nederlands
- c telefoondialogen opgenomen met behulp van een telefooncentrale
- d telefoondialogen opgenomen met behulp van een minidiskrecorder
- e zakelijke onderhandelingen
- f interviews en discussies uitgezonden op radio en televisie
- g discussies, debatten, vergaderingen (met name politieke)
- h lessen (middelbare school met focus op leerkracht)
- i spontane commentaren (onder andere sport) uitgezonden op radio en televisie
- j actualiteitenrubrieken en reportages uitgezonden op radio en televisie
- k nieuwsbulletins uitgezonden op radio en televisie
- l beschouwingen en commentaren uitgezonden op radio en televisie
- m missen, lezingen, plechtige toespraken
- n colleges, voordrachten, lezingen
- o voorgelezen teksten

Bij de samenstelling van het corpus is zo veel mogelijk rekening gehouden met wensen en behoeftes van diverse groepen potentiële gebruikers, in plaats van één specifieke doelgroep te bedienen zoals dat bij eerder samengestelde corpora vaak het geval was. Er bleek vooral vraag te zijn naar spontane spraakdata waarin interactie als belangrijk bestanddeel gezien werd. In het corpus zijn spontane dialogen en multilogen daarom in ruime mate vertegenwoordigd. In eerste instantie was het de bedoeling om een corpus samen te stellen dat uit-

sluitend spontane spraak bevatte. Door de grote behoefte die er onder spraaktechnologen bestaat aan voorgelezen teksten voor de training van spraakherkenners, is ervoor gekozen om deze vorm van gesproken Nederlands ook op te nemen in de data (Oostdijk 2000).

De uiteindelijke structuur van het corpus is het resultaat van een 'getrapte sampling'. Tabel 1 laat zien hoe de componenten tot stand gekomen zijn door enerzijds *dialoog/multiloog* en anderzijds *monoloog* uit te splitsen naar steeds specifiekere componenten.

Sampling					***Component***
dialoog / multiloog	privé		spontaan	direct	*a.* spontane conversaties
					b. interviews leraren Nederlands
				indirect	*c.* telefoondialogen (centrale)
					d. telefoondialogen (minidisk)
					e. zakelijke onderhandelingen
	publiek	uitgezonden	min of meer voorbereid		*f.* interviews en discussies
		niet uitgezonden	spontaan		*g.* discussies, debatten, vergaderingen
					h. lessen
monoloog	privé		min of meer voorbereid		*
	publiek	uitgezonden	spontaan		*i.* spontaan commentaar
			min of meer voorbereid		*j.* actualiteitenrubrieken, reportages
					k. nieuwsbulletins
					l. beschouwingen, commentaren
		niet uitgezonden	min of meer voorbereid		*m.* lezingen, toespraken
					n. colleges, voordrachten
					o. voorgelezen tekst

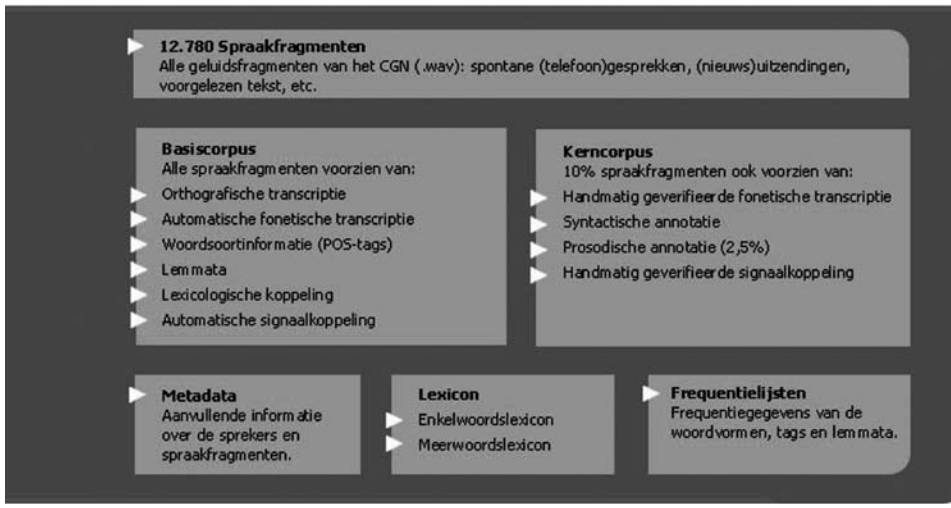
Tabel 1: Het definitieve corpusontwerp.

In het oorspronkelijke ontwerp was er nog een component voorbereid op de plaats van de asterisk: beschrijvingen van route of plaatjes. De realisatie van het corpus liep echter vertraging op waardoor de omvang en samenstelling van sommige componenten bijgesteld moest worden. Deze component kwam daarmee te vervallen. De tweedeling in de component telefoondialogen is ook later ingevoerd door onvoorziene technische problemen met de telefooncentrale.

De omvang van de fragmenten loopt per component tamelijk uiteen. Omdat het moeilijk bleek de optimale lengte voor een fragment te bepalen, hebben vooral intuïtieve factoren een rol gespeeld. De 'natuurlijke' lengte van een bepaald soort spraak speelde mee:

een nieuwsbericht op de radio is bijvoorbeeld van kortere duur dan een gesproken tekst bij een documentaire (Oostdijk 2000).

Aan de spraakfragmenten is een aanzienlijke hoeveelheid taalkundige informatie toegevoegd in diverse annotatielagen. Echter niet elk fragment is even 'rijk' geannoteerd. Op basis van de verschillende annotaties werd onderscheid gemaakt tussen het basiscorpus en het kerncorpus. De term *basiscorpus* verwijst naar het volledige corpus en omvat alle fragmenten van het CGN voorzien van orthografische en automatisch gegenereerde fonetische transcripties, woordsoortinformatie, lemma-informatie en de automatische verbinding van transcripties met het geluidssignaal (signaalkoppeling). Onder het *kerncorpus* verstaat men een uitvoeriger verrijkte selectie van het basismateriaal, waaronder de handmatig geverifieerde fonetische transcripties en signaalkoppeling, de syntactische annotaties en de prosodische informatie. In figuur 1 is de indeling van het complete CGN schematisch weergegeven.



Figuur 1: Schematische weergave van het CGN.

Het basiscorpus bevat alle orthografisch getranscribeerde spraakfragmenten. Alle spraak is per spreker uitgeschreven inclusief herhalingen, versprekingen en afgebroken woorden. Voor onder andere dialectwoorden, niet-Nederlandse woorden en sprekergeluiden zoals lachen of hoesten, is een afzonderlijke markering toegepast. De transcripties zijn vervaardigd met behulp van het programma Praat (Boersma & Weenink 1992-2007), en kunnen hierin – onafhankelijk van de exploitatiesoftware Corex (zie paragraaf 4) – samen met het geluidsbestand geopend worden.

Op basis van de orthografische transcripties is het volledige corpus automatisch verrijkt met lemmata en woordsoortinformatie (POS-tags). Binnen het project werd daarvoor een eigen CGN-tagset gedefinieerd die driehonderdzestien verschillende tags omvat, en aansluit bij de praktijk van de ANS (Haeseryn en anderen 1997). De automatisch toegekende tags zijn naderhand nagekeken en waar nodig gecorrigeerd.

De fonetische transcripties zijn eveneens automatisch gegenereerd voor het gehele corpus. Een deel ervan, ongeveer tien procent, is handmatig geverifieerd. De gebruikte

foneemset vertoont veel overeenkomsten met gebruikelijke fonetische alfabetten zoals IPA en SAMPA. De fonetische transcripties zijn evenals de orthografie te bekijken in het programma Praat.

Syntactische annotaties zijn vervaardigd voor ongeveer tien procent van het opgenomen materiaal. Alle afhankelijkheidsrelaties binnen een geannoteerde zin zijn hiërarchisch in kaart gebracht in boomstructuren. De boomstructuren zijn te visualiseren met behulp van het programma *Tiger* (zie paragraaf 4) dat geïntegreerd is in het exploitatieprogramma Corex.

Voor ongeveer tweeënhalf procent van het corpus zijn prosodische verschijnselen gemarkeerd. De markering betreft prominente lettergrepen, prosodische grenzen en abnormale klankverlengingen en is aangebracht in de orthografische transcriptielaag.

De annotaties zijn op woordniveau gekoppeld aan het geluidsmateriaal. Deze automatische segmentatie is nadien voor tien procent van de data gecontroleerd en handmatig gecorrigeerd. Het doel van de signaalkoppeling is om woorden in verbonden spraak van elkaar te scheiden, door grenzen of markeringen te plaatsen in het spraaksignaal. Hierdoor is het mogelijk om het spraaksignaal direct terug te vinden bij een bepaalde annotatie of andersom.

3.2 Lexica, frequentielijsten en metadata

Zoals uit figuur 1 blijkt behoren naast de spraakdata met annotaties ook een lexicon, frequentielijsten en bestanden met metadata tot het CGN. Het CGN-lexicon is een woordenlijst die vrijwel alle unieke woordvormen uit het corpus bevat. Het ontwerp is gebaseerd op bestaande elektronische bronnen zoals CELEX (Baayen en anderen 1993), het RBN (Referentiebestand Nederlands 1998) en het WNT (Woordenlijst Nederlandse Taal 1995). De *entries* in het lexicon zijn voorzien van uitgebreide lexicale informatie zoals woordsoort, lemma, syntactisch complementatiepatroon, uitspraak en morfologische segmentatie.

Het lexicon is onderverdeeld in een enkelwoordslexicon en een meerwoordslexicon. Het enkelwoordslexicon (ook wel standaardlexicon genoemd) bevat uitsluitend aaneengeschreven woordvormen. Het meerwoordslexicon bestaat uit meerledige expressies zoals scheidbaar samengestelde werkwoorden (*nemen op, halen adem*), ingeburgerde vreemdtalige uitdrukkingen (*et cetera, wishful thinking*), eigennamen en titels (*Berg En Dal, De Pfaffs*). Door middel van een lexicologische koppeling zijn in het CGN de los geschreven delen van meerledige uitdrukkingen met elkaar verbonden. Via de meerwoordslemma's zijn verwijzingen naar het lexicon opgenomen. Hierdoor zijn zoekacties op de afzonderlijke delen van een expressie mogelijk.

In verschillende lijsten zijn frequentiegegevens uit het CGN opgenomen. Naast een complete lijst van alle woordvormen die in het CGN voorkomen ('totrank'), zijn er aparte frequentielijsten waarbij onderscheid gemaakt wordt tussen Vlaamse en Nederlandse data. Daarnaast zijn frequentielijsten beschikbaar waarbij een uitsplitsing is gemaakt naar de vijftien componenten die in het corpus worden onderscheiden. Al deze frequentielijsten zijn op alfabet of frequentievolgorde gesorteerd. Van de woordsoortinformatie (POS-tags), de lemmata en woordvormen met bijbehorende fonetische transcripties zijn tevens afzonderlijke frequentielijsten beschikbaar.

RANK	TOT	TOKEN
1:	941685	.
2:	309774	ja
3:	262631	dat
4:	248660	de
5:	226189	en
6:	205181	uh
7:	184619	ik
8:	172044	een
9:	141417	is
10:	138235	die

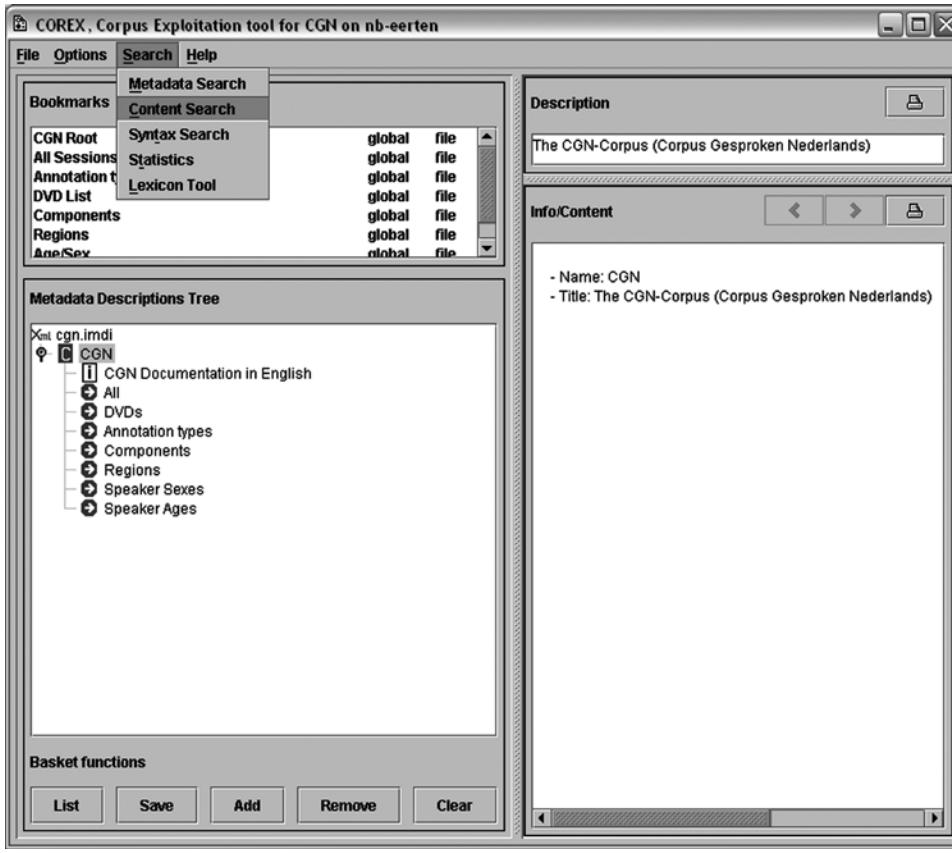
Figuur 2: Top tien van de frequentielijst 'totrank'.

Het onderdeel metadata geeft nadere informatie over de sprekers en de fragmenten die in het corpus voorkomen: van zo veel mogelijk sprekers is sociolinguïstisch relevante informatie verzameld zoals leeftijd, geboorteplaats of geboorteregio en opleidingsniveau. Met betrekking tot de opnames wordt onder andere informatie verschaft over de locatie, het gemiddelde spreektempo en de datum van opname. Dit soort aanvullende gegevens maakt het mogelijk om diverse variabelen te selecteren voor een specifiek onderzoek. Denk hierbij bijvoorbeeld aan spontane conversaties tussen Nederlandse vrouwen ouder dan 25 jaar, waarin een hoog spreektempo gehanteerd wordt. Meer voorbeelden van onderzoeksvragen en zoekacties worden aan de hand van Corex behandeld in de volgende paragraaf.

4 Het zoekprogramma Corex[1]

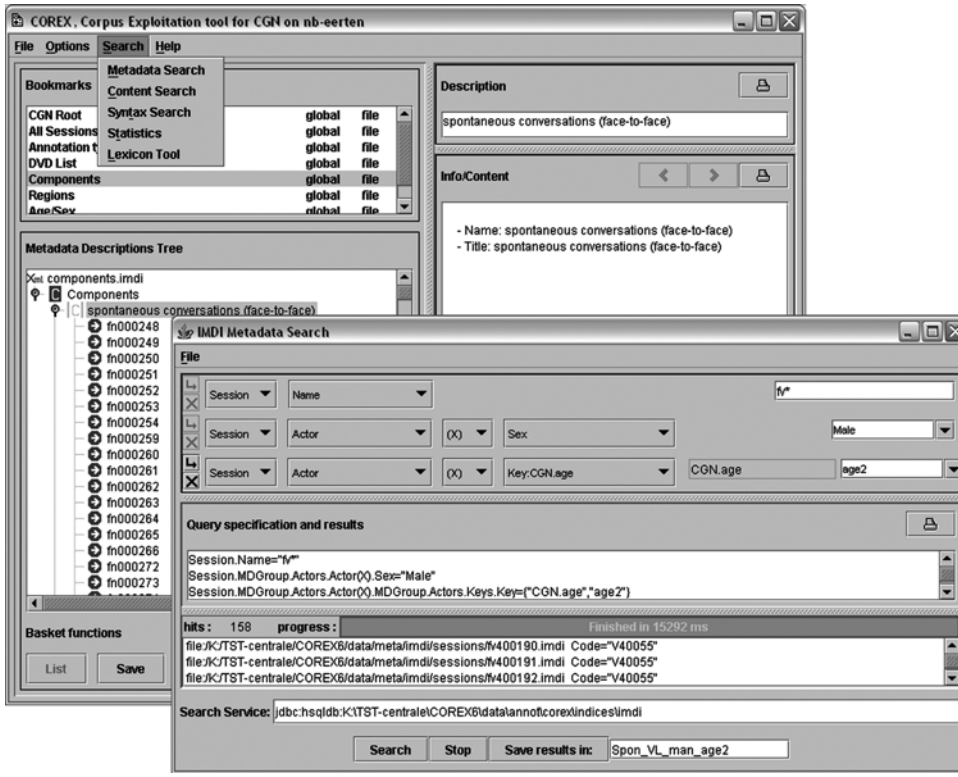
Om op een eenvoudige en efficiënte manier door de grote hoeveelheid data te kunnen navigeren, is parallel aan het CGN-project een speciaal zoekprogramma ontwikkeld: de *cor*pus*ex*ploitatiesoftware Corex. In figuur 3 wordt het openingsvenster van Corex afgebeeld van waaruit alle zoekacties starten.

1 In dit artikel is gebruikgemaakt van versie 6.1 van Corex. Het is mogelijk dat de schermen en zoekacties zoals geïllustreerd in dit artikel er anders uitzien in een oudere of nieuwere versie van het programma.



Figuur 3: Het hoofdvenster van Corex met uitgeklapt zoekmenu.

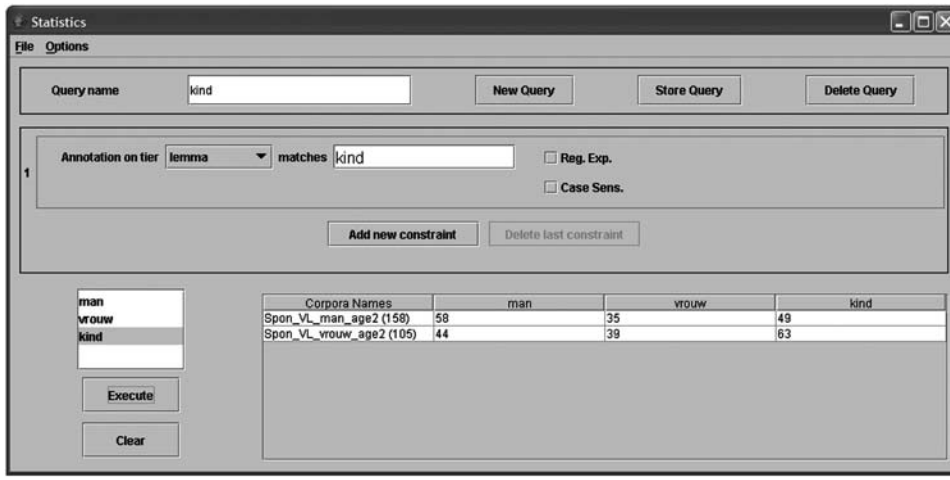
Het programma biedt veel mogelijkheden doordat er gewerkt kan worden met subcorpora, gebaseerd op voorgedefinieerde of eigen criteria, zoals het geslacht en de leeftijd van de spreker en diverse andere metadata. In bijvoorbeeld een onderzoek naar spontane conversaties tussen uitsluitend Vlaams sprekende mannen boven de vijfentwintig, kan er met behulp van Corex een selectie (subcorpus) gemaakt worden van deze groep alvorens de meer specifieke zoekacties uit te voeren. Figuur 4 illustreert hoe de samenstelling van een dergelijk subcorpus in zijn werk gaat.



Figuur 4: Samenstelling van een subcorpus van spontane conversaties tussen Vlaamse mannen in de leeftijdscategorie 25 - 34 jaar.

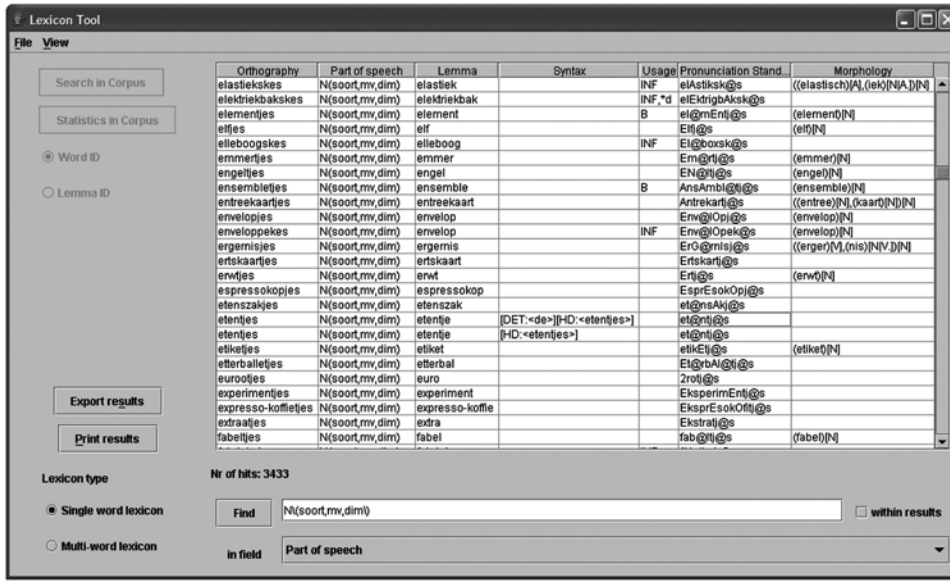
In het hoofdvenster is een eerste selectie gemaakt op basis van de component met spontane conversaties. Binnen de *Metadata Search* wordt het subcorpus verder gespecificeerd door alleen Vlaamse fragmenten ('fv*'), mannelijke sprekers en de gewenste leeftijdscategorie ('age2') te selecteren (Corex maakt hier gebruik van zogenaamde *keywords* waarvan verklarende lijsten opgenomen zijn in de data; *age2* betreft de leeftijdscategorie 25 - 34). De zoekresultaten kunnen vervolgens binnen Corex opgeslagen worden als subcorpus.

Met de zoekoptie *Statistics* kan een grote variatie aan frequentiegegevens opgevraagd worden van woordvormen, lemma's, tags, prosodische en fonetische fenomenen of een combinatie daarvan. In figuur 5 is een voorbeeld weergegeven van frequentieresultaten van de lemma's *man, vrouw* en *kind*, onder andere binnen het subcorpus uit figuur 4. De *queries* worden eerst één voor één opgeslagen en vervolgens tegelijkertijd uitgevoerd. Let er bij een zoekactie in *Statistics* wel op dat de frequenties absolute getallen zijn en daardoor een vertekend beeld kunnen geven.



Figuur 5: Frequentiegegevens van de lemma's 'man' 'vrouw' en 'kind' in spontane conversaties van Vlaamse mannen tussen de 25 en 34 jaar tegenover spontane conversaties van Vlaamse vrouwen tussen de 25 en 34 jaar.

Naast de *Metadata Search* en de *Statistics Tool* is er binnen Corex een speciale *Lexicon Tool* om in het lexicon te kunnen zoeken. In figuur 6 wordt een schermafbeelding getoond van een lijst diminutieven in de *Lexicon Tool*. Onder de menuoptie *View* kan er een keuze gemaakt worden in het aantal kolommen met woordinformatie dat getoond wordt.



Orthography	Part of speech	Lemma	Syntax	Usage	Pronunciation Stand...	Morphology
elastiekskes	N(soort,mv,dim)	elastiek		INF	elAstiksk@s	((elastisch)[A],(iek)[N]A,])[N]
elektriekbakskes	N(soort,mv,dim)	elektriekbak		INF,*d	elEktrigbAksk@s	
elementjes	N(soort,mv,dim)	element		B	el@mEntj@s	(element)[N]
eltjes	N(soort,mv,dim)	elf			Eltj@s	(elf)[N]
elleboogskes	N(soort,mv,dim)	elleboog		INF	El@boxsk@s	
emmertjes	N(soort,mv,dim)	emmer			Em@rtj@s	(emmer)[N]
engeltjes	N(soort,mv,dim)	engel			EN@ltj@s	(engel)[N]
ensembletjes	N(soort,mv,dim)	ensemble		B	AnsAmbl@tj@s	(ensemble)[N]
entreekaartjes	N(soort,mv,dim)	entreekaart			Antrekartj@s	((entree)[N],(kaart)[N])[N]
envelopjes	N(soort,mv,dim)	envelop			Env@lOpj@s	(envelop)[N]
enveloppekes	N(soort,mv,dim)	envelop		INF	Env@lOpek@s	(envelop)[N]
ergernisjes	N(soort,mv,dim)	ergernis			ErG@rnisj@s	((erger)[V],(nis)[N]V.])[N]
ertskaartjes	N(soort,mv,dim)	ertskaart			Ertskartj@s	
erwtjes	N(soort,mv,dim)	erwt			Ertj@s	(erwt)[N]
espressokopjes	N(soort,mv,dim)	espressokop			EsprEsokOpj@s	
etenszakjes	N(soort,mv,dim)	etenszak			et@nsAkj@s	
etentjes	N(soort,mv,dim)	etentje	[DET:<de>][HD:<etentjes>]		et@ntj@s	
etentjes	N(soort,mv,dim)	etentje	[HD:<etentjes>]		et@ntj@s	
etiketjes	N(soort,mv,dim)	etiket			etikEtj@s	(etiket)[N]
etterballetjes	N(soort,mv,dim)	etterbal			Et@rbAl@tj@s	
eurootjes	N(soort,mv,dim)	euro			2rotj@s	
experimentjes	N(soort,mv,dim)	experiment			EksperimEntj@s	
expresso-koffietjes	N(soort,mv,dim)	expresso-koffie			EksprEsokOftj@s	
extraatjes	N(soort,mv,dim)	extra			Ekstratj@s	
fabeltjes	N(soort,mv,dim)	fabel			fab@ltj@s	(fabel)[N]

Figuur 6: Weergave van een lijst diminutieven in de Lexicon Tool.

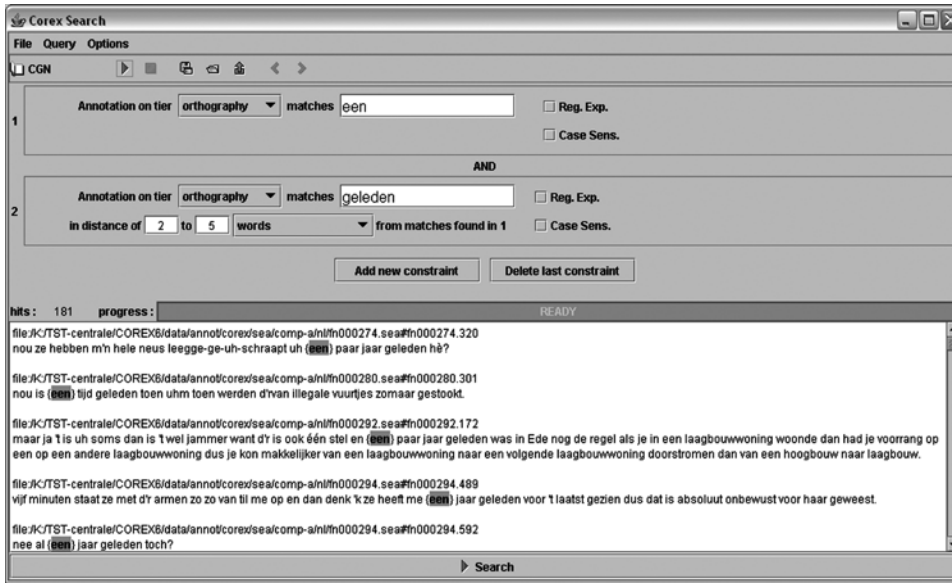
Zoals uit bovenstaande voorbeelden blijkt, benadert iedere zoekfunctie de data op een eigen, zo efficiënt mogelijke wijze. Het verschilt daarom per zoekfunctie welke informatie ingevoerd moet worden om het optimale resultaat te verkrijgen. Er zijn vaak meerdere manieren om resultaten te vinden. Daarvoor moet een tussenweg gevonden worden tussen de hoeveelheid ruis die optreedt bij een zoekactie en de hoeveelheid data die doorzocht wordt. Aangezien de syntactische annotaties toegevoegd zijn aan maar tien procent van de data, wordt bij een syntactische zoekactie alleen dit deel van het corpus doorzocht. Hierover meer in de uitgewerkte onderzoeksvraag in 4.3.

Corex kan met veel onderzoeksvragen uit de voeten, maar uiteraard zijn er beperkingen. Het is niet mogelijk om negatieve zoekacties uit te voeren, zoals zoeken op woorden die eindigen op *–ig* en die geen bijvoeglijke naamwoorden zijn. De frequentiegegevens zijn ook maar beperkt toegankelijk (zie figuur 5); voor het samenstellen of herschikken van een bepaalde frequentielijst moet er buiten Corex gewerkt worden in een teksteditor of in Excel.

Hierboven zijn drie zoekfuncties van Corex kort behandeld: *Metadata Search, Statistics* en *Lexicon Tool.* Twee andere belangrijke zoekfuncties die nog niet eerder genoemd zijn – *Content Search* en *Syntax Search* – worden in 4.1 en 4.2 uitgebreider toegelicht.

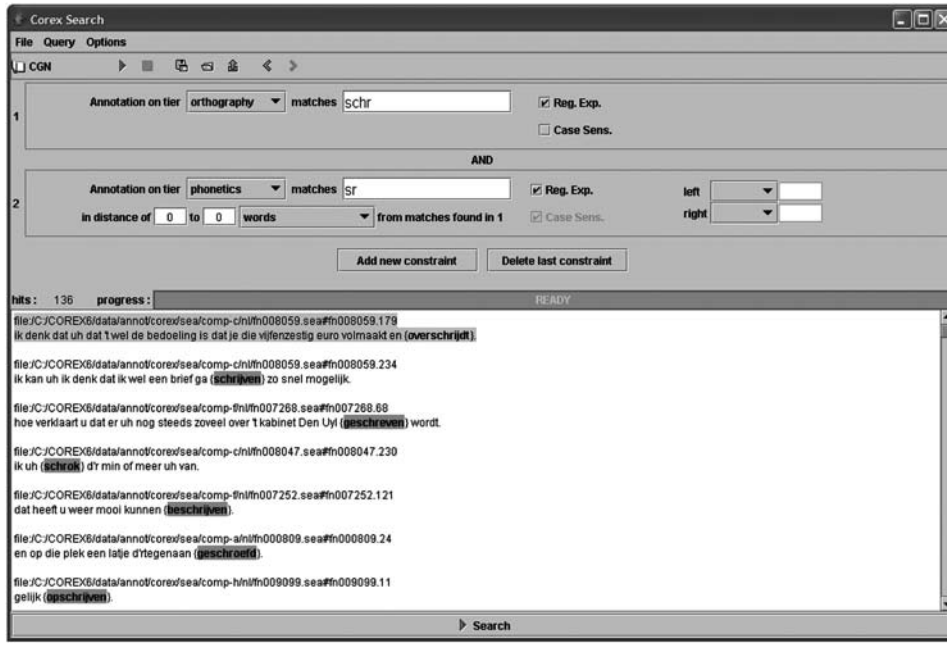
4.1 Content Search

Eén van de meest gebruikte en veelomvattende zoekfuncties binnen Corex, is de *Content Search.* Met uitzondering van de syntactische annotaties, kan met deze functie in elke annotatielaag gezocht worden op specifieke woordvormen, lemma's, POS-tags, prosodische of fonetische fenomenen of apart gemarkeerde spraak zoals dialectwoorden of versprekingen. Daarnaast kunnen meerdere parameters tegelijkertijd gedefinieerd worden: in de orthografie kan het hele corpus doorzocht worden op een enkel woord of op woordcombinaties. Neem bijvoorbeeld de woordcombinatie *een … geleden.* De afstand tussen de woorden *een* en *geleden* kan variabel zijn afhankelijk van de zin waarin de combinatie voorkomt (*een jaar geleden, een hele tijd geleden, een maand of zes geleden*). Corex biedt in dit geval de mogelijkheid om het minimale en/of maximale aantal woorden tussen *een* en *geleden* in te voeren in het zoekscherm. Hoe groter de ingestelde afstand tussen de twee woorden wordt, hoe meer ruis er ontstaat. In dat geval moet er handmatig meer filtering plaatsvinden. In figuur 7 is gekozen voor een afstand van twee tot vijf woorden tussen de twee doelwoorden.



Figuur 7: Het zoekscherm van de Content Search waarin gezocht wordt op de woordcombinatie ' een ... geleden'.

In de *Content Search* zijn meerdere annotatielagen met elkaar te combineren. In de orthografische annotatielaag kunnen we bijvoorbeeld zoeken op woorden beginnend met de cluster <schr>. In de fonetische transcriptielaag kan vervolgens parallel gezocht worden op de uitspraakvariant /sr/. De zoekresultaten zoals weergegeven in figuur 8, geven nu alle 'schr'-clusters weer die uitgesproken worden als /sr/. Let hierbij wel op dat alleen de handmatig gecontroleerde fonetische transcripties toegankelijk zijn via Corex; slechts tien procent van het corpus wordt doorzocht.



Figuur 8: Het zoekscherm van de Content Search waarin gezocht wordt op uitspraakvariant /sr/ van de cluster <schr> (de optie Reg. Exp. zorgt ervoor dat er woordintern gezocht wordt).

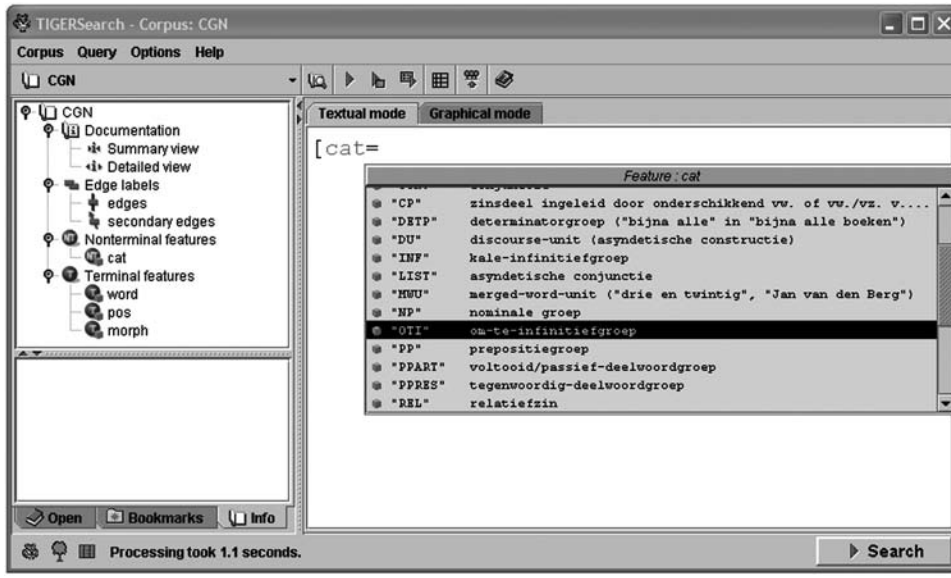
Doordat de annotaties door middel van een signaalkoppeling verbonden zijn aan de spraakfragmenten, zijn de zoekresultaten van de *Content Search* direct te bekijken in combinatie met de geluidsbestanden. In de zogenaamde *Corex-viewer* (zie figuur 9) kan het spraaksignaal synchroon worden afgespeeld met de gewenste annotaties. Deze *viewer* kan direct worden geopend vanuit het zoekvenster met de resultaten.

Figuur 9: De corex-viewer, gekoppeld aan het eerste resultaat in figuur 8.

4.2 Syntax Search

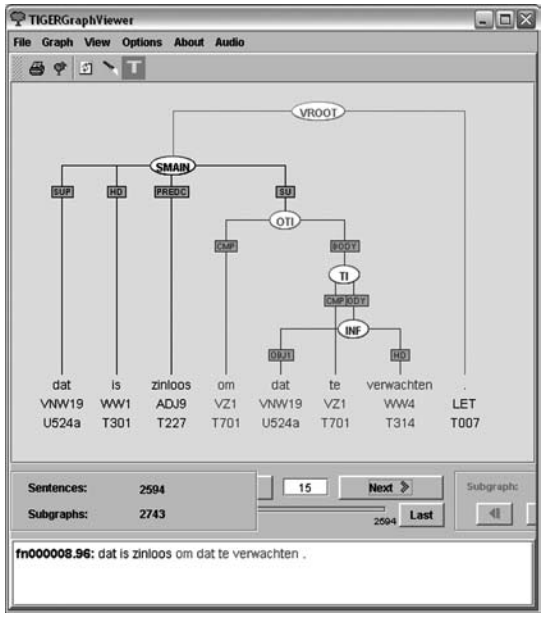
Om de syntactische annotaties te bekijken en te analyseren, is er een speciale applicatie in Corex ingebouwd, genaamd *TigerSearch*. De applicatie wordt via het hoofdscherm geopend door in het menu *Syntax Search* te kiezen. Binnen *Tiger* is alleen dát deel van het corpus toegankelijk, waarvan een syntactische transcriptie beschikbaar is. In *Tiger* kan gezocht worden op woorden, woordsoort en morfologische informatie. Complexere zoekopdrachten in verband met adjacentie, dominantie of grammaticale functie behoren ook tot de mogelijkheden. In 4.3 worden deze begrippen verder verduidelijkt aan de hand van een concrete onderzoeksvraag,

Er zijn twee manieren om in *TigerSearch* te zoeken. De tekstuele modus wordt ten opzichte van de grafische modus het meeste gebruikt. In de tekstuele modus kunnen syntactische zoekvragen gedefinieerd worden met behulp van de *TigerSearch query language* (zie ook paragraaf 4.3). Met behulp van deze taal kan er gezocht worden op dependentiestructuren en syntactische labels. In *TigerSearch* is een beschrijving van alle syntactische labels geïntegreerd en de mogelijke opties verschijnen automatisch wanneer het begin van een zoekvraag in het scherm ingetikt wordt. De syntactische constructie *om te + infinitief* is bijvoorbeeld standaard opgenomen in *Tiger* als categorie 'OTI'.



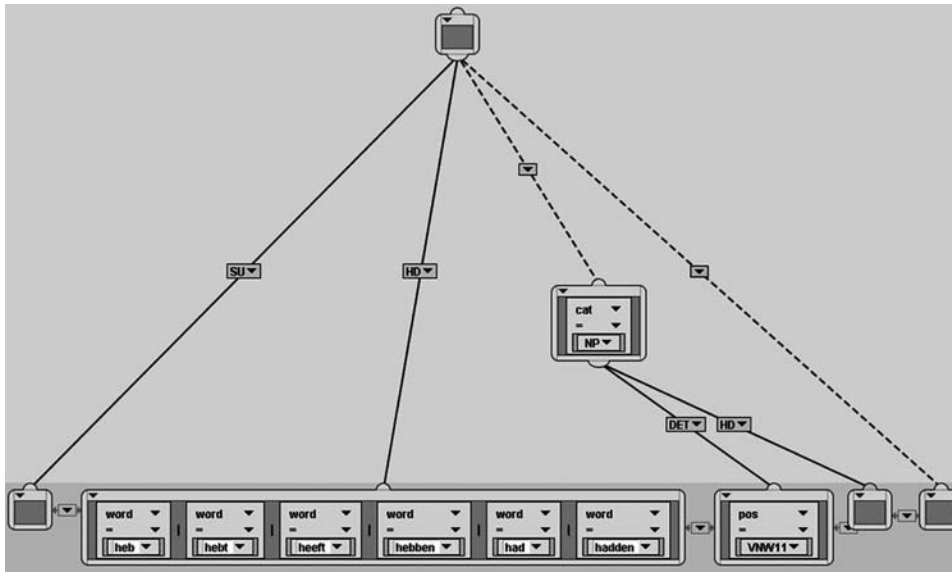
Figuur 10: Zoekvraag in de tekstuele modus van Tiger; om te+infinitiefgroep.

Na uitvoering van de zoekvraag [cat="OTI"] (figuur 10) verschijnt de *Tigergraph-viewer* waarin de boomstructuren te zien zijn. In figuur 11 wordt één van de resultaten getoond.



Figuur 11: De Tigergraph-viewer; om te+ infinitiefgroep.

In de grafische modus kunnen de afhankelijkheidsrelaties en hiërarchie van een syntactische zoekvraag visueel in kaart gebracht worden. In figuur 12 is de grafische modus afgebeeld waarin gezocht wordt op constructies vergelijkbaar met *ik heb mijn band lek.* De afbeelding uit de grafische modus kan direct omgezet worden naar de tekstuele modus. Andersom is dit niet mogelijk.



Figuur 12: Zoekvraag in de grafische modus van Tiger; SU - hebben - VNMW11 (bezittelijk voornaamwoord) - HD.

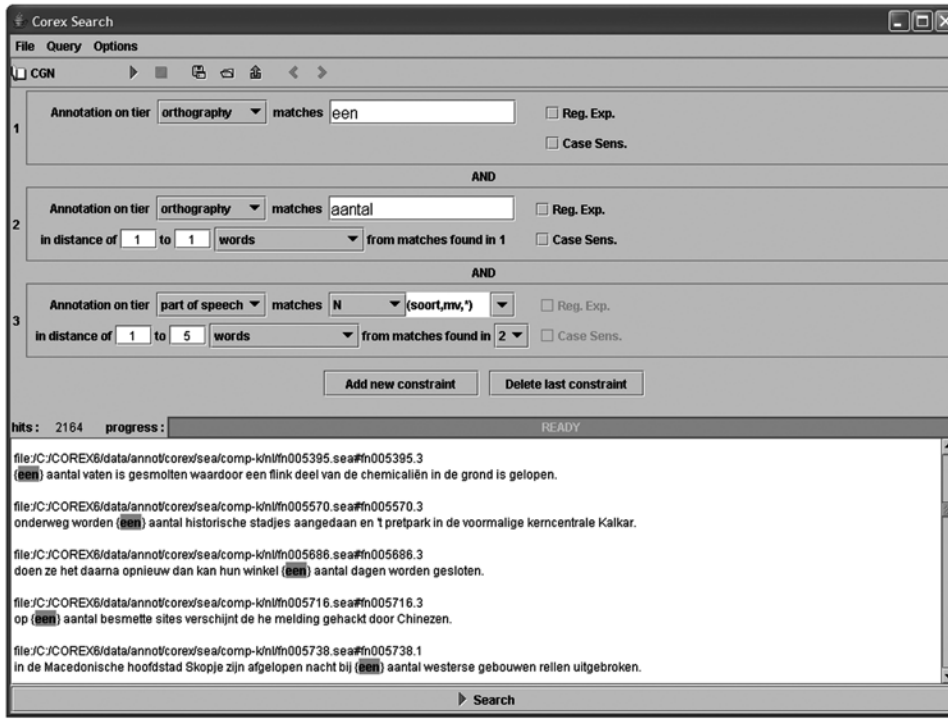
4.3 Uitwerking van een onderzoeksvraag

Om te illustreren hoe het verwerkingsproces van een onderzoeksvraag in Corex in zijn werk gaat, wordt in deze paragraaf de volgende casus behandeld: wanneer in een zin een subject als *een aantal vragen* voorkomt, zoals in (2), wordt in gesproken taal dan de norm gevolgd dat het finiete werkwoord congrueert met het enkelvoudige nomen *aantal,* of ontstaat er meervoudige congruentie met *vragen*? Is daarnaast de woordvolgorde nog van invloed op de congruentie: finiete werkwoord - onderwerp in (2a) tegenover onderwerp - finiete werkwoord in (2b,c)?

(2) a Er moet/moeten nog een aantal vragen beantwoord worden.
b Een aantal vragen moet/moeten nog beantwoord worden.
c ...dat een aantal vragen nog beantwoord moet/moeten worden.

Voor het zo goed mogelijk beantwoorden van deze vraag beginnen we – om een beeld te vormen van de relevante data – met een algemene zoekactie in de *Content Search.* In de orthografische transcriptielaag zoeken we het woord *een* direct gevolgd door *aantal,* dus

met een afstand van precies één woord. Deze woordcombinatie moet gevolgd worden door een meervoudig zelfstandig naamwoord. In de *Content Search* wordt dit aangegeven door een specifieke POS-tag op te vragen. In figuur 13 is gekozen voor de tag van een meervoudige soortnaam. De zoekactie levert meer dan tweeduizend resultaten op.



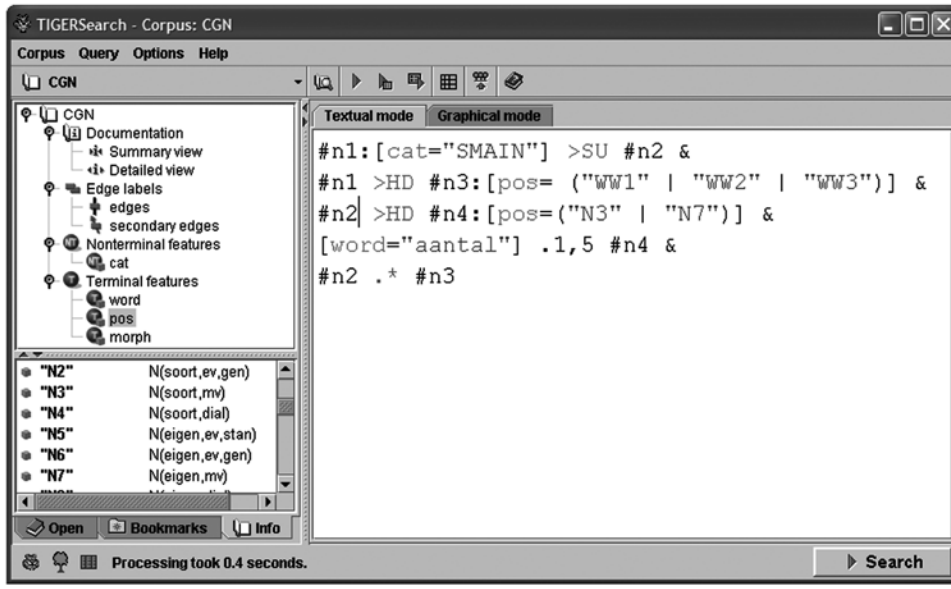
Figuur 13: Content Search, resultaat van de zoekactie 'een aantal + N (soort, mv)'.

In de voorbeeldzin (2) volgt het nomen direct op de woordcombinatie *een aantal*. Er kan natuurlijk ook een bijvoeglijk naamwoord voor het zelfstandig naamwoord voorkomen (*een aantal moeilijke vragen*) of zelfs meerdere (*een aantal heel erg moeilijke vragen)*. Om dat te ondervangen wordt de woordafstand ingesteld op vijf verwijderd van *aantal* (zie ook figuur 13). Hoe groter de afstand tussen de doelwoorden hoe groter de ruis; niet alle weergegeven resultaten zullen voldoen aan de gewenste zinsconstructie. De resultaten zullen nu handmatig gefilterd moeten worden op de relevante zinnen.

Een constructie met een eigennaam (*een aantal Vlamingen)* valt ook onder de te onderzoeken categorie. Aangezien er in de *Content Search* maar op één POS-tag tegelijkertijd gezocht kan worden, moet hiervoor een aparte zoekactie worden verricht. De afzonderlijke zoekactie op eigennamen (*eigen, mv*) levert vijftien extra resultaten op.

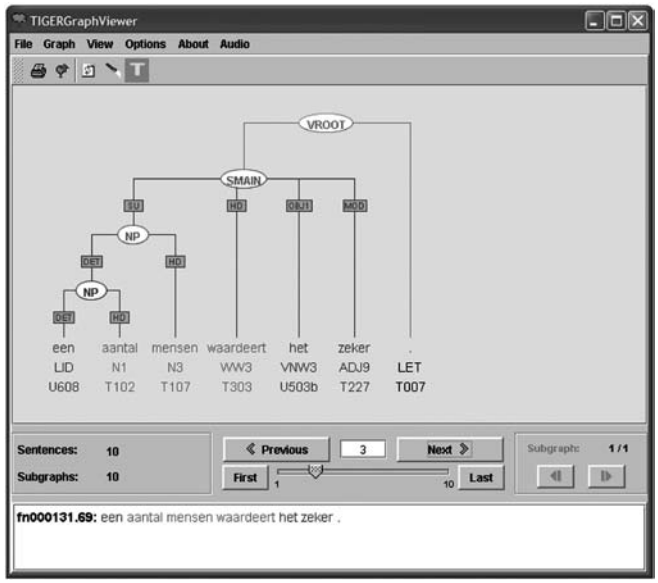
Met *TigerSearch* kunnen specifiekere zoekacties geformuleerd worden, maar die worden alleen toegepast op de fragmenten die over een syntactische annotatie beschikken. Er kan bijvoorbeeld bepaald worden dat de constructie *een aantal + N* onderwerp moet zijn van een zin. Het voordeel hiervan ten opzichte van de *Content Search* is dat er minder

ongewenste resultaten teruggeven zullen worden. Een nadeel is dat *Tiger* alleen werkt met de fragmenten die syntactisch geannoteerd zijn (tien procent van alle data) waardoor er in totaal minder resultaten gevonden zullen worden.



Figuur 14: Syntactische zoekvraag in ' TigerSearch query language'.

In de tekstuele modus afgebeeld in figuur 14 is door middel van de *query language* een syntactische zoekvraag gedefinieerd. Linksonder in het deelscherm is een fragment te zien van de afgekorte tags die hiervoor gebruikt moeten worden. De eerste regel van de *query* geeft aan dat we op zoek zijn naar een hoofdzin ('SMAIN') met een subject ('SU') ('#n1' en '#n2' zijn variabelen). De tweede regel geeft aan dat het hoofd ('HD') van de hoofdzin een persoonsvorm moet zijn (enkelvoud, meervoud of de vorm met -t). Daarnaast moet het hoofd van het onderwerp een meervoudig zelfstandig naamwoord zijn (soortnaam N3 of eigennaam N7). Dit vereiste is gedefinieerd in de derde regel van de zoekvraag. Ten slotte wordt in de vierde regel aangegeven dat het woord *aantal* onderdeel moet zijn van de zin, gevolgd door het eerder gedefinieerde zelfstandig naamwoord ('#n4') met een maximale afstand van vijf woorden. De afsluitende regel bepaalt dat het subject voorafgaat aan de persoonsvorm. Deze zoekactie levert tien zinnen op, waarvan één in de *Tigergraph-viewer* weergegeven is in figuur 15.



Figuur 15: Een resultaat van de syntactische zoekactie uit figuur 14.

Voor de andere twee zinsconstructies (2a, c) moet een vergelijkbare nieuwe *query* gemaakt worden. In totaal worden er dan tweeënvijftig relevante zinnen gevonden. Er is een tendens te ontdekken in de voorkeur voor enkelvoudige congruentie met *aantal* in hoofdzinnen met de volgorde onderwerp - werkwoord (70%, *een aantal mensen waardeert het zeker)* en juist een voorkeur voor meervoudige congruentie met het zelfstandig naamwoord in hoofdzinvolgorde werkwoord - onderwerp (76%, *worden een aantal historische stadjes aangedaan)* en in de bijzin (76%, *dat een aantal dieren nog zoiets hebben van).* Opmerkelijk is dat de twee zinnen met de woordvolgorde onderwerp - finiete werkwoord een andere voorkeur vertonen voor de congruentie.

Zoals uit het voorbeeld blijkt is een syntactische zoekactie in *TigerSearch* vrij bewerkelijk. Een grondige verdieping in de mogelijke functies en tags is vereist, voordat een goede zoekvraag gedefinieerd kan worden die tot de gewenste resultaten leidt. Bestudering van de uitgebreide *Tiger*-handleiding strekt tot aanbeveling.

Het hier geïllustreerde onderzoekje dient slechts als voorbeeld voor dit artikel. Op basis van de genoemde resultaten zijn op een tendens na geen betrouwbare conclusies te trekken. Meer factoren spelen een rol. Denk bijvoorbeeld aan verschillen tussen Vlamingen en Nederlanders of vervanging van *aantal* met vergelijkbare woorden met een collectieve aanduiding zoals *heleboel.* Het uitvoeren van meerdere verschillende soorten zoekvragen geeft een betrouwbaarder resultaat.

5 Onderzoeken met het CGN

Het Corpus Gesproken Nederlands wordt vandaag de dag ingezet bij uiteenlopende onderzoeken. De geluidsbestanden worden door verschillende taal- en spraaktechnologiebedrijven gebruikt voor het trainen van spraakherkenners. In het wetenschappelijke veld wordt onderzoek gedaan naar de verbetering van huidige spraakherkenners. In Van Dalen, Wiggers & Rothkrantz (2006) wordt beschreven hoe de automatische herkenning van lexicale klemtoon in continue spraak onderzocht wordt. Een bijzonder aspect aan het onderzoek is dat niet alleen van klinkers, maar ook van consonanten aangenomen wordt dat deze klanken een indicatie van klemtoon met zich meedragen. In het experiment krijgen consonanten daarom in sommige gevallen ook lexicale klemtoon toegewezen. Een dergelijk experiment kan tot aanzienlijke verbeteringen van de Nederlandse spraakherkenning leiden.

Ook binnen de taaltheoretische gebieden van de wetenschap wordt het corpus frequent gebruikt. Van Son, Wesseling & Pols (2006) voerden een experiment uit met betrekking tot beurtwisselingen in spontane gesprekken. De hypothese werd getest, dat de aanwezigheid van geaccentueerde woorden vlak voor een beurtwisseling tot snellere reacties leidt van de gesprekspartner. Uit het CGN zijn hiervoor de spontane dialogen gebruikt. De sprekers werden van elkaar gescheiden met behulp van de woordsegmentatie. In de experimentopstelling kon nu een dialoogsituatie gecreëerd worden, door de combinatie van een opname van één spreker uit het corpus en de proefpersoon. Ook de prosodische annotatie uit het CGN werd gebruikt om de accenten te bepalen.

Naast deze praktische experimenten zijn er ook voorbeelden van meer theoretisch gerichte onderzoeken, zoals dat van Kloots, Gillis & Swerts (2006). Hierin wordt crosslinguïstisch onderzoek gedaan naar onbeklemtoonde vocalen aan het syllabe-einde in het Nederlands van Vlamingen en Nederlanders. In de Nederlandse fonologie wordt in tegenstelling tot de Vlaamse onderscheid gemaakt tussen zogenaamde ‘lange’ en ‘korte’ vocalen. De intuïtie bestaat dat een open syllabe in het Nederlands steeds eindigt op een fonologisch ‘lange’ vocaal. Een perceptie-experiment werd uitgevoerd aan de hand van spontane gesprekken uit het corpus, waaronder interviews met leraren Nederlands.

Bovenstaande beschrijvingen zijn slechts een greep uit de verschillende soorten onderzoek die uitgevoerd zijn met het CGN. Dit nummer van *Nederlandse Taalkunde* bevat verdere illustraties van gebruiksmogelijkheden van het CGN: op het terrein van de semantiek (zie de bijdrage van Oosterhof en Coussé), in onderzoek naar gebruiksverschillen tussen Belgisch Nederlands en Nederlands Nederlands (zie het artikel van Boogaart e.a.), en bij het vaststellen van ‘Extended Lexical Units’ (Van der Wouden).

6 Het CGN bij de TST-centrale

Na afronding van het project werd het CGN aanvankelijk gedistribueerd door ELDA (Evaluations and Language resources Distribution Agency) maar sinds 2004 is het corpus beschikbaar via de centrale voor Taal- en Spraaktechnologie (TST-centrale). De TST-centrale is op initiatief van de Nederlandse Taalunie opgericht omdat er behoefte was aan een kennis- en distributiecentrum voor digitale taalmaterialen zoals het CGN. De TST-centrale is op

dit moment als afdeling ondergebracht bij het Instituut voor Nederlandse Lexicologie met een vestiging in Leiden en Antwerpen. De taalmaterialen worden bij de centrale niet alleen gedistribueerd maar ook onderhouden. Wat het CGN betreft, houdt dat in dat er onder andere talloze fouten in de annotaties, de software en de documentatie verbeterd werden. Deze verbeteringen hebben in 2006 geleid tot de lancering van versie 2.0 van het corpus.

Naast distributie en onderhoud, stimuleert de TST-centrale het gebruik van het CGN en fungeert de centrale tevens als helpdesk. Zo worden er aan universiteiten regelmatig workshops en gastcolleges verzorgd, over de gebruiksmogelijkheden van het CGN en Corex. Momenteel wordt er ook gewerkt aan een online CGN-webcursus, een nieuwe en verbeterde versie van Corex (door Polderland, Nijmegen) en uiteindelijk zal er een online webversie van het CGN beschikbaar komen. Dit alles wordt naar verwachting binnen een jaar gerealiseerd.

De webversie van het CGN is op dit moment in ontwikkeling en er wordt naar gestreefd om zo veel mogelijk functionaliteit van Corex in deze versie te herbergen. Tot in hoeverre álle functies en achterliggende data gratis online beschikbaar komen, moet nog onderzocht worden. Het complete CGN wordt op dit moment voor onderzoek en commerciële doeleinden aangeboden op drieëndertig dvd's: een annotatie-dvd en tweeëndertig geluids-dvd's. De annotatie-dvd bevat onder andere alle transcripties, de metadata, lexica en de frequentiegegevens. Ook de exploitatiesoftware Corex en alle documentatie en handleidingen zijn aanwezig op de annotatie-dvd. Op de tweeëndertig geluids-dvd's staan alle 12.780 geluidsbestanden. De annotatie-dvd is ook te gebruiken zonder de dvd's met spraakbestanden.

De Nederlandse Taalunie heeft een commissie aangesteld die zich bezig gaat houden met het vaststellen van marktconforme prijzen voor taalmaterialen. Voor individuele onderzoekers is de aanschaf van het gehele CGN nu nog een kostbare aangelegenheid. De 'prijzencommissie' zal onderzoeken of zo veel mogelijk materialen voor onderzoeksdoeleinden voor een minimale prijs beschikbaar gesteld kunnen worden. Voor commercieel gebruik zullen de (marktconforme) prijzen echter wel gehandhaafd worden.

En wat staat er verder nog te gebeuren? Het Corpus Gesproken Nederlands is een afgerond project en zal in zijn huidige staat waarschijnlijk niet verder worden aangevuld. Er zijn wel projecten die in lijn van het CGN ontwikkeld worden, zoals bijvoorbeeld het JASMIN-CGN (Cucchiarini en anderen 2006). De afkorting JASMIN staat voor Jongeren, Anderstaligen, Senioren, mens-Machine-Interactie voor het Nederlands. Dit corpus wordt onder leiding van de Radboud Universiteit Nijmegen ontwikkeld in het kader van STEVIN (Spraak- een Taaltechnologische Essentiële Voorzieningen In het Nederlands), een onderzoek- en stimuleringsprogramma voor de Nederlandse en Vlaamse taal- en spraaktechnologie. Het doel van het project is om het CGN uit te breiden, door een nieuw corpus samen te stellen van hedendaags Nederlands zoals gesproken door kinderen van verschillende leeftijdsgroepen, niet-moedertaalsprekers en ouderen. Een bijzondere taalgebruiksituatie die onderdeel vormt van het project JASMIN-CGN, is communicatie tussen mens en computer, oftewel mens-machine-interactie. Een dialoog met een spraakcomputer leidt tot typische fenomenen zoals hyperarticulatie, syllabeverlenging, stemverheffing of klemtoonverschuiving. Deze kunstmatige dialogen vormen samen met de andere kenmerkende spraak uit het JASMIN-CGN een interessante bron voor toekomstig onderzoek naar gesproken Nederlands. Het Corpus Gesproken Nederlands is nog maar het begin.

Bibliografie

Baayen, R.H., R. Piepenbrock & Rijn, H. van (1993). *The CELEX Lexical Database*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Boersma, P.& D. Weenink, (1992-2007). Praat: doing phonetics by computer. <http://www.praat.org/>.

Cucchiarini, **C., H. van Hamme**, **O. van Herwijnen**, **& F. Smits (2006).** JASMIN-CGN: Extension of the Spoken Dutch Corpus with Speech of Elderly People, Children and Non-natives in the Human-Machine Interaction Modality. In: ***Proceedings of the 5th International Conference on Language Resources and Evaluation***. Genoa, Italy, 135-138.

Dalen, R.C. van, P. Wiggers & L.J.M. Rothkrantz (2006). Lexical Stress in Continuous Speech Recognition. In: *Proceedings Interspeech 2006*, 2382–2385.

Kloots, H., S. Gillis & M. Swerts (2006). Onbeklemtoonde vocalen aan het syllabeeinde in het Standaardnederlands van Vlamingen en Nederlanders. In: T. Koole e.a. (red.), *Artikelen van de Vijfde Sociolinguïstische Conferentie*. Delft: Eburon, 296-307.

Haeseryn, W., K. Romijn, G. Geerts, J. de Rooij, & M.C. van den Toorn (1997). *Algemene Nederlandse Spraakkunst*. Groningen: Martinus Nijhoff.

Oostdijk, N. (2000). The Spoken Dutch Corpus. Overview and first Evaluation. In: *Proceedings LREC 2000,* Genoa, Italy.

Oostdijk, N. (2004). De website van het Corpus Gesproken Nederlands. <http://www.tst.inl.nl/cgndocs/doc_Dutch/start.htm>

***Referentiebestand Nederlands (RBN)* (1998).** Samengesteld onder auspiciën van de Commissie Lexicografische Vertaalvoorzieningen (CLVV). Beschikbaar bij het Instituut voor Nederlandse Lexicologie.

Son, R.J.J.H. van, W. Wesseling & L.C.W. Pols (2006). Prominent Words as Anchors for TRP Projection. In: *Proceedings Interspeech 2006,* paper 1235.

***Woordenlijst Nederlandse taal* (1995).** Samengesteld door het Instituut voor Nederlandse Lexicologie in opdracht van de Nederlandse Taalunie. Met een Leidraad van Jan Renkema. Den Haag/Antwerpen: Sdu Uitgever/Standaard Uitgeverij.

Corpusgebaseerd semantisch onderzoek: troeven en beperkingen van het CGN

ALBERT OOSTERHOF EN EVIE COUSSÉ*

Abstract

This paper discusses the advantages and drawbacks of the *Spoken Dutch Corpus* (CGN) for semantic research. Most of the discussion is based on phenomena associated with polarity, genericity and partitivity. We present corpus studies into the distribution of two polarity sensitive items (*doorgaan* 'be held' and *zot zijn van* 'be crazy about') and into the frequencies of different noun phrase types in sentences with the kind predicate *uitsterven* ('die out'). An important advantage of the CGN is that a number of components of the corpus contain spontaneous spoken material, which makes it possible to investigate semantic properties of phenomena that appear in informal language use only. Furthermore, CGN data show a (relatively) low level of awareness of the standard. It will be shown that as a consequence of this the CGN is a more appropriate tool for investigating (geographical) variation playing a role in semantic phenomena than a number of other corpora. On the other hand, a disadvantage of the CGN is that the corpus is relatively small (9 million words) and in many cases simply too small to draw statistically reliable conclusions.

1 Inleiding

In deze bijdrage worden aan de hand van enkele casestudies enkele troeven en beperkingen van het *Corpus Gesproken Nederlands* (CGN) voor semantisch onderzoek besproken.[1] In paragraaf 2 gaan we in op de moeilijke relatie tussen corpusonderzoek en semantisch onderzoek. In veel theoretisch semantisch onderzoek staat (vooral) introspectie centraal bij de semantische analyse van zinnen en constructies. We zullen echter betogen dat de formele semantiek ook kan profiteren van bevindingen die voortkomen uit corpusonderzoek. De geschetste uiteenzetting zal de achtergrond vormen voor onze evaluatie van het CGN als bron voor semantisch onderzoek. Paragraaf 3 belicht enkele voordelen van het CGN voor semantisch corpusonderzoek. We zullen argumenteren dat het CGN door zijn gevarieerde samenstelling (gevarieerder dan andere corpora met geschreven Nederlands)

* Beide auteurs zijn verbonden aan de Vakgroep Nederlandse Taalkunde van de Universiteit Gent, Blandijnberg 2, 9000 Gent, België. E-mail: albert.oosterhof@ugent.be en evie.cousse@ugent.be. Albert Oosterhof is postdoctoraal onderzoeker aan de UGent en deeltijds docent aan de Universiteit Antwerpen. Evie Coussé is aspirant-onderzoeker bij het Fonds voor Wetenschappelijk Onderzoek – Vlaanderen (FWO). We danken de reviewers voor hun commentaar bij een eerdere versie van dit artikel.

1 Alle casestudies zijn uitgevoerd op versie 1.0 van het CGN via de meegeleverde exploitatiesoftware COREX. Het programma werd vanaf de annotatie-dvd gedraaid onder het besturingssysteem Windows XP Professional.

bij uitstek geschikt is voor semantisch onderzoek. Interessant is dat het corpus spontane en informele teksten bevat waarbij sprake is geweest van een minimum aan normdwang. Verder vertoont het taalgebruik dat opgeslagen is in het CGN veel regionale en stilistische variatie. Die troeven zullen geïllustreerd worden aan de hand van twee casestudies. Het gaat om studies naar de polaire gevoeligheid van *doorgaan* ('gehouden worden'), zoals in *De bijeenkomst gaat waarschijnlijk niet/toch door*, en van *zot zijn van* ('gek zijn op'). Daarnaast zullen we ook kijken naar enkele beperkingen van het CGN voor semantisch corpusonderzoek (paragraaf 4). In vergelijking met bestaande geschreven corpora is het CGN relatief beperkt in omvang (ca. 9 miljoen woorden), waardoor minder frequente constructies nauwelijks aangetroffen worden in het corpus. In paragraaf 4 zullen we deze beperking illustreren aan de hand van casestudies over de *wemel*-constructie (zie Hoeksema 2005) en over de frequentie van verschillende typen naamwoordgroepen als subject van het werkwoord *uitsterven*. Paragraaf 5 bevat de conclusies van deze bijdrage.

2 Op intuïties gebaseerd onderzoek versus corpusonderzoek

2.1 Voor- en nadelen van intuïties voor theoretisch taalonderzoek

In deze sectie bespreken we de voordelen en beperkingen van op intuïties gebaseerd onderzoek voor theoretisch-taalkundig onderzoek. In 2.2 gaan we in op de voordelen en bezwaren van corpusonderzoek. De geschetste discussie zal de achtergrond vormen van onze evaluatie van het CGN.

In de formele semantiek wordt grote waarde gehecht aan de intuïtie van een moedertaalspreker. Door middel van introspectie kan een linguïst oordelen over de interpretatie en acceptabiliteit van taaluitingen. Zo kan inzicht verworven worden in de structuur van de grammatica van een taal. Deze manier van werken is niet uniek voor formele semantici. Syntactici die werken in de generatieve traditie volgen doorgaans dezelfde methodologische procedures. Borsley & Ingham (2002:1) verwijzen naar de 'frameworks' die deze benadering kiezen als "mainstream theoretical linguistics":

> "Mainstream work is mainly concerned with I-language, the cognitive system underlying the ordinary use of language. Various applied linguists have suggested that this is unsatisfactory and that texts are a more appropriate object of study. (...) Kress (1993) is under the impression that mainstream linguists believe that all members of a speech community have the same I-language and use it in the same way, but of course no one believes this."

De nadruk op *I-language*, de mentaal gerepresenteerde linguïstische kennis die een moedertaalspreker van een taal heeft (vgl. Chomsky 1988), ligt aan de basis van veel theoretische syntactische en semantische studies. Soms wordt dit perspectief expliciet gemaakt, zoals door Haegeman (2004: 120):

> "Uiteraard is niet uitgesloten dat andere sprekers (…) andere oordelen hebben, maar dit is op zich te verwachten vanuit de generatieve visie op de grammatica met haar nadruk op de I-language, de interne taal, d.w.z. de competentie van de individuele taalgebruiker."

Acceptabiliteitsoordelen impliceren dus niet noodzakelijk dat ze gedeeld worden door (een meerderheid van de) andere sprekers van een taal. Als we echter een beeld willen krijgen van het geheel van acceptabiliteitsoordelen dat correspondeert met de I-language van een spreker van een bepaalde taal (en de interne consistentie in dat systeem), is de op intuïties gebaseerde methode geschikter dan corpusonderzoek.

De subjectiviteit van intuïties (i.e. het feit dat intuïties niet noodzakelijk gedeeld worden door andere sprekers) houdt natuurlijk tegelijkertijd een belangrijke beperking in. We zullen dit illustreren aan de hand van een voorbeeld. De zinnen en de bijbehorende grammaticaliteitsoordelen in (1) zijn afkomstig van Zwart (1997: 28). Zwart stelt dat topicalisatie een trigger is van inversie van het finiete werkwoord (*kussen*) en het subject (*de jongens*), zoals geïllustreerd in (1b). Hij schrijft over deze zin: "The finite verb and the subject no longer have to be adjacent." Dit blijkt uit de positie die *vandaag* in kan nemen in (1b). In zin (1a), waarin er geen sprake is van topicalisatie, moeten het finiete werkwoord en het subject volgens Zwart wél adjacent zijn.

(1) a Marie (*vandaag) kussen de jongens vaak.
b Marie kussen (vandaag) de jongens vaak.

Voor ons betoog is vooral zin (1b) van belang. Volgens Zwart is de zin *Marie kussen vandaag de jongens vaak* blijkbaar welgevormd. Toch vinden veel sprekers van het Nederlands deze zin niet acceptabel.[2] In een kleinschalig onderzoekje (zie Oosterhof 2006a) werd negen moedertaalsprekers van het Nederlands gevraagd de acceptabiliteit van deze zin te beoordelen. Acht van die negen informanten bleken de zin onacceptabel te vinden. Wanneer een linguïst aanneemt dat een zin als (1b) welgevormd is, wordt dus niet (noodzakelijk) de syntaxis van het Nederlands beschreven, maar veeleer de syntaxis van een bepaald idiolect (maar zie ook noot 2). Het is van belang om voldoende rekening te houden met het subjectieve karakter van op introspectie gebaseerd onderzoek.

De eigen intuïties over de acceptabiliteit en interpretatie van taaluitingen kunnen natuurlijk wel vergeleken worden met de intuïties van andere moedertaaltaalgebruikers, bijvoorbeeld door middel van enquêtes. Zo kan achterhaald worden of er variatie is in de oordelen, hetgeen hierboven geïllustreerd werd naar aanleiding van de voorbeeldzinnen in (1).

2 Hier moet opgemerkt worden dat er een verschil is tussen 'acceptabiliteit' en 'grammaticaliteit'/'welgevormdheid'. De bewering dat een zin acceptabel is, heeft betrekking op de intuïties van moedertaalsprekers over linguistische data. 'Grammaticaliteit' en 'welgevormdheid' zijn daarentegen theoretische noties (vgl. Chomsky 1965). Een zin is grammaticaal/welgevormd als hij gevormd is volgens de regels van de grammatica van de taal zoals die geformuleerd worden door de linguïst. Haegeman (1994: 8) schrijft dat "[t]he native speaker who judges a sentence cannot decide whether it is grammatical. He only has **intuitions** about **acceptability**. It is for the linguist to determine whether the unacceptability of a sentence is due to grammatical principles or whether it may be due to other factors". Grammaticale zinnen kunnen bijvoorbeeld onacceptabel gevonden worden omdat de zin moeilijk te verwerken is. Dat Zwart de zin *Marie kussen vandaag de jongens vaak* welgevormd noemt is dus niet *a priori* in strijd met het feit dat sprekers de zin onacceptabel vinden. De vraag is hier of Zwart van mening is dat de zin onacceptabel is om andere redenen, bijvoorbeeld omdat de zin moeilijk te verwerken zou zijn. Zwart maakt in zijn bespreking in het geheel niet duidelijk dat volgens hem de zin inderdaad onacceptabel is om zulke redenen. Als hij toch een dergelijk standpunt in zou nemen, zou enige explicatie in die richting op zijn plaats geweest zijn.

2.2 Voor- en nadelen van corpora voor theoretisch taalonderzoek

Naast enquêteonderzoek kan ook corpusonderzoek een interessante aanvulling vormen op de eigen intuïties als bron van informatie voor semantisch onderzoek. Via corpusonderzoek krijgen we inzicht in de distributie en het gedrag van taalelementen in concreet taalgebruik van verschillende sprekers. In vergelijking met de resultaten van enquêteonderzoek naar (acceptabiliteits)oordelen van groepen sprekers, is het minder duidelijk wat we uit corpusfrequenties kunnen afleiden over de interpretatie of acceptabiliteit van taaluitingen. Er is immers geen één-op-één-relatie tussen het al dan niet voorkomen van een taaluiting in een corpus en de acceptabiliteit ervan (vgl. bijvoorbeeld McEnery & Wilson 2001 en Meurers 2005). De afwezigheid van een taaluiting in een corpus betekent niet dat die uiting onacceptabel is en omgekeerd zijn niet alle zinnen in een corpus per se acceptabel volgens sprekers van de taal.

Het eerste punt, namelijk dat de afwezigheid van een zinstype niet betekent dat het betreffende type onacceptabel is, kan worden geïllustreerd aan de hand van corpusgegevens over de distributie van een negatief polaire uitdrukking (zie Hoeksema 2004, geciteerd in Oosterhof 2003-2004). De uitdrukking *het feest gaat door*, waarbij *feest* figuurlijk geïnterpreteerd moet worden, is een voorbeeld van een idioom dat gevoelig is voor polariteit. Hoeksema's materiaal (vgl. Hoeksema 2004) bevat 56 voorbeelden van deze uitdrukking. In al die zinnen komt de trigger *niet* voor. Een voorbeeldzin uit het CGN is gepresenteerd in (2).[3]

In deze zin verwijst *'t feest* naar een eerdere mededeling: *Arjan zou gisteren komen*. Hoewel het niet uitgesloten kan worden dat deze gebeurtenis letterlijk een feest is, heeft de zin in elk geval een alternatieve lezing waarbij *'t feest* figuurlijk bedoeld is.

(2) en uh ja nou heb ik gister heel veel gedaan en uh dat kwam ook eigenlijk omdat Arjan zou gisteren komen dus ik ben uh ja eerst ging van alles gaan kokkerellen en toen ben ik uh de tuin gaan uh vegen en stof gaan ruimen en zo. nou en op gegeven ogenblik dacht ik van nou hij had er toch allang moeten zijn. en ik denk nou dat uh ik zal 'ns even 't antwoordapparaat afluisteren. (…) en jawel hoor had ie zaterdag ingesproken dat uh Tijmen ziek was geworden. (…) en dat 't feest dus niet doorging. [fn008062]

Het feit dat in alle corpuszinnen van Hoeksema (2004) de trigger *niet* voorkomt, is natuurlijk een belangrijke indicatie dat *het feest gaat door* een negatief polair item is. Dit illustreert meteen de bruikbaarheid van corpusresultaten voor theoretisch semantisch onderzoek. Als een onderzoeker beweert dat een item *x* een negatief gevoelige uitdrukking is, dan is de voorspelling gerechtvaardigd dat *x* frequent[4] voorkomt in zinnen met een ontkenning of een andere trigger (vgl. bijvoorbeeld Giannakidou 1999).

3 Zin (2) is afkomstig uit een spontane telefoondialoog uit 2002 (Nederland).

4 Dit roept natuurlijk de volgende vraag op: wanneer kunnen we spreken van 'frequent'? Deze vraag kunnen we omzeilen door de frequentie van een negatief polair item x te relateren aan een item y, waarvan aangenomen wordt dat het niet negatief polair is. De voorspelling die we dan kunnen doen, is de volgende: we verwachten dat het item x frequenter voorkomt in zinnen met een ontkenning (of een andere trigger) dan het item y. In Oosterhof (2004-2005) wordt het werkwoord *doorgaan* met als betekenis 'gehouden worden' bijvoorbeeld vergeleken met het werkwoord *doorgaan* in de betekenis 'voortduren, aanhouden'. Uit de resultaten van dat corpusonderzoek blijkt dat *doorgaan* 'gehouden worden' frequenter voorkomt met een ontkenning of een andere trigger dan *doorgaan* 'voortduren, aanhouden'. Dergelijke evidentie ondersteunt de bewering dat *doorgaan* 'gehouden worden' een negatief polaire uitrukking is, in tegenstelling tot andere gebruikswijzen van *doorgaan*.

Voorbeelden van andere triggers zijn conditionele zinnen of bijzinnen bij intensionele werkwoorden als *hopen*. Volgens onder meer Giannakidou (1999) kunnen negatief polaire uitdrukkingen in principe ook in zulke contexten en in nog een aantal andere door haar gedefinieerde contexten (zie ook noot 11) gebruikt worden. Het hier beschreven resultaat maakt echter niet duidelijk of *het feest gaat door* al of niet voorkomt met andere triggers, zoals in conditionele zinnen (vgl. (3a)) en bij intensionele werkwoorden zoals *hopen* in (3b). Uit het feit dat zulke zinnen niet voorkomen in Hoeksema's corpus (en evenmin in het CGN) kunnen we immers niet concluderen dat sprekers van het Nederlands ze onacceptabel vinden. Om dergelijke conclusies te kunnen trekken, zullen we sprekers van het Nederlands moeten confronteren met dergelijke zinnen en hen moeten vragen naar hun oordelen.

(3) a Als dat feest doorgaat, slaan we een flinke slag.
b Ik hoop dat het feest doorgaat.

Het tweede punt, namelijk dat niet alle uitingen in een corpus noodzakelijk grammaticaal zijn, is geïllustreerd in (4).

(4) (…) onlangs is er een toeristische boycot afgekondigd tegen Noorwegen. (…) De reden is bekend: Noorwegen (…) weigert de internationale verdragen te ondertekenen die de walvisvaart verbieden. Zo'n boycot roept een gevoel op dat verdacht veel op nostalgie lijkt. Denk aan de jaren zestig en zeventig en het lijstje van verboden vakantielanden ontrolt zich als vanzelf in je hoofd. (…) En nu dan Noorwegen. Het is geen toeval dat het om dieren gaat. Een walvis is weerloos en bijna uitgestorven en wie voor zijn lot opkomt, raakt niet in allerlei onoplosbare morele dilemma's verstrikt; die paar werkeloze vissers kunnen wel omgeschoold worden.

In (4) is een passage gegeven uit het *INL 27 miljoen woorden corpus* (oorspronkelijk afkomstig uit het *NRC Handelsblad*, maart 1994). In deze passage komt een zin voor waarin één en dezelfde indefiniete enkelvoudige naamwoordgroep, namelijk *een walvis* wordt gecombineerd met het predicaat *weerloos* en met het predicaat *uitgestorven* (vgl. Cohen 1999: 40 voor vergelijkbare voorbeelden, zie De Vries 2005 voor een discussie over de syntaxis van dergelijke gecoördineerde structuren). *Uitgestorven* is echter een *soortpredicaat*. Dergelijke predicaten hebben als kenmerkende eigenschap dat ze alleen kunnen worden toegeschreven aan naamwoordgroepen die verwijzen naar soorten. Een algemeen geaccepteerde observatie (zie bijvoorbeeld de volgende grammatica's: Haeseryn e.a. 1997 en Broekhuis e.a. 2003) is dat zulke predicaten niet kunnen worden gecombineerd met indefiniete enkelvouden. Dat betekent dat de onderstreepte zin in (4) ongrammaticaal is.

Het feit dat *uitgestorven* niettemin gecombineerd wordt met een indefiniet enkelvoud hangt uiteraard samen met de aanwezigheid van het predicaat *weerloos*, dat geen soortpredicaat is. Dergelijke predicaten kunnen wel gecombineerd worden met indefiniete enkelvouden. Het probleem is echter dat in zinnen als (4) daarnaast een soortpredicaat wordt toegeschreven aan het indefiniete enkelvoud. Op grond van de beschrijving in de bestaande literatuur voorspellen we dan zo'n zin onwelgevormd is. Het feit dat de zin toch voorkomt in concreet taalgebruik toont aan dat op grond van de aanwezigheid van een bepaald type zin niet geconcludeerd kan worden dat de zin grammaticaal is.

Het is natuurlijk verleidelijk om zinnen die ongrammaticaal gevonden worden te verwijderen uit de verzameling van zinnen waarop de resultaten worden gebaseerd. Een dergelijke werkwijze heeft echter belangrijke nadelen. Deze praktijk roept natuurlijk meteen de vraag op hoe bepaald wordt dat een zin ongrammaticaal is. Worden daarvoor ook andere sprekers geconsulteerd? Zo ja, hoeveel en wat voor sprekers? Of worden er bijvoorbeeld woordenboeken of andere bronnen geraadpleegd? Zo ja, hoeveel en welke bronnen? Meer in het algemeen leidt de methodologische keuze om ongrammaticale zinnen te verwijderen ertoe dat de resultaten van corpusonderzoek net als op intuïtie gebaseerde bevindingen subjectief zijn. Het verschil is echter dat het risico bestaat dat corpusdata op een oncontroleerbare manier subjectief zijn, omdat het niet duidelijk is welke voorbeelden verwijderd zijn (tenzij er een lijst met geweerde zinnen wordt bijgeleverd). Terwijl op intuïties gebaseerde resultaten normaal gesproken een beeld geven van een bepaald idiolect, zijn corpusresultaten in onderzoek waarin ongrammaticale zinnen verwijderd worden dus géén weergave van een idiolect en géén weergave van gebruiksfrequenties in een corpus. Dergelijke corpusresultaten zijn een weergave van gebruiksfrequenties in een bepaald corpus, waaruit de zinnen die door de betreffende linguïst ongrammaticaal gevonden worden, zijn verwijderd. Zodoende bestaat het gevaar dat het slechtste van twee werelden gecombineerd wordt.

De aanwezigheid van zinnen die door (een percentage van de) sprekers van de taal in kwestie onacceptabel gevonden worden, is dus een valkuil bij het uitvoeren van corpusonderzoek voor theoretische doeleinden. De beste manier om hiermee om te gaan is door i) ongrammaticale zinnen gewoon te betrekken in de resultaten en ii) onder ogen te zien dat er geen één-op-één relatie is tussen corpusfrequenties en acceptabiliteit. Op die manier kan de hierboven beschreven valkuil vermeden worden.

Voor semantisch onderzoek geldt dat het in een bepaald opzicht nog moeilijker is dan voor syntactisch, morfologisch of fonologisch onderzoek om op een verantwoorde manier gebruik te maken van corpora. Het is namelijk de taak van de semanticus om uit te maken welke interpretatie een zin of constructie krijgt. Dergelijke afwegingen hebben een subjectieve component.[5] Dat zal geïllustreerd worden aan de hand van de zinnen in (5), die afkomstig zijn uit het CGN.[6]

5 Paradoxaal genoeg kan de subjectiviteit van intuïties tot grotere methodologische problemen leiden voor corpusonderzoek dan voor enquêteonderzoek (en ander op intuïties gebaseerd onderzoek). Als een onderzoeker of een andere taalgebruiker een oordeel geeft over de acceptabiliteit of interpretatie van een zin dan is dat oordeel uiteraard subjectief. Die subjectiviteit is inherent aan de betreffende intuïties. Dat leidt verder niet tot methodologische problemen. Als een taalgebruiker een bepaald oordeel heeft over de acceptabiliteit of de interpretatie van een zin, dan is de onderzoeksbevinding dat de betreffende taalgebruiker dat oordeel heeft niet subjectief. Bij corpusonderzoek ligt de situatie anders. Het feit dat oordelen over acceptabiliteit en interpretatie van zinnen subjectief zijn, betekent dat de procedure die leidt tot het resultaat van het corpusonderzoek een subjectieve component bevat. Het oordeel over de interpretatie van een zin is hier dus niet het object van studie maar maakt deel uit van de methodologische keuzen die worden gemaakt.

6 Zin (5a) is afkomstig uit een nationaal radionieuws (Nederland) uit 2001; (5b) komt uit een voorgelezen tekst (Nederland) uit 2001.

(5) a het oormerken van dieren die preventief zijn ingeënt tegen mond- en klauwzeer mag doorgaan. dat heeft de rechtbank in Den Haag bepaalt in het kort geding dat was aangespannen door de dierenbescherming. [fn001643]
b aan Godfried was gevraagd eens iets te schrijven voor het reclameblad Op De Solex. ze hadden hem twee van die fietsen gestuurd. [fn001288]

Zin (5a) bevat het werkwoord *doorgaan*. Door Haeseryn e.a. (1997) (in paragraaf 29.3) wordt aangenomen dat *doorgaan* in de betekenis 'gehouden worden, plaatsvinden' in het algemeen (en dus niet alleen in de uitdrukking *het feest gaat* door, vgl. zin (2) en (3)) een negatief polair item is. Merk op dat *doorgaan* in (5a) gebruikt wordt met een modaal hulpwerkwoord, hetgeen volgens Giannakidou (1999) een trigger is voor negatief polaire items (zie noot 11).[7]

Het werkwoord *doorgaan* ('gehouden worden') kan regionaal – met name in België – wel zonder ontkenning of een andere trigger gebruikt worden. In Oosterhof (2003-2004) worden corpusresultaten gepresenteerd waaruit blijkt dat *doorgaan* in Nederlands materiaal in ongeveer 90% van de gevallen voorkomt in negatieve omgevingen of met andere triggers, terwijl *doorgaan* in Belgisch materiaal inderdaad minder vaak in zulke omgevingen aangetroffen wordt. Dit bevestigt de beschrijving in de ANS.

Een probleem bij de uitvoering van dergelijk corpusonderzoek is dat het niet in 100% van de gevallen met zekerheid uit te maken is of *doorgaan* inderdaad 'gehouden worden' betekent. In een zin als (5a) is dit een mogelijke interpretatie, maar daarnaast kan *doorgaan* in (5a) ook een andere betekenis krijgen, namelijk 'voortduren, aanhouden' (vgl. Oosterhof 2003-2004). De bewering in de ANS heeft geen betrekking op die alternatieve betekenis. In veel gevallen zal uit de context opgemaakt kunnen worden welke betekenis *doorgaan* krijgt. Toch is het onvermijdelijk dat er twijfelgevallen overblijven. In die gevallen is de beslissing dat we met de juiste betekenis van doen hebben tot op zekere hoogte subjectief.

Een ander voorbeeld van de subjectieve component in semantisch corpusonderzoek is gegeven in (5b). Zin (5b) bevat een partitieve constructie, namelijk *twee van die fietsen.* In de literatuur worden echter verschillende typen partitieve constructies besproken. Een zin als (5b) is ambigu: de zin kan zowel een gewone partitieve lezing krijgen als een verbleekte partitieve lezing (vgl. De Hoop e.a. 1990, Oosterhof 2005 en Le Bruyn 2007). In de eerste lezing verwijst *die fietsen* naar een verzameling van fietsen die eerder in de tekst geïntroduceerd is. Deze lezing wordt noodzakelijk indien we bijvoorbeeld een telwoord inserteren als "ingebedde determinator", zoals in *twee van die drie fietsen*. De tweede interpretatie (i.e. de verbleekte partitieve lezing) kan als volgt geparafraseerd worden: 'twee van die fietsen, weet je wel' (zie De Hoop e.a. 1990: 81). In deze interpretatie verwijst *die fietsen* eerder naar een bepaald soort van fietsen, die ook onafhankelijk van de context bekend is bij taalgebruikers. Deze interpretatie wordt noodzakelijk indien we de "inbeddende" determinator, namelijk het telwoord *twee*, weglaten zoals in *van die fietsen*.

7 Een reviewer merkt terecht op dat in (5a) wellicht ook meespeelt dat de journalist kennis bij de lezers veronderstelt, namelijk dat het er naar uitzag dat het oormerken NIET zou doorgaan. Zin (5a) kan dus worden gelezen als de nadrukkelijke bevestiging van het feit dat het oormerken WEL door mag gaan. In dat soort contrastieve interpretaties is de aanwezigheid van een trigger (vaak) überhaupt niet nodig. Dit verklaart dat er ook zinnen als (i), waarin zelfs geen modaal hulpwerkwoord voorkomt, aangetroffen worden in concreet taalgebruik (zie ook Van der Wal 1996).

Stel nu dat we de intentie hebben corpusonderzoek uit te voeren naar de syntaxis en semantiek van verbleekte partitieven. Dan worden we geconfronteerd met het probleem dat een zin als (5b) ambigu is. In Oosterhof (2005) wordt geconstateerd dat 4% van de in het corpus[8] gevonden partitieve constructies uiteindelijk ambigu is en zowel een gewone als een verbleekte partitieve lezing kan krijgen, zelfs als we rekening houden met de context waarin de zin zich bevindt. Uiteindelijk is de beslissing of de zin al of niet ambigu is en welke lezing(en) de zin krijgt/kan krijgen tot op zekere hoogte subjectief.

Beide voorbeelden illustreren dat ook in corpusgebaseerd semantisch onderzoek de subjectieve rol van de semanticus niet uitgesloten kan worden. De selectie en interpretatie van de corpusresultaten hangen immers in grote mate af van de inzichten en intuïties van de linguïst. Om toch een betrouwbaar beeld te geven van de frequentie en het gebruik van taalelementen in concreet taalgebruik is het essentieel dat de corpuslinguïst alle stappen in het evaluatieproces voldoende expliciteert. Zo verdient het de voorkeur om ongrammaticale zinnen niet zomaar te verwijderen uit de dataset. Eventueel kunnen ze in een aparte lijst voorgelegd worden aan de lezer (vgl. hierboven). Anders loopt de semanticus het risico dat de resultaten van corpusonderzoek (net als op intuïtie gebaseerde bevindingen, vgl. echter noot 5) subjectief zijn. Meer zelfs, de corpusdata worden op een oncontroleerbare manier subjectief, omdat het niet duidelijk is welke voorbeelden verwijderd zijn. Als er voldoende rekening wordt gehouden met de inbreng van de linguïst in de analyse van de corpusdata, kan semantisch corpusonderzoek op intuïtie gebaseerde analyses echter uitstekend aanvullen en nuanceren.

3 Troeven van het CGN

Nu we de bruikbaarheid van corpusdata voor semantisch onderzoek hebben besproken, zullen we specifieker ingaan op de troeven en beperkingen van het CGN voor semantisch onderzoek aan de hand van enkele casestudies. Eerst bespreken we in 3.1 drie pluspunten, die we vervolgens in 3.2 en 3.3 zullen illustreren aan de hand van twee casestudies.

3.1 Drie troeven

Een eerste pluspunt van het CGN heeft te maken met één van de argumenten die door Verkuyl (1998) worden ingebracht tegen corpusonderzoek. Verkuyl (1998:61) schrijft:

> "Er ontstaan direct tal van (…) vragen over de betrouwbaarheid van het corpus. In het INL-corpus komen grote excerpten uit NRC Handelsblad [voor] (…). We weten dat kranten er zo hun eigen schrijfdictatoren op na houden. Die van De Volkskrant gaat in zijn taalpedanterie het verst, maar ik ken ook een chef bij een van de redacties bij NRC die absoluut niet houdt van Een aantal V+en, dus als het even kan wordt een tekst waarin dat wel staat, aangepast aan die nogal particuliere regel."

8 Het corpus bestaat uit 12,1 miljoen woorden en is samengesteld uit materiaal van het *INL 27 miljoen woorden corpus* en het *38 miljoen woorden corpus*. Het gaat om tekst die afkomstig is uit kranten, tijdschriften, boeken en televisiejournaals. Zie Oosterhof (2005).

Dit soort invloed van de taalnorm (en de rol die 'schrijfdictatoren' hierin spelen) kan uiteraard invloed hebben op de resultaten van corpusonderzoek. Het CGN bevat echter een aantal onderdelen die bestaan uit min of meer spontaan taalgebruik, zoals spontane conversaties, interviews, telefoondialogen, discussies, lessen en spontane commentaren. In totaal gaat het om ca. 7 miljoen woorden. Hierdoor is het CGN geschikter voor onderzoek naar verschijnselen waarbij **normdwang** een rol speelt dan bijvoorbeeld corpora die uitsluitend bestaan uit krantenmateriaal. Dit punt wordt in 3.2 geïllustreerd aan de hand van een casestudie naar *doorgaan* ('gehouden worden').

Een tweede pluspunt heeft ermee te maken dat sommige uitdrukkingen, constructies, woorden of vormen alleen of vooral voorkomen in de **spreektaal** en/of in **informeel taalgebruik**. In Oosterhof (2006b) worden bijvoorbeeld corpusresultaten gepresenteerd waaruit blijkt dat de uitdrukking *zot zijn van* ('gek zijn op') in zinnen zoals (6) vooral voorkomt in combinatie met negatie en dus geanalyseerd kan worden als een uitdrukking die gevoelig is voor polariteit.[9]

(6) behalve spagetti me kaassaus, daar ben ik niet zot van.

Deze Vlaamse uitdrukking vinden we vooral in de spreektaal en dan vooral in informele contexten. Corpusonderzoek naar een dergelijke uitdrukking is daardoor alleen mogelijk met behulp van corpora die (ook) informele spreektaal bevatten. Het CGN is een voorbeeld van zo'n corpus. We komen hierop terug in 3.3, waar corpusresultaten op basis van het CGN gepresenteerd worden.

Een derde pluspunt is dat het CGN in tegenstelling tot veel andere corpora geschikt is om onderzoek te doen naar **variatie** in het Nederlands. Door een zorgvuldige selectie van sprekers en tekstgenres in het CGN is het mogelijk om zicht te krijgen op geografisch bepaalde variatie, zoals Noord/Zuid-verschillen, en registervariatie. Noord/Zuid-verschillen kunnen worden ondersszocht door frequenties van verschillende typen zinnen, constructies en fenomenen in Nederlands materiaal te vergelijken/contrasteren met Belgisch materiaal. In 3.2 wordt met een casestudie naar *doorgaan* geïllustreerd dat het CGN mogelijkheden biedt voor onderzoek naar Noord/Zuid-verschillen. Registervariatie kan onderzocht worden door componenten die informeler taalgebruik bevatten, zoals spontane conversaties en telefoondialogen, te vergelijken met componenten die formeler taalgebruik bevatten, zoals nieuwsbulletins en plechtige toespraken.[10]

9 Zin (6) is afkomstig van www.noxa.net/caitje (oktober 2005).

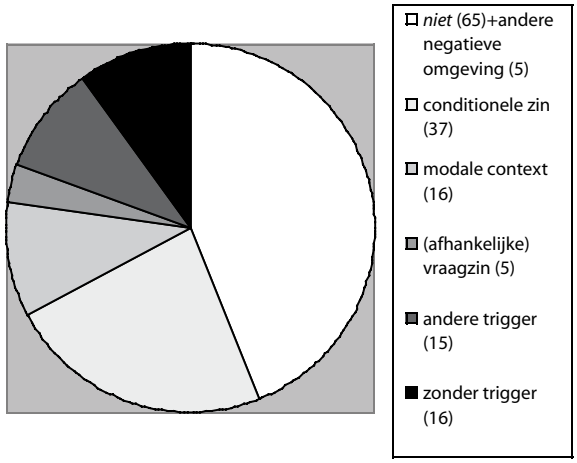
10 In (i) is een voorbeeld gegeven van een contrast dat samenhangt met registervariatie. Zinnen als (ia), waarin een definiet enkelvoud gebruikt is in een karakteriserende zin (i.e. een zin die een generalisatie uitdrukt die van toepassing is op tijgers in het algemeen), zijn gebruikelijker in formele contexten (en in schrijftaal) dan in informele contexten (en in spreektaal). In informele contexten en spreektaal vinden we vaker zinnen als (ib), waarin een indefiniet lidwoord is gebruikt. Om te onderzoeken of corpusresultaten bevestigen dat er een dergelijk verband bestaat tussen register en het gebruik van het definiet lidwoord zouden we gebruik kunnen maken van het CGN. Zinnen als die in (i) kunnen geëxtraheerd worden uit het corpus door te zoeken op verschillende soortnamen zoals tijger. Het probleem is echter dat dit enorm veel tijd kost en uiteindelijk zeer weinig voorkomens oplevert. Daardoor was het niet mogelijk de resultaten van een dergelijk corpusonderzoek te presenteren.

(i) a De tijger is solitair.
b en jij zegt een tijger is (…) solitair die he leeft in z'n eentje. [fn000506]

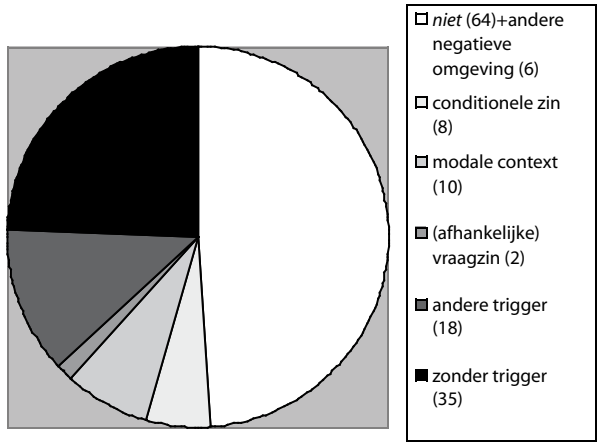
(Voorbeeld (ib) is afkomstig uit een spontane conversatie uit 2000 (Nederland).)

3.2 Casestudie 1: doorgaan *('gehouden worden')*

In Oosterhof (2003-2004) worden de resultaten gepresenteerd van een onderzoek naar de gebruiksomgevingen van *doorgaan* ('gehouden worden') in krantenmateriaal (afkomstig uit de *27 miljoen* en het *38 miljoen woorden corpora* van het INL en het CONDIV-corpus, i.e. het corpus dat werd samengesteld en gebruikt in het kader van het VNC-project *Lexicale variatie in het Standaardnederlands*). De resultaten van die corpusstudie zijn in beknopte vorm gepresenteerd in de figuren 1 en 2.



Figuur 1: Omgevingen van doorgaan *('gehouden worden') in Nederlands krantenmateriaal (N=159).*



Figuur 2: Omgevingen van doorgaan *('gehouden worden') in Belgisch krantenmateriaal (N=142).*

Figuur 1 laat zien dat *doorgaan* ('gehouden worden') in Nederlands krantenmateriaal in slechts 10% van de gevallen voorkomt in omgevingen zonder trigger voor negatief polaire uitdrukkingen (NPUs).[11] Uit figuur 2 blijkt dat in Belgisch krantenmateriaal *doorgaan* frequenter voorkomt in omgevingen zonder trigger, namelijk in 25% van de gevallen. Oosterhof (2003-2004) concludeert op grond van de resultaten in figuur 1 en 2 dat *doorgaan* in het Nederlandse Nederlands een 'semi-NPU' is. Een semi-NPU is een uitdrukking die wel gevoelig is voor polariteit, maar in een beperkt aantal gevallen toch voorkomt zonder trigger en daarom in de strikte zin des woords niet negatief polair is (vgl. Van der Wal 1996). Omdat *doorgaan* in Nederlands materiaal toch nog in 10% van de gevallen in omgevingen zonder trigger voorkomt, gebruiken we de term 'semi-NPU'. De stelling dat *doorgaan* zo'n semi-NPU is, gaat echter niet (of in veel mindere mate) op voor het Belgische Nederlands, omdat het werkwoord in Belgisch materiaal veel vaker, namelijk in 25% van de gevallen, voorkomt zonder trigger.

De conclusie dat *doorgaan* in het Nederlandse Nederlands in tegenstelling tot het Belgische Nederlands een semi-NPU is, is gebaseerd op het aantal voorkomens zonder trigger (zie figuren 1 en 2). Tegelijkertijd laten de figuren echter zien dat *doorgaan* in Belgisch materiaal relatief vaker dan in Nederlands materiaal voorkomt met *niet* of in andere negatieve contexten (i.e. met negatieve uitdrukkingen zoals *geen* en *nooit* of inherent negatieve elementen zoals *alleen* en *slechts*, Van der Wal 1996). Dus in het algemeen komt *doorgaan* in Nederlands materiaal vaker voor met een trigger (waarbij ook conditionele zinnen, modale contexten en een aantal andere omgevingen meetellen, zie noot 11), maar daarbinnen ligt het aantal voorkomens in negatieve contexten hoger in het Belgische materiaal dan in het Nederlandse materiaal.

Een mogelijke verklaring voor dat onverwachte resultaat zou kunnen zijn dat Belgen vooral onder invloed van normdwang vaker geneigd zijn *doorgaan* te gebruiken met *niet* of een ander negatief element. Uit het CGN-voorbeeld in (7) blijkt dat deze verklaring niet zo vergezocht is.[12]

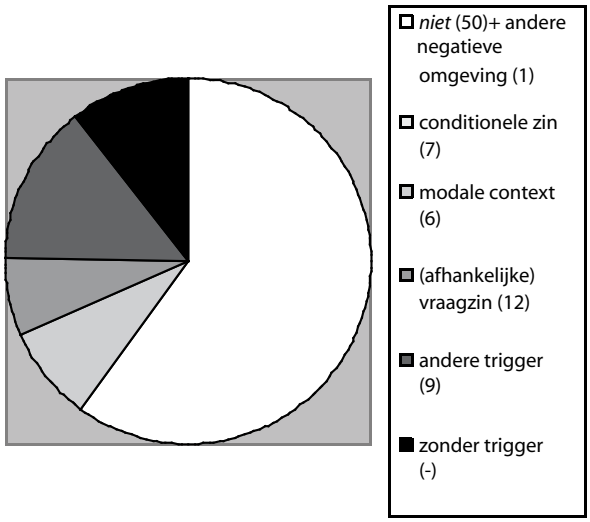
(7) A: 'k zal eerst eens informeren voor uh die langlauftocht. als die doorgaat
B: uhu
A: uhm… klopt dat doorgaan? plaats… ik mis altijd. als ie doorgaat. als ie plaatsvindt. plaatsheeft. [fv400320]

11 Met omgevingen 'zonder trigger' doelen we op omgevingen zonder nonveridicale operatoren. In navolging van Giannakidou (1999) definiëren we nonveridicale operatoren als operatoren waarvoor geldt dat uit het feit dat Op p, waarin Op staat voor een logische operator en p voor de propositie die wordt gemodificeerd door die operator, waar is niet volgt dat de propositie p waar is. Om een voorbeeld te geven: modale werkwoorden zijn nonveridicale operatoren, aangezien een zin als *Die rechtzaak moet/kan doorgaan* niet impliceert dat de betreffende rechtszaak daadwerkelijk doorgaat. In dit geval staat Op p voor een modale operator die correspondeert met *moet/kan* en die van toepassing is op de volgende propositie: *die rechtzaak gaat door*. In dit geval volgt uit het feit dat Op p waar is dus niet dat p (*i.e. de rechtzaak gaat door*) waar is. Voor meer informatie over polariteitsverschijnselen en nonveridicaliteit verwijzen we naar Giannakidou (1999).

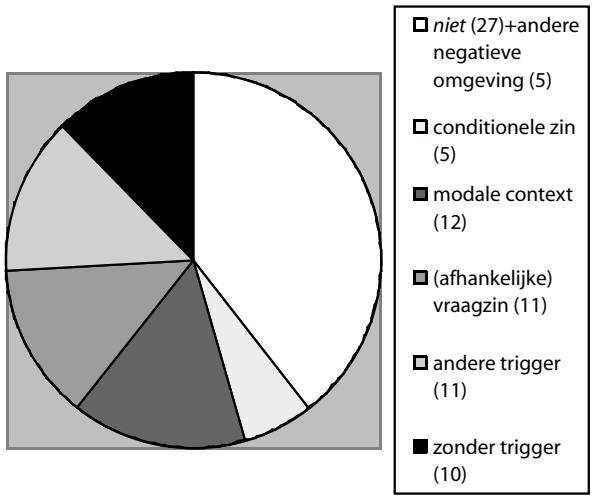
12 Het voorbeeld komt uit een spontane conversatie uit 2001 (Vlaanderen).

In (7) gebruikt spreker A *doorgaat* zonder een ontkenning, waarna hij zichzelf corrigeert in de volgende uiting en andere werkwoorden, namelijk *plaatsvindt* en *plaatsheeft* gebruikt. Een dergelijk geval van zelfcensuur illustreert dat Vlaamse sprekers de indruk hebben dat het gebruik van *doorgaan* ('gehouden worden') zonder ontkenning niet aan de taalnorm voldoet.

Daarnaast is dit stukje dialoog ook interessant omdat het aantoont dat zelfs in spontane conversaties normdwang een rol blijft spelen. Toch is het waarschijnlijk dat het CGN een betrouwbaarder beeld geeft van het Vlaamse gebruik van *doorgaan* (i.e. een beeld dat minder wordt beïnvloed door normgevoelens) dan corpora met krantenmateriaal. Om het verschil tussen Nederlands en Belgisch Nederlands verder te analyseren, betrekken we de voorkomens van *doorgaan* (in de betekenis 'gehouden worden') in het CGN in het verhaal (figuren 3 en 4). Deze resultaten hebben betrekking op het volledige CGN. Er is dus voor gekozen ook de componenten die geen of relatief weinig spontaan taalgebruik bevatten, zoals nieuwsbulletins, plechtige toespraken en colleges (bij elkaar circa 2 miljoen woorden) in het onderzoek te betrekken. We claimen dus niet dat de hier gepresenteerde data uitsluitend een beeld geven van taalgebruik waarbij normdwang geen rol speelt. Wel is het zo dat het CGN-materiaal voor het grootste deel bestaat uit relatief spontaan taalgebruik, dat slechts in beperkte mate onder invloed staat van normdwang. Daardoor zijn deze data geschikter voor onderzoek naar het Vlaamse gebruik van *doorgaan* dan krantencorpora.



Figuur 3: Omgevingen van doorgaan *('gehouden worden') in Nederlands CGN-materiaal (N=85).*



Figuur 4: Omgevingen van doorgaan *('gehouden worden') in Belgisch CGN-materiaal (N=80).*

De resultaten uit het CGN bevestigen deels het beeld dat ook al te zien is in figuur 1 en 2: in Belgisch materiaal komt *doorgaan* vaker voor zonder trigger dan in Nederlands materiaal (Fisher's Exact Test, two-tailed, $p \leq 0.001$). Daarnaast is er in tegenstelling tot de resultaten in Oosterhof (2003-2004) ook een significant verschil tussen Belgisch en Nederlands materiaal wat het aantal voorkomens met *niet* en in andere negatieve omgevingen betreft: in Nederlands materiaal ligt dit aantal nu wel hoger dan in Belgisch materiaal ($\chi^2=6.6$, $p \leq 0.025$).

Het CGN voegt dus duidelijk iets toe aan de resultaten van ander corpusonderzoek. Het is veelzeggend dat in het Belgische materiaal uit het CGN *doorgaan* minder vaak voorkomt in negatieve omgevingen dan in het Belgische krantenmateriaal. Het is aannemelijk dat dit te maken heeft met normdwang, die in de spreektaal en met name in spontaan taalgebruik een minder grote rol speelt dan in (formele) geschreven taal. CGN-data maken het dus mogelijk een completer beeld te geven van reëel taalgebruik en bieden zo een goede basis voor onderzoek naar Noord/Zuid-verschillen.

3.3 Casestudie 2: Zot zijn van *('gek zijn op')*

De uitdrukking *zot zijn van* is duidelijk een uitdrukking die vooral in de spreektaal en dan met name in informele contexten wordt gebruikt. Een corpusvoorbeeld is gegeven in (6), hier herhaald als (8).

(8) behalve spagetti me kaassaus, daar ben ik niet zot van.

Om corpusonderzoek te kunnen doen naar een dergelijke uitdrukking (en naar de polaire gevoeligheid van de uitdrukking), hebben we een corpus nodig dat dergelijk taalgebruik bevat. Het CGN is één van de corpora die aan die eis voldoen, omdat het veel taalgebruik bevat dat relatief spontaan tot stand is gekomen.

In zin (8) verwijst het complement van *zot van* naar voedsel (in dit geval *kaassaus*). We zullen resultaten presenteren waarbij zinnen als (8) worden vergeleken met zinnen als (9), waarin het complement verwijst naar een persoon.[13]

(9) maar ik was ook zot van haar hè. [fv801368]

Het hier gepresenteerde onderzoek is gebaseerd op drie corpora, namelijk een corpus van zinnen van het internet (geëxtraheerd in oktober 2005), een corpus van chatmateriaal uit het CONDIV-corpus (15,2 miljoen woorden) en het volledige CGN. Het aantal voorkomens van de uitdrukking in die corpora is weergegeven in tabel 1.

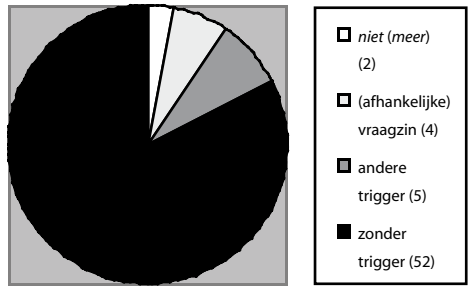
	internet (met Google)	chatmateriaal CONDIV	CGN	N
voedsel	36	3	1	40
personen	46	14	3	63
totaal	82	17	4	103

Tabel 1: Aantal voorkomens van zot zijn van *in 3 corpora.*

In de figuren 5 en 6 zijn de resultaten weergegeven van het onderzoek naar de omgevingen waarin *zot zijn van* voorkomt in de drie corpora .

Figuur 5: Omgevingen van zot van*, waarbij het complement verwijst naar voedsel (N=40).*

13 Het voorbeeld is afkomstig uit voorgelezen tekst uit 2001 (Vlaanderen).



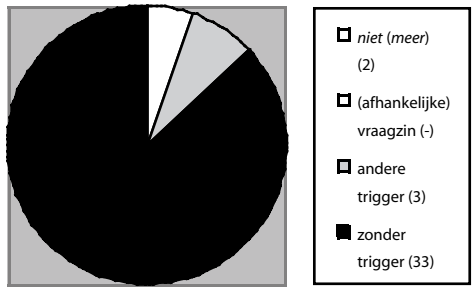
Figuur 6: Omgevingen van zot van, *waarbij het complement verwijst naar personen (N=63).*

Deze figuren maken duidelijk dat *zot zijn van* als het gezegd wordt over voedsel (vgl. (8)) meestal voorkomt in negatieve omgevingen, terwijl *zot zijn van* als het gezegd wordt over personen (vgl. (9)) veel minder vaak voorkomt met negatie. Het verschil is significant (Fisher's Exact Test, two-tailed, $p \leq 0.001$). Dit resultaat wijst erop dat *zot zijn van* in zinnen als (8) polair gevoelig is (zie Oosterhof 2006b voor details en een verklaring).[14]

Het is interessant om na te gaan of hetzelfde resultaat gevonden wordt voor een uitdrukking met een vergelijkbare betekenis, namelijk *gek zijn op*. Deze uitdrukking wordt zowel in Vlaanderen als Nederland gebruikt. Voorbeeldzinnen zijn gegeven in (10) en (11).[15]

(10) oh ik persoonlijk ben gek op balsamico. [fv700093]
(11) Annelies was toen ja uh vier jaar geleden of zo was zij gek op Sjoerd. [fn000390]

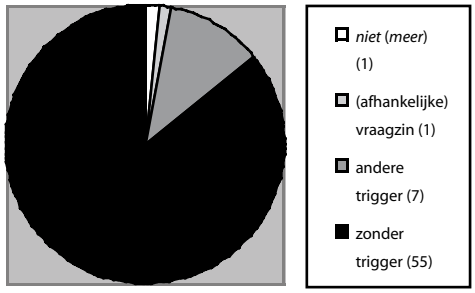
In de figuren 7 en 8 worden de resultaten van een (kleinschalig) corpusonderzoek naar de omgevingen van *gek zijn op* gepresenteerd.



Figuur 7: Omgevingen van gek op *waarbij het complement verwijst naar voedsel (N=38).*

14 De eerlijkheid gebiedt te zeggen dat het effect relatief zwak is. Daarom is het de vraag of de gevoeligheid voor polariteit die we op het spoor zijn gekomen sterk genoeg is om zot zijn van te beschouwen als een NPU (of zelfs als een semi-NPU). Het bestaande onderzoek naar NPUs heeft nog niet geleid tot antwoorden op dergelijke vragen.

15 Zin (10) is afkomstig uit een spontaan telefoongesprek uit 2002 (Vlaanderen). Zin (11) komt uit een spontane conversatie uit 2000 (Nederland).



Figuur 8: Omgevingen van gek op *waarbij het complement verwijst naar personen (N=64).*

Deze resultaten zijn gebaseerd op het volledige CGN (in totaal 26 zinnen) en een corpus van zinnen van het internet (in totaal 79 zinnen, geëxtraheerd in februari 2007). Uit deze figuren blijkt dat *gek zijn op* slechts in een minderheid van de gevallen gebruikt wordt in negatieve omgevingen of in zinnen met een andere trigger (vgl. noot 11). Dat geldt zowel voor zinnen als (10), waarin het complement verwijst naar voedsel als voor zinnen zoals (11), waarin het complement verwijst naar een persoon.

Het hier gepresenteerde corpusonderzoek toont aan dat de uitdrukking *zot zijn van* zoals gebruikt in zinnen als (8), waarin het complement verwijst naar voedsel, gevoelig is voor polariteit. Deze gevoeligheid is een eigenschap die niet gevonden wordt voor de uitdrukking *gek zijn op* met vergelijkbare betekenis.

Het CGN heeft dus als voordeel dat het een aantal componenten bevat waarin de uitdrukking *zot zijn van* kan worden aangetroffen. Tegelijkertijd laat tabel 1 zien dat de meeste voorkomens van *zot zijn van* gevonden werden in het CONDIV-corpus en op het internet. Op dit punt zijn het nut en de bruikbaarheid van het CGN nog relatief beperkt.

4 Beperkingen van het CGN

We zullen de (onvermijdelijke) beperkingen van het CGN voor bepaalde semantische vragen wat verder uitwerken. We beginnen met het bespreken van een drietal beperkingen (4.1). De derde beperking, die belangrijker is dan de eerste twee, zal geïllustreerd worden aan de hand van twee concrete voorbeelden van corpusonderzoek (4.2 en 4.3).

4.1 Drie beperkingen

Een eerste (praktisch) probleem is dat de uitvoering van zoekopdrachten in de exploitatiesoftware van het CGN een aantal minuten tot enkele uren in beslag kan nemen. Vooral zoekopdrachten waarbij naar verscheidene trefwoorden gezocht wordt op verschillende annotatieniveaus (bv. orthografische transcriptie en lemmatisering) vragen erg veel tijd. Om een vergelijking te maken met het *5 Miljoen Woorden Corpus 1994*, het *27 Miljoen Woorden Krantencorpus 1995* en het *38 Miljoen Woorden Corpus 1996* van het INL: bij die corpora neemt een zoekopdracht slechts enkele seconden in beslag. Hetzelfde geldt voor zoekopdrachten op het internet met behulp van Google of een andere zoekmachine. Daar staat

tegenover dat het CGN naast een orthografische transcriptie ook meer informatie bevat dan de meeste bestaande corpora geschreven Nederlands, zoals woordsoortinformatie, lemmatisering en een foneemtranscriptie. Voor veel semantisch onderzoek is dergelijke informatie echter niet van groot belang waardoor de lange **zoektijd** een vervelende hinderpaal blijft voor het meeste semantisch onderzoek.

Een tweede minpuntje is dat de **annotatie** van de geluidsfragmenten niet steeds 100% consequent is. Dit wordt geïllustreerd in (12).[16] In beide zinnen komt een vorm van het werkwoord *uitsterven* voor. Maar volgens de lemma-informatie van het CGN gaat het in zin (12a) om het lemma *uitsterven* en in (12b) om het lemma *sterven*. Stel dat we op basis van het CGN willen onderzoeken welke typen nominale constituenten met welke frequentie voorkomen in subjectpositie van het predicaat *uitsterven* (een zgn. *soortpredicaat*, zie 2.2), dienen we dus te zoeken naar voorkomens van zowel het 'lemma' *uitsterven* als het 'lemma' *sterven*.

(12) a als er niets gebeurt is het gevaar groot dat de kievit in ons land uitsterft. [fn006287]
b nou zijn d'r van allerlei dieren zijn d'r trends hè. ze komen soms vaak voor soms minder vaak. soms uh sterven ze helemaal uit. hoe is het met de vleermuizen eigenlijk. [fn007494]

Beide geschetste beperkingen van het CGN zijn echter niet onoverkomelijk voor semantisch onderzoek. Met wat geduld en creativiteit valt een mouw te passen aan die problemen. Een fundamentelere beperking van het CGN voor bepaalde semantische vragen is de relatief **beperkte omvang** van het corpus, namelijk 9 miljoen woorden. Dat is weinig in vergelijking met geschreven corpora, die samen tientallen miljoenen woorden bevatten, en het internet. We illustreren dit aan de hand van een tweetal voorbeelden (zie 4.2 en 4.3).

4.2 Casestudie 3: uitsterven

Een eerste voorbeeld gaat over zinnen zoals in (12). In Oosterhof (te verschijnen) worden de resultaten gegeven van een corpusonderzoek naar zinnen waarin een vorm van het werkwoord *uitsterven*[17] is gebruikt. Kenmerkend aan dit werkwoord is dat het een soortpredicaat is. In Oosterhof (te verschijnen) wordt nagegaan hoe frequent een viertal typen telbare naamwoordgroepen (definiete enkelvouden, indefiniete enkelvouden, definiete meervouden en kale meervouden) wordt gebruikt in combinatie met *uitsterven*. De resultaten zijn gebaseerd op het *27 Miljoen Woorden Corpus* en het *38 Miljoen Woorden Corpus* van het INL en het CONDIV-corpus. Alles bij elkaar gaat het om een corpus van circa 110 miljoen woorden. In dit corpus werden 60 relevante zinnen gevonden. Het resultaat is weergegeven in tabel 2.

16 Zin (3a) is afkomstig uit een nationaal radionieuws (Nederland) uit 2000; (3b) komt uit een radio-uitzending (Vroege Vogels) uit 2000 (Nederland).

17 Hierbij zijn ook die gevallen betrokken waar *uitgestorven* volgens beschrijvende grammatica's, zoals Haeseryn e.a. (1997:110), ontleed moet worden als een predicatief gebruikt adjectief.

N	definiet enkelvoud		indefiniet enkel-voud		definiet meervoud		kaal meer-voud	
60	23	38%	1	2%	28	47%	8	13%

Tabel 2:Frequenties van vier typen naamwoordgroepen in zinnen met uitsterven *(INL en CONDIV).*

Als een corpus van 110 miljoen woorden slechts 60 relevante zinnen oplevert, valt het te verwachten dat het aantal relevante zinnen in het CGN te laag zal zijn om betrouwbare conclusies te kunnen trekken. Het resultaat voor het CGN is gegeven in tabel 3.

N	definiet enkelvoud	indefiniet enkelvoud	definiet meervoud	kaal meervoud
9	4	0	3	2

Tabel 3: Frequenties van vier typen telbare naamwoordgroepen in zinnen met uitsterven *(CGN).*

Het is van belang om op te merken dat het beperkte aantal relevante zinnen niet betekent dat het niet lonend zou zijn om gebruik te maken van het CGN. We zullen twee redenen geven waarom CGN-data wel degelijk relevant kunnen zijn. Een eerste punt is dat we data uit ander corpusonderzoek kunnen aanvullen met de CGN-data. De data uit tabel 3 kunnen worden toegevoegd aan die in tabel 2, zodat de kans groter wordt dat het onderzoek leidt tot statistisch significante resultaten. Daarnaast is het CGN (zoals elk corpus) een bron van relevante voorbeeldzinnen die allerlei beweringen kunnen ondersteunen. Een voorbeeld van zo'n zin is gegeven in (13).[18]

(13) wilde appels zijn bijna uitgestorven. [fn007498]

In een aantal grammatica's, zoals de *Modern Grammar of Dutch* (Broekhuis e.a. 2003:609), wordt gesteld dat kale meervouden niet gebruikt kunnen worden bij soortpredicaten. Corpusvoorbeelden zoals zin (13) illustreren dat dergelijke zinnen in het Nederlands wel degelijk voorkomen. CGN-voorbeelden laten bovendien zien dat zulke zinnen ook mogelijk zijn in de spreektaal. Hoewel de CGN-data in kwantitatief opzicht slechts in beperkte mate bruikbaar zijn, kunnen ze in kwalitatief opzicht dus wel van belang zijn.

4.3 Casestudie 4: de wemel-*constructie*

Een tweede illustratie van de beperkte omvang van het CGN en de consequenties daarvan betreft een constructie die door Hoeksema (2005:5-6) wordt aangeduid als de *wemel*-constructie. Een tweetal voorbeelden, overgenomen van Hoeksema (2005:5), is te vinden in (14).

(14) a Het wemelt hier van de zwervers
b Het krioelt er van de muizen.

18 Zin (13) is afkomstig uit een radio-uitzending (Vroege Vogels) uit 2000 (Nederland).

Voor inhoudelijke bevindingen over deze constructie verwijzen we naar het werk van Hoeksema (vgl. Hoeksema 2005, 2007). Een voor ons relevante vraag is echter hoeveel relevante voorbeeldzinnen het CGN bevat. In tabel 4 is weergegeven hoe vaak de constructie met een aantal werkwoorden (of werkwoordelijke uitdrukkingen) voorkomt in het CGN. De tien werkwoorden in tabel 4 komen overeen met de werkwoorden die door Hoeksema (2005:5) opgesomd worden.

wemelen	7	*zwart zien*	4	*stikken*	8	*vergeven zijn*	1
krioelen	6	*sterven*	0	*bol staan*	4		
barsten	12	*leven*	0	*ritselen*	0	totaal	42

Tabel 4: Voorkomens van de wemel-*constructie in het CGN.*

Al met al is het duidelijk dat het aantal voorkomens in het corpus vrij beperkt is. Wie op grond van corpusonderzoek de syntactische en semantische eigenschappen van deze constructie wil beschrijven, zou de CGN-data aan moeten vullen met data uit andere corpora.

5 Conclusie

In deze bijdrage zijn enkele troeven en beperkingen van het Corpus Gesproken Nederlands voor semantisch onderzoek besproken aan de hand van een aantal casestudies. Het CGN blijkt vooral door zijn gevarieerde samenstelling interessant te zijn voor semantisch onderzoek. Zo bevat het corpus een selectie spontane en informele teksten waarbij een minimale invloed van normdwang aangenomen mag worden. Daarnaast is de regionale en stilistische diversiteit van de teksten een troef voor onderzoek naar variatie in de semantiek van sommige constructies. Een belangrijke beperking van het CGN voor semantisch onderzoek is de relatief beperkte omvang van het corpus. Dat betekent dat een aantal minder frequente constructies niet of nauwelijks onderzocht kunnen worden door alleen het CGN te gebruiken, maar in combinatie met andere grotere corpora geanalyseerd moeten worden om statistisch significant resultaten te behalen. Al met al betekent het CGN een welkome aanvulling voor bestaand semantisch onderzoek. De winst zit niet zozeer in een kwantitatieve uitbreiding, maar veeleer in een nuancering van de bestaande resultaten, met name ten aanzien van regionale en stilistische variatie.

Bibliografie

Borsley, Robert D. & Richard Ingham (2002). Grow your own linguistics? On some applied linguists' Views of the Subject. *Lingua* 112, 1-6.

Broekhuis, Hans, Evelien Keizer & Marcel den Dikken. (2003). *Modern Grammar of Dutch. Nouns and Noun Phrases. Occasional Papers 4.* Tilburg.

Bruyn, Bert Le (2007). 'Van die dingetjes.' Over verbleekte partitieve constructies. *Over Taal* 46, 45-47.

Chomsky, Noam. (1965). *Aspects of the theory of syntax.* Cambridge, Mass.: MIT Press.

Chomsky, Noam (1988). *Generative Grammar: Its basis, development and prospects.* Kyto: Kyto University of Foreign Studies.

Giannakidou, Anastasia (1999). Polariteitsverschijnselen en (non)veridicaliteit. *Nederlandse Taalkunde* 2, 93-110.

Haegeman, Liliane. (1994). *Introduction to Government and Binding Theory*. Second edition. Oxford: Blackwell.

Haegeman, Liliane (2004). Verdubbeling van subjectpronomina in de Zuid-Nederlandse dialecten: een reactie uit Lapscheure. *Taal en Tongval* 56, 119-159.

Haeseryn, Walter, Karin Romijn, Guido Geerts, Jaap de Rooy & Maarten van den Toorn. (1997). *Algemene Nederlandse Spraakkunst*. tweede, geheel herziene druk. Groningen: Wolters Noordhoff.

Hoeksema, Jack (2004). *De negatief-polaire uitdrukkingen van het Nederlands. Inleiding en lexicon*. Manuscript. University of Groningen.

Hoeksema, Jack (2005). Rijkdom en weelde van het Nederlands. *TABU* 34, 1-12.

Hoeksema, Jack (2007). *The SWARM-alternation revisited.* Manuscript. Swarthmore College & Rijksuniversiteit Groningen.

Hoop, Helen de, Guido Vanden Wyngaerd & Jan-Wouter Zwart. (1990). Syntaxis en semantiek van de *van die*-constructie. *Gramma* 14, 81-106.

Kress, Günther. (1993). Cultural Considerations in Linguistic Description. In: David Graddol, Linda Thompson & Mike Byram (red.), *Language and Culture.* Clevedon: Multilingual Matters, 1-22.

McEnery, Tony & Andrew Wilson (2001). *Corpus linguistics.* 2nd edition. Edinburgh: Edinburgh University Press.

Meurers, Walt D. (2005). On the Use of Electronic Corpora for Theoretical Linguistics. Case Studies from the Syntax of German. *Lingua* 115: 1619-1639.

Oosterhof, Albert (2003-2004). Polariteitsgevoeligheid van *doorgaan* ('gehouden worden'). *TABU* 33, 131-150.

Oosterhof, Albert (2005). Verbleekte partitieven: Descriptieve, syntactische en semantische aspecten. *Neerlandistiek.nl* 5, 1-28.

Oosterhof, Albert (2006a). *Generics in Dutch and related languages. Theoretical and empirical perspectives.* Proefschrift, Universiteit Gent.

Oosterhof, Albert (2006b). Twee polair gevoelige items in het Belgische Nederlands. *Over Taal* 45, 10-13.

Oosterhof, Albert. (te verschijnen). De Empirische Basis van Semantisch Onderzoek. *Gramma/ttt* 10.

Verkuyl, Henk J. (1998). O corpora, O mores. *Nederlandse Taalkunde* 3, 60-63.

Vries, Mark de. **(2005).** Ellipsis in nevenschikking: voorwaarts deleren maar achterwaartsdelen.' *TABU* 34: 13-46.

Wal, Sjoukje van der. (1996). *Negative Polarity Items and Negation: Tandem Acquisition.* Proefschrift, Rijksuniversiteit Groningen.

Zwart, C. Jan-Wouter (1997). *Morphosyntax of Verb Movement. A Minimalist Approach to the Syntax of Dutch.* Dordrecht: Kluwer.

Modale verschillen tussen België en Nederland in het CGN

RONNY BOOGAART, PIETER BYLOO, JANNEKE DIEPEVEEN, THEO JANSSEN EN JAN NUYTS*

Abstract

The Corpus of Spoken Dutch was used to investigate differences in the expression of modality between two varieties of Dutch, namely Belgian Dutch and Netherlandic Dutch. Methodological considerations include the selection of the relevant subcorpora, the distinction between global counts and detailed analyses, and the variables used in statistical analysis. While a number of supposed modal differences between Belgian Dutch and Netherlandic Dutch could be confirmed or rejected on the basis of corpusdata, there were also issues that remained unresolved because the corpus did not provide enough data for a reliable statistical test to be possible. Some of the modal expressions that we were interested in turned out to be relatively infrequent in language use and, therefore, in the corpus. In other instances, however, the lack of data can be ascribed to specific properties of the corpus and its search program Corex.

1 Inleiding

In het VNC-project *Modale hulpwerkwoorden en andere uitdrukkingen van modaliteit in het Nederlands* hebben we geprobeerd voor modale verschillen tussen het Belgisch-Nederlands en het Nederlands-Nederlands die her en der in de literatuur worden genoemd, of die we door verkennend corpusonderzoek en eigen intuïties op het spoor waren gekomen, vast te stellen of het hier daadwerkelijk significante verschillen tussen de twee taalvariëteiten betreft. In dit artikel doen we verslag van dat deel van het onderzoek dat is uitgevoerd met behulp van het Corpus Gesproken Nederlands (CGN, versie 1.0).[1]

Gezien het doel van het onderzoek was het noodzakelijk noch zinvol om de categorie modaliteit, en de subcategorieën daarvan, van tevoren precies af te bakenen. We hebben ons dus niet beperkt tot epistemisch-modale uitdrukkingen, die de mate van waarschijnlijkheid van een propositie betreffen, maar hebben ook uitdrukkingen bestudeerd die de

* Het onderzoek waarvan we hier verslag uitbrengen, is uitgevoerd in het kader van het project 'Modale werkwoorden en andere uitdrukkingen van modaliteit in het Nederlands. Een corpusgebaseerd onderzoek', gesubsidieerd door het *Vlaams Nederlands Comité voor Nederlandse taal en cultuur* (VNC). Correspondentieadres: r.j.u.boogaart@let.leidenuniv.nl. De auteurs bedanken de redactie van *Nederlandse Taalkunde* en de anonieme beoordelaar voor waardevolle opmerkingen bij een eerdere versie.

1 In het tweede deel van het onderzoek zijn de modale verschijnselen waarvoor in het CGN niet voldoende gegevens beschikbaar waren, nader getoetst in een enquêteonderzoek. In de discussie in paragraaf 4 presenteren we een aantal uitkomsten daarvan, zie verder Diepeveen e.a. (2006).

ANS (1997:1617 e.v.) 'oneigenlijk modaal' zou noemen of helemaal niet onder modaliteit behandelt. Voorbeelden zijn het deontische of directieve gebruik van *kunnen* in Belgisch-Nederlandse uitingen als *je kunt zien dat je het morgen af hebt*, het 'volitionele' gebruik van *moeten* voor de taalhandeling aanbod (*moeten jullie nog wat drinken?*), maar ook adviserende constructies met *beter* en *best* (*je kunt beter/(het) best(e) de bus nemen*). Het laatste voorbeeld laat al zien dat we ook wat de formele kant van modaliteit betreft het net wijd hebben uitgegooid: naast modale hulpwerkwoorden zijn onderzocht deontische hoofdwerkwoorden als *horen te*, *dienen te*, *toelaten* en *toestaan*, modale adverbia zoals *vast*, *zeker* en *waarschijnlijk* en de modale infinitiefconstructie, die bijvoorbeeld voorkomt in *het is/valt te verwachten*.

We gebruiken de aanduidingen 'Belgisch-Nederlands' en 'Nederlands-Nederlands', die we verder afkorten tot *BN* en *NN*, voor het Nederlands gesproken in het Nederlandstalige deel van België en het Nederlands gesproken in Nederland omdat daarmee duidelijk wordt aangegeven dat het gaat om nationale variëteiten van hetzelfde Nederlands (Geeraerts 2001:338). Regionale verschillen binnen België en binnen Nederland blijven buiten beschouwing aangezien ons onderzoek geen dialectonderzoek is: we houden ons bezig met bovenregionaal taalgebruik in beide landen zoals dat in het CGN is gerepresenteerd (maar zie 4.3).

In paragraaf 3 geven we een beknopt overzicht van de belangrijkste uitkomsten van het onderzoek. Daarnaast besteden we in dit themanummer over het CGN speciaal aandacht aan methodologische aspecten (paragraaf 2) en aan problemen die we in de loop van het corpusonderzoek hebben ondervonden (paragraaf 4).

2 Methodologie

2.1 Selectie subcorpora

Aangezien we geïnteresseerd zijn in verschillen tussen taalgebruik in Nederland en in België, ligt het voor de hand om in ons onderzoek gebruik te maken van het onderscheid tussen 'Dutch regions' en 'Flemish regions' zoals dat op het openingsscherm van het zoekprogramma Corex wordt gemaakt. Na selectie van één van de twee subcorpora – 'Dutch' dan wel 'Flemish' – kunnen we met *Content Search* laten zoeken naar een specifieke modale vorm. We herhalen de zoekopdracht voor het andere subcorpus en moeten bij het vergelijken van de twee uitkomsten dan alleen nog rekening houden met het verschil in de omvang van het BN en het NN subcorpus (zie 2.3).

Toch hebben we van deze door Corex aangeboden zoekstrategie geen gebruik gemaakt. Het onderscheid tussen 'Flemish regions' en 'Dutch regions' in het CGN is gebaseerd op de afkomst van de sprekers: sessies met sprekers uit België zijn bij de subcategorie 'Flemish regions' ondergebracht; sessies met sprekers uit Nederland bij de subcategorie 'Dutch regions'. Het probleem is dat het corpus ook opnamesessies bevat waarin zowel sprekers uit België als sprekers uit Nederland aan het woord zijn. Zulke 'gemengde' sessies zijn bijvoorbeeld nieuwsuitzendingen van de Vlaamse openbare omroep VRT waarin een Nederlandse correspondent verslag uitbrengt. Als gevolg daarvan is er enige overlap tussen de subcorpora Flemish regions en Dutch regions: de 'gemengde' sessies komen dubbel voor.

We hebben er daarom voor gekozen om de plaats van opname als criterium te nemen. Sessies die in België zijn opgenomen, zijn in het CGN met *fv* gelabeld, terwijl sessies opge-

nomen in Nederland het label *fn* gekregen hebben. Deze subcorpora zijn niet aanklikbaar op het openingsscherm van Corex, maar moeten voorafgaand aan de *Content Search* met *Metadata Search* worden gedefinieerd (zie de bijdrage van Van Eerten in dit nummer voor verdere toelichting en illustraties). Ook hier blijft het probleem dat het materiaal sessies bevat waarin sprekers van beide variëteiten voorkomen. Onze voorkeur gaat desondanks uit naar de *fv*/*fn*-verdeling omdat daar de 'gemengde' sessies niet dubbel voorkomen en de verschillen maximaal blijven.

In beide subcorpora hebben we op één na alle tekstcategorieën uit het CGN gebruikt. Uitgesloten is de categorie 'Read speech', die bestaat uit voorgelezen passages van romans. Zulke teksten hebben voor publicatie meestal een nauwkeurige (normaliserende) revisie ondergaan, waardoor bijvoorbeeld eventuele BN-getinte eigenschappen mogelijk verwijderd zijn. Dat zou vanzelfsprekend het taalvariatieonderzoek vertroebelen.

2.2 Grove telling tegenover detailanalyse

Voor tellingen in de subcorpora is gebruik gemaakt van het zoekprogramma Corex. Een telling die uitsluitend gebaseerd is op het resultaat van een zoekopdracht, zonder verdere ingrepen van de onderzoeker, noemen we een grove telling. Het is echter niet altijd juist om op een grove telling af te gaan. Het zoekresultaat kan aanzienlijk van omvang zijn, maar erg veel ruis bevatten, dat wil zeggen uitingen die om een of andere reden niet relevant zijn voor het fenomeen dat we onderzoeken. Die redenen kunnen zijn: een betekenis die niet relevant is, het ontbreken van interactie tussen de relevante elementen in de uiting (bijv. er is wel een negatief element aanwezig, maar dat staat niet in verband met de modale vorm), en zo meer. In bepaalde gevallen blijkt het onmogelijk om de zoekopdracht zó te verfijnen dat ruis in het zoekresultaat uitgesloten of tot een minimum gereduceerd wordt. Een bijkomend probleem is dat het toevoegen van extra zoekrestricties natuurlijk juist tot missers kan leiden: relevante treffers die we mislopen doordat de zoekopdracht te beperkend is geworden.

Een oplossing voor het probleem van de ruis, en een alternatief voor de grove telling, is de methode van de detailanalyse. Daarbij bekijken we alle afzonderlijke uitingen in het zoekresultaat. We filteren de ruis eruit en alleen de relevante treffers worden meegeteld. Als het aantal treffers heel groot is (het loopt tot in de honderd- of duizendtallen), is een detailanalyse van het volledige zoekresultaat niet haalbaar. In dat geval voeren we een detailanalyse uit op een steekproef. Bij de samenstelling van een steekproef uit treffers van meer dan één zoekopdracht, zoals bij de vergelijking van twee verschillende vormen, moet uiteraard rekening gehouden worden met de verhouding tussen het aantal treffers van beide zoekopdrachten: de verhoudingen in het aantal te analyseren treffers zijn gelijk aan de verhoudingen in het totale aantal treffers van de zoekopdrachten. Als zoekopdracht 1 in een taalvariëteit 1500 treffers oplevert en zoekopdracht 2 er 2250 te zien geeft, is het aantal te analyseren treffers voor die taalvariëteit bijvoorbeeld 20 voor zoekopdracht 1 en 30 voor zoekopdracht 2.

2.3 Methodes van vergelijking

Het resultaat van de zoekopdrachten, in absolute aantallen, biedt op zich weinig informatie over verschillen tussen NN en BN. Naast de frequentiegegevens van een vorm X in beide

subcorpora, is steeds een tweede paar gegevens nodig ter vergelijking. Aan de hand van die vergelijking kan met behulp van een chikwadraattoets bepaald worden of de frequentie van vorm X significant groter is in het NN dan in het BN subcorpus. In overeenstemming met de gangbare praktijk in taalkundige corpusstudies nemen we als significantiedrempel 0,05 aan, dat wil zeggen dat de p-waarde kleiner moet zijn dan 0,05 ($p < 0,05$). We hebben in het onderzoek drie vergelijkingsmethodes gebruikt, die we hier kort toelichten.

In de eerste plaats kan het aantal treffers voor een bepaalde vorm X worden vergeleken met het totale aantal woorden in het subcorpus van de desbetreffende taalvariëteit. Het BN subcorpus bevat in totaal (los van de categorie 'Read speech') 2.910.209 woorden; het NN subcorpus bestaat uit 5.103.020 woorden. Op die manier kan bijvoorbeeld eenvoudig worden aangetoond dat de modale bijwoorden *waarschijnlijk, wellicht* en *allicht* frequenter zijn in BN dan in NN (voor alle drie geldt: $p = 0,000$).[2] Een probleem van deze methode is dat hij eigenlijk niet toegepast kan worden op uitdrukkingen die uit meer dan één woord bestaan: de vraag is namelijk of het dan wel terecht is het totale aantal woorden als vergelijkingsbasis te nemen. Het zou dan logischer zijn om te contrasteren met het totale aantal uitingen (of annotation units), en wel als volgt: het aantal uitingen waarin de uitdrukking voorkomt, wordt vergeleken met het aantal uitingen waarin ze niet voorkomt. Helaas brengt dat weer andere moeilijkheden met zich mee, bijvoorbeeld als de uitdrukking meer dan eens voorkomt in een uiting. Bovendien is de lengte van een uiting als eenheid in het CGN vrij arbitrair en verschilt die sterk per sessie.

Een tweede vergelijkingsmethode wordt ingezet als we alternatieve uitdrukkingsvormen voor 'dezelfde' betekenis met elkaar vergelijken: het aantal treffers voor vorm X wordt dan vergeleken met het aantal treffers voor vorm Y. De chikwadraattoets laat zien of de verhouding tussen X en Y in de ene taalvariëteit significant verschilt van de verhouding tussen X en Y in de andere taalvariëteit. Een voorbeeld is het gebruik van *moest*(*en*) en *mocht*(*en*) in conditionele constructies van het type *moest/mocht je me nodig hebben*. De variant met *moeten* komt in het NN subcorpus niet voor, terwijl we in het BN subcorpus 51 gevallen hebben gevonden. (Dat zijn bovendien zeker niet alle gevallen uit het corpus, zie 4.2.) Het verschil tussen BN en NN is significant ($p = 0,000$). Een probleem bij deze methode is dat we in het corpus niet van betekenis naar vorm kunnen zoeken, dus naar alle verschillende uitdrukkingsmogelijkheden van een bepaalde betekenis of functie als 'conditionaliteit'. De vormen die we met elkaar vergelijken hebben we zelf gekozen op grond van onze intuïtie of opmerkingen in de literatuur dat de twee vormen in gelijksoortige contexten worden gebruikt. We kunnen echter niet uitsluiten dat ook functionele criteria van belang zijn bij bijvoorbeeld de keus tussen *mocht*(*en*) en *moest*(*en*). In ieder geval kan conditioneel *moest*(*en*) in BN niet altijd vervangen worden door *mocht(en)* in NN en dus is het de vraag in hoeverre de twee uitdrukkingen 'dezelfde betekenis' hebben (Boogaart 2006b, 2007). Bovendien bestaat natuurlijk altijd de mogelijkheid dat we alternatieve uitdrukkingsmogelijkheden voor een bepaalde betekenis over het hoofd hebben gezien.

De derde en laatste methode wordt gebruikt als we vermoeden dat er verschillen zijn tussen de betekenissen die BN en NN toeschrijven aan één en dezelfde vorm, dus voor gebruiksverschillen. Deze methode houdt in dat een bepaalde betekenis of een bepaald

2 In absolute aantallen is de uitkomst als volgt. *Waarschijnlijk*: 1335 (NN) tegenover 1604 (BN); *allicht*: 21 (NN) tegenover 49 (BN); *wellicht*: 124 (NN) tegenover 146 (BN).

gebruik van een vorm vergeleken wordt met andere gebruikwijzen van diezelfde vorm. De intuïtie bestond bijvoorbeeld dat specifiek het epistemische gebruik van het modale bijwoord *vast* frequenter is in NN dan in BN; voorbeelden (uit het corpus) zijn *'t lukt vast wel* en *die lui die geloven vast niet dat jij een universitaire opleiding hebt gehad*. Om dat aan te tonen zouden we het aantal treffers voor epistemisch *vast* kunnen contrasteren met het aantal voorkomens van *vast* in alle overige betekenissen. De vraag is echter of zo'n vergelijking tot interessante vaststellingen leidt. Het lijkt weinig zinvol epistemisch *vast* te contrasteren met uitdrukkingen die volledig andere betekenissen dekken enkel en alleen omdat ze homoniem zijn. Bovendien is het mogelijk dat een van de 'overige' betekenissen frequent voorkomt in een van de taalvariëteiten, waardoor ten onrechte het vermoeden gewekt wordt dat de onderzochte betekenis er relatief zeldzaam is. Een voorbeeld is het frequente voorkomen van *ik stel vast* in BN: dat in BN *vast* minder vaak epistemisch gebruikt is, zou dan eerder iets zeggen over de uitdrukking *ik stel vast* dan over het epistemische gebruik.[3]

Een probleem bij de analyse van gebruiksverschillen is verder dat de grens tussen verschillende modale lezingen van één en dezelfde vorm niet altijd gemakkelijk te trekken is. Dat geldt bijvoorbeeld voor het onderscheid tussen volitioneel *willen* en frequentatief *willen* (als in *hij wil nog wel eens te laat komen*), als ook voor de grens tussen volitioneel *willen* en 'doelaanduidend' *willen* (als in *wil het project slagen, dan moeten we hard doorwerken*). Het is waarschijnlijk dat in zulke gevallen een proces van grammaticalisatie aan de gang is: het frequentatieve en doelaanduidende gebruik van *willen* is niet incompatibel met het letterlijke, volitionele gebruik van het werkwoord en is daar waarschijnlijk uit ontstaan. Als gevolg daarvan is het in veel concrete gevallen bijna onmogelijk te besluiten of de letterlijke betekenis van het werkwoord (ook) nog in het geding is (volgens Goossens 2000 is dat bij doelaanduidend *willen* bijvoorbeeld altijd het geval). Een voorbeeld is het gebruik van *wil* in (1).

(1) d'r zal heel wat bij Heerenveen moeten veranderen wil de ploeg nog voorkomen dat Amsterdam vandaag de eerste titel sinds negentien vijfentachtig binnenhaalt. (fn008957)

Hier is een parafrase van het type 'Als de ploeg wil voorkomen (…)' niet uitgesloten en het werkwoord *willen* lijkt in de parafrase letterlijk, dus volitioneel, bedoeld te zijn; (1) zou dan een 'normaal' geval van de conditionele inversie-constructie kunnen zijn (zoals in: *wil ik naar de film, dan ga ik ook*). Het geval in (2) is een duidelijker voorbeeld van de zuiver doelaanduidende *wil*-constructie aangezien daar in de bijzin geen (expliciet) bezield subject aanwezig is waarmee een volitionele agens wordt aangeduid.

(2) wil de evaluatie zinnig zijn dan zullen we de wet echt wat tijd moeten gunnen (fn000163)

3 In dit specifieke geval hebben we desondanks wel aan kunnen tonen dat epistemisch *vast* veel frequenter is in NN dan in BN, maar dan via de eerste vergelijkingsmethode. In een steekproef van 100 BN en 100 NN gevallen hebben we de aantallen epistemisch *vast* bepaald (resp. 2 en 30) en die gerelateerd aan proporties van de overige woorden in het CGN (p = 0,000). (Vgl. 2.2 over het nemen van een steekproef.)

Detailanalyse laat zien dat de doelaanduidende constructie met *willen* in BN minder vaak voorkomt dan in NN (p = 0,011). Zo'n analyse van gebruiksverschillen levert veel interessante gegevens op, maar is zeer ingewikkeld en tijdrovend. Daar komt nog bij dat bij een 'minimale' telling, waarin alle tussen- en twijfelgevallen worden uitgesloten, het aantal overblijvende gegevens in het CGN al gauw te klein wordt om een betrouwbare statistische toets toe te laten (zie verder 4.1).

3 Resultaten

Voor een groot aantal mogelijke verschillen tussen BN en NN in de uitdrukking van modaliteit hebben we met behulp van het CGN weten aan te tonen dat het in het corpus statistisch significante verschillen tussen de twee taalvariëteiten betreft ($p < 0,05$), of juist dat het verschil niet significant is ($p > 0,05$). We maken in het onderstaande overzicht een onderscheid tussen uitdrukkingen met modale werkwoorden en uitdrukkingen met modale bijwoorden.[4] Voor verdere toelichting en getalsmatige uitkomsten per kwestie verwijzen we naar Diepeveen e.a. (2006:13-71).

Uitdrukkingen met modale werkwoorden

- Een van de bekendste verschillen tussen BN en NN op het gebied van modaliteit wordt in het CGN bevestigd: het werkwoord *hoeven* is frequenter in NN dan in BN; BN gebruikt *niet moeten* met de betekenis waarvoor NN *niet hoeven* gebruikt. Zoals Van der Wouden (1998:90-91) terecht opmerkt over het BN: "het gevolg is dat een zin als *je moet niet komen* dubbelzinnig of vaag is tussen 'je hoeft niet te komen' en 'je bent verplicht, niet te komen'".
- Het werkwoord *moeten* kent in BN nog andere gebruikswijzen die in NN ongewoon zijn. Zo heeft BN de conditionele constructie met *moest(en)* die we in 2.3 al noemden. Het conditionele gebruik van *moeten* komt voor in *als*-zinnen (*als ik rechter moest zijn ze zouden d'r hier in België van beleven*) en in de conditionele inversieconstructie waarin *moeten* zelf het eerste woord van de voorwaardelijke bijzin is (*moest het niet lukken dan belt de maar efkes*). (In 4.2 bespreken we een belangrijke complicatie bij het zoeken naar de laatste groep gevallen). In het NN subcorpus ontbreekt de conditionele constructie met *moest(en)* geheel. Het NN gebruikt 'ter compensatie' van het ontbreken van conditioneel *moeten* overigens niet meer dan BN werkwoorden als *mocht(en)* of *zou(den)* in conditionele constructies (zie verder Boogaart 2007).
- BN gebruikt vaker dan NN evidentiële en epistemische constructies met *moeten* of *zullen* en een voltooid complement (*hij zal niet veel tijd gehad hebben zeker hè*). Voor de werkwoorden *moeten* en *zullen* als zodanig, dus los van de aard van het complement (voltooid/onvoltooid), kan geen significant verschil worden vastgesteld.

4 Daarnaast hebben we, zoals gezegd in de inleiding, nog gekeken naar de modale infinitiefconstructie (*het bedrag is te betalen*). Die komt significant meer voor in BN dan in NN, maar we vonden geen evidentie voor de claim uit de literatuur dat de constructie in BN vaker dan in NN een deontische verplichtingslezing krijgt (zie verder ook Boogaart 2006a).

- De combinatie '*graag mogen* + infinitief' om een favoriete activiteit aan te duiden (*ik mag wel graag een beetje met computers klooien en lezen en uh fietsen*) is een typische NN constructie. De constructie komt in het BN subcorpus zelfs niet één keer voor.
- Ook de conditionele constructie met 'doelaanduidend' *willen*, geïllustreerd in (1) en (2), blijkt typisch voor NN, al blijft, als gezegd in 2.3, de classificatie van deze gevallen een complexe taak.
- *Durven* is frequenter in BN dan in NN en kent een frequentatief gebruik, voor een zo nu en dan voorkomend gebeuren, dat in het NN subcorpus niet voorkomt (*'t enige is dat de die beesten nogal wat lawaai durven maken 's nach enfin 's morgens vroeg*); in NN wordt hiervoor uitsluitend *willen* gebruikt (*hier thuis spreken we dan Nederlands nou en dan wil het wel eens krom gaan*). Net als bij *willen* (zie 2.3) is ook bij *durven* de precieze grens tussen letterlijk gebruik ('het lef hebben om') en frequentatief gebruik lastig te bepalen.
- NN gebruikt veel frequenter dan BN het werkwoord *zullen*, met name *zal ik/zullen we*, voor taalhandelingen als aanbod, voorstel, en suggestie. Alleen in het BN subcorpus wordt in zulke contexten een enkele maal *willen* gebruikt, zoals in *willen we toch maar naar 't restaurant gaan?*, maar daarvan zijn in het CGN te weinig gevallen om een betrouwbare statistische toets mogelijk te maken.[5]
- NN en BN verschillen van elkaar in de aan- of afwezigheid van het werkwoord *kunnen* bij bepaalde infinitieven, zoals in *ik vind het niet* (frequenter in BN) tegenover *ik kan het niet vinden* (frequenter in NN). Dergelijke verschillen zijn eerder opgemerkt voor het Nederlands en het Engels bij 'verbs of involuntary perception' (Goossens 1983): het Engels gebruikt bijvoorbeeld *I can see* waar het Nederlands eerder kiest voor *ik zie het*, dus zonder modaal hulpwerkwoord. Voor het werkwoord *zien*, net als voor *begrijpen*, is er geen verschil tussen BN en NN, maar voor een aantal andere werkwoorden wel: NN gebruikt vaker *kunnen* met *niet vinden*, *zich niet herinneren* en *niet verstaan*, terwijl het modaal werkwoord hier in BN vaker ontbreekt. BN gebruikt wel weer vaker *kunnen* bij *niet geloven* en *nooit weten*.[6]
- Ondanks claims in de literatuur dat BN eerder *kunnen* dan *kennen* gebruikt om de kennis van een taal aan te duiden (Peeters 1930:268, Goossens 1983:150,152), hebben we hiervoor in het CGN geen significant verschil gevonden. Overigens komt zowel de constructie met *kunnen* als die met *kennen* met aanduidingen van talen significant meer voor in BN dan in NN. Dat kan betekenen dat talenkennis sowieso vaker onderwerp van gesprek is in het BN subcorpus, maar ook dat NN hier vaker kiest voor een constructie zonder modaal werkwoord zoals in *hij spreekt Arabisch*.
- De deontische constructie met *horen te* wordt meer gebruikt in NN dan in BN. Voor *dienen te* is er geen significant verschil.
- Het deontische hoofdwerkwoord *toelaten* kent een ruimer gebruik in BN dan in NN; het wordt er met name gebruikt met de betekenis 'toestemming geven', waar NN eerder voor *toestaan* kiest. (Ook het zelfstandig naamwoord *toelating* is frequenter in BN.)

5 Mogelijke BN alternatieven in deze context zijn *moeten* (vgl. 'volitioneel' *moeten* in 4.1) en *gaan*, dat in België veel van de functies van *zullen* overneemt. Maar het is natuurlijk ook mogelijk dat BN-sprekers voor deze taalhandelingen formuleringen kiezen die helemaal geen expliciet modaal element bevatten.

6 De beperking tot negatieve contexten bleek nodig om aan voldoende relevante gevallen te komen. Het lijkt erop dat 'ability' in positieve contexten minder relevant of informatief is dan in ontkennende zinnen (en vragen).

Uitdrukkingen met adverbia

- De uitdrukking *zeker en vast* wordt frequent gebruikt in BN, maar komt in het NN subcorpus niet voor. Overigens is de omgekeerde uitdrukking *vast en zeker* in NN niet significant frequenter dan in BN. *Zeker* is op zich frequenter in BN, terwijl *vast* juist meer in NN wordt gebruikt. (Waarbij weer de vraag gesteld kan worden of beide vormen wel 'dezelfde' betekenis hebben, vgl. 2.3.)
- Met name het epistemisch gebruik van *vast* is typisch NN. Van de in NN frequent gebruikte combinaties *vast wel* en *vast niet* komt geen enkel geval in het BN subcorpus voor. In BN is *waarschijnlijk*, en ook de combinatie *waarschijnlijk wel/niet*, frequenter dan in NN.
- Zowel *wellicht* als *allicht* wordt meer gebruikt in BN dan in NN. Bovendien zijn er in het gebruik van deze woorden subtiele verschillen tussen BN en NN (Janssen & Diepeveen ms.)
- In adviserende uitingen met *beter* gebruikt NN systematisch het werkwoord *kunnen* (*ja dan kun je beter diesel nemen*), terwijl dat in BN vaker ontbreekt (*'k had het misschien beter eerder gedaan*; *dan zou ik beter (…) morgen al bellen*).
- BN heeft een voorkeur voor *liefst*, terwijl NN vaker voor *het liefst* kiest.
- *Hopelijk* is frequenter in BN dan in NN.

Het bovenstaande overzicht laat zien dat het CGN zeer geschikt is om taalvariatie tussen België en Nederland, in ons geval op het specifieke terrein van modaliteit, in kaart te brengen. Echter, voor een ongeveer even groot aantal vermoede verschillen tussen BN en NN dan in het overzicht vermeld gaf ons corpusonderzoek – veelal letterlijk – geen uitkomst te zien.

4 Problemen bij het corpusonderzoek

Voor alle in de vorige paragraaf genoemde modale verschijnselen kon in het CGN worden vastgesteld of het hier significante verschillen tussen BN en NN betreft. Wanneer een mogelijk verschil in het corpus niet significant blijkt te zijn, is dat natuurlijk net zo goed een onderzoeksresultaat, althans zolang we genoeg data hebben gevonden om een betrouwbare statistische toets uit te voeren. Het laatste bleek in een aantal andere gevallen een probleem te zijn. De oorzaak daarvan is dan ofwel dat de modale uitdrukking niet frequent genoeg voorkomt in het corpus (4.1), ofwel dat de uitdrukking in het corpus lastig op te sporen is (4.2). Die twee oorzaken zijn natuurlijk niet goed van elkaar te scheiden: als het moeilijk is een bepaalde modale constructie in het CGN te vinden, bijvoorbeeld door beperkingen van het zoekprogramma of door het ontbreken van relevante codering in het corpus, weet je in principe niet of de constructie wel frequent genoeg voorkomt om conclusies mogelijk te maken. We bespreken in de volgende paragrafen enkele illustratieve gevallen. Een laatste probleem dat we aan de orde stellen, betreft de regionale variatie in de uitdrukking van modaliteit *binnen* België en *binnen* Nederland (4.3).

De kwesties waarvoor we in het CGN niet voldoende data vonden, hebben we opgenomen in een aanvullend enquêteonderzoek: een schriftelijke groepsenquête onder ongeveer 900 studenten (geen talenstudenten) aan zes verschillende universiteiten in

België en Nederland. In de volgende subparagrafen verwijzen we een aantal malen naar de uitkomsten van dat enquêteonderzoek, de volledige beschrijving en alle resultaten zijn te vinden in Diepeveen e.a. (2006:73-117).

4.1 De modale uitdrukking komt in het CGN niet frequent genoeg voor

Twee interessante kwesties waarin we vanwege te weinig data in het CGN geen conclusie konden trekken over een verschil tussen NN en BN, zijn het deontisch gebruik van *kunnen* en het volitioneel gebruik van *moeten*.

Op grond van opmerkingen in de literatuur vermoedden we dat het BN *kunnen* gebruikt met een deontische betekenis in contexten waarin het NN bijvoorbeeld eerder voor een imperatief zou kiezen. Zo geeft Van Dale (2005:1866) aan het voorbeeld *Je kunt zien dat je klaar komt* de labels 'Belgisch-Nederlands' en 'spreektaal' (vgl. ook Goossens 1983:154 en Van Es 1973:413). Er zijn nauwelijks beperkende zoekcriteria te bedenken om in het corpus specifiek het deontisch gebruik van *kunnen* op te sporen: eerder onderzoek, van Nuyts (2001) en Van Ostaeyen en Nuyts (2004), heeft al laten zien dat het deontisch gebruik van dit werkwoord marginaal is. We beperken ons daarom tot de combinatie van *kunnen* met *zien* en *zorgen*. In NN kan de letterlijke, 'visuele' betekenis van *zien* plaats maken voor een deontische zonder dat er *kunnen* bij wordt gebruikt (*je ziet maar dat je het af krijgt, zie maar dat je het af krijgt*). Hetzelfde geldt voor *zorgen*, dat in bepaalde contexten eveneens een deontisch gebruik kent (*je zorgt maar dat je het af krijgt, zorg maar dat je het af krijgt*). We zoeken dus in het corpus naar de combinatie van *kunnen* met de werkwoorden *zien* en *zorgen*. Een detailanalyse is nodig om gevallen van dynamisch *kunnen* uit te sluiten, van het type *je kunt zien dat hij ziek is: hij is erg bleek*. In het hele corpus vinden we slechts twee relevante (BN) gevallen van *kunnen zien dat*:

(3) dat is zo de ne stempel van van gij zijt germanist dus ge kunt maar zien dat je dat je geen fouten spreekt. (fv400549)

(4) ja en 't was al van kunt zien dat je niet valt en dat je u niet pijn doet want 'k ga kwaad zijn. (fv400265)

Voor *kunnen zorgen dat* is er zelfs geen enkel geval. We kunnen hieruit niet concluderen dat het verschil tussen BN en NN niet statistisch significant is: met slechts twee relevante gevallen is er geen statistisch betrouwbare toets mogelijk. In ons aanvullende enquête-onderzoek bleek dat het deontisch gebruik van *kunnen*, overigens ook met andere werkwoorden dan *zien* en *zorgen*, in België wel degelijk veel frequenter is dan in Nederland (zie 4.3).

Doorgaans biedt men mensen iets te eten of te drinken aan met *willen*, zoals *in willen jullie nog koffie?*, maar het is ook mogelijk om *moeten* te gebruiken. De *ANS* (1997:996) noemt het gebruik van *moeten* in uitingen als *moeten jullie nog koffie?* 'informeel' en Van Dale (2005:2188) noemt 'willen' als oudste betekenis van *moeten*, zonder regionaal label. Volitioneel *moeten* zou stilistisch gemarkeerd of contextueel bepaald zijn. Wij sluiten echter niet uit dat het gebruik (ook) samenhangt met de nationale taalvariëteit. Misschien wordt dit gebruik van *moeten* door BN-sprekers minder als substandaard of informeel aangevoeld dan door NN-sprekers. Het is moeilijk om via een zoekopdracht in het CGN relevante uitingen van dit type te vinden. We hebben onze zoekopdracht beperkt tot uitingen die

eindigen met een vraagteken, omdat de kans dan het grootst is dat we met een aanbod te maken hebben. Door daarbij *moeten* te combineren met aanduidingen van gangbare soorten drank en met de werkwoorden *eten* en *drinken* hebben we gevallen gevonden als in (5).

(5) hé moet je wat drinken onderwel of zo? (fn007836)

Maar de aantallen blijven ook in deze kwestie te klein om een betrouwbare statistische toets mogelijk te maken. Opnieuw laat enquêteonderzoek wel een resultaat zien: de BN-respondenten kiezen in een dergelijke context vaker voor *moeten* dan de NN respondenten, die eerder *willen* zouden gebruiken. (Er is binnen Nederland wel regionale variatie vast te stellen, zie 4.3).

4.2 De modale uitdrukking is lastig op te sporen in het CGN

Terwijl het voor deontisch *kunnen* en volitioneel *moeten* voorstelbaar is dat het type taalhandeling waarin het specifieke gebruik van deze werkwoorden optreedt, sowieso relatief infrequent is in het taalgebruik en/of in het corpus, is er ook een aantal gevallen waarin het tekort aan data primair toe te schrijven lijkt aan beperkingen van het CGN en met name van het zoekprogramma Corex.[7] We geven hiervan drie voorbeelden: (i) het is niet mogelijk om een zoekopdracht te beperken tot het begin van (deel)zinnen, (ii) de codering voor woordsoort is voor ons doel soms niet specifiek genoeg, (iii) er kunnen geen negatieve zoekrestricties worden opgegeven.

Omdat de modale werkwoorden zeer frequent gebruikte werkwoorden zijn, is het vaak moeilijk om specifieke gebruikswijzen van die werkwoorden uit het corpus te filteren. We merkten dat hierboven al op voor het deontisch gebruik van *kunnen*. Een ander voorbeeld: in een steekproef van 50 BN en 50 NN gevallen van *moeten*, vinden we slechts één geval van epistemisch (of misschien liever 'evidentieel', zie Nuyts 2001) *moeten*. Om zoveel mogelijk ruis te verwijderen, zou het voor een aantal kwesties handig zijn om de zoekopdracht te kunnen beperken tot de optredens van modale werkwoorden aan het begin van (deel)zinnen. Dat geldt bijvoorbeeld voor de al eerder genoemde conditionele constructies met *moest(en)* en *mocht(en)*, voor de doelaanduidende *willen*-constructie, voor volitioneel *moeten* en voor het gebruik van modale werkwoorden in de taalhandeling 'aanbod' (*zal ik/zullen we*). Zo'n beperkte zoekopdracht zou nog steeds ruis bevatten, die hoofdzakelijk uit vraagzinnen bestaat, maar die kan vervolgens in een detailanalyse handmatig worden verwijderd. Het probleem is dat zoeken op zinsbegin in het CGN eigenlijk niet mogelijk is.[8] Met enig kunst- en vliegwerk lukt het wel om te zoeken naar het begin van 'annotatie-eenheden', maar niet alle annotatie-eenheden in het CGN zijn zinnen en niet alle zins-

7 Het CGN kan in principe ook met andere zoekprogramma's dan Corex worden doorzocht. In dit artikel beperken we ons echter tot een evaluatie van het CGN zoals dat, inclusief Corex, 'op de markt' is.

8 In het syntactisch geannoteerde deel van het CGN is dat wel mogelijk, maar vanwege de beperkte omvang van dat subcorpus komen we daar zeker niet aan genoeg gevallen.

9 Alle optredens van dezelfde vorm aan het begin van een annotatie-eenheid komen bij elkaar te staan als je het zoekresultaat laat alfabetiseren. In Corex zelf is dat overigens niet mogelijk.

grenzen vallen samen met een grens tussen twee annotatie-eenheden.[9] Met deze methode komen we in het geval van de conditionele *moeten*-constructie en de doelaanduidende *willen*-constructie wel (net) aan voldoende relevante gegevens, maar we weten zeker dat we veel gevallen hebben gemist en dat de constructie feitelijk veel frequenter is dan onze absolute aantallen laten zien.[10]

Een tweede zoekprobleem heeft te maken met de woordsoortlemmatisatie. Het is bij het werken met een bestaand corpus onvermijdelijk dat het corpus in bepaalde gevallen niet de onderscheidingen maakt die de onderzoeker goed van pas zouden komen. Zo wilden we nagaan of BN, behalve met *niet*, ook vaker dan NN het werkwoord *moeten* gebruikt met restrictief *maar*. We verwachten natuurlijk dat de trend zich doorzet en dat NN dan de voorkeur aan *hoeven* geeft. Relevante voorbeelden van dit verschil, zoals in (6) en (7), zijn wel te vinden in het CGN.

(6) ja ja ja wel wel honderd tachtig vragen maar je hoeft er *maar* tien goed te hebben of zo. (fn000188)
(7) hij moet nog *maar* zijn bek opendoen en je hoort onmiddellijk dat ie van Antwerpen is. (fv400325)

Toch is het lastig aan te tonen dat het verschil in gebruik van *hoeven* en *moeten* in combinatie met *maar* een significant verschil is. Een grove telling van de combinatie *hoeven/moeten* en *maar* bevat veel ruis, die bestaat uit ten minste drie groepen niet-relevante gevallen: (i) gevallen met *maar* als voegwoord, (ii) gevallen met *maar* als niet-beperkend partikel, en (iii) gevallen waarin het moeilijk vast te stellen is of we er al dan niet met beperkend *maar* te maken hebben. Voorbeelden van de laatste categorie zijn (8) en (9), waarin *maar* ambigu lijkt tussen beperkend en 'aansporend' gebruik.

(8) maar ja zeg Nadine als ge ziek zijt ge moet *maar* bellen hè. (fv700182)
(9) maar ik denk dat ja mmm mensen ik geloof graag in de goedheid van de mensen maar 't wordt dagdagelijks je moet de krant *maar* lezen bewezen dat er mensen zijn die intrinsiek niet goed zijn. (fv600613)

Een beperking van de zoekopdracht tot *maar* als partikel zou waarschijnlijk wel voldoende gegevens hebben opgeleverd, maar die mogelijkheid biedt het corpus niet. Een detailanalyse van een steekproef van 100 BN en 100 NN gevallen weerlegt onze hypothese niet, maar levert nog steeds te weinig relevante gegevens op voor een betrouwbare statistische toets.[11]

In de kwestie van restrictief *maar* zou veel ruis vermeden kunnen worden door tenminste het voegwoordelijk gebruik van *maar* in de zoekopdracht uit te sluiten. Maar hier stuiten we op een laatste, algemener probleem van het zoekprogramma Corex, namelijk dat het niet mogelijk is om negatieve zoekrestricties op te geven. Die beperking speelde

10 De methode veronderstelt bovendien dat we in het BN subcorpus ongeveer evenveel gevallen mislopen als in het NN subcorpus en dat de afbakening van annotatie-eenheden in beide subcorpora vergelijkbaar is.

11 In de enquête kiest meer dan 70% van de BN-sprekers voor *moeten* in een restrictieve context met *maar*, terwijl de NN-respondenten daar met 96% massaal voor *hoeven* kiezen

ons vooral parten bij het opsporen van 'hypercorrect' *hoeven*. In zijn woordenboek van belgicismen wijst Peeters (1930) al op het gebruik van hoeven "in bevestigenden zin", met citaten als *Om half twee hoefde ze in de kerk te zijn* (zie verder ook Daman 1946:34, Veering 1959:50). Recent vermeldt ook Haeseryn (1996:120) nog dat *hoeven* in België uit hypercorrectie "wel eens voorkomt in positieve contexten". Van der Wouden (1998:71) doet een poging om dat te verklaren: *hoeven* "is een moeilijk werkwoord voor de meeste Belgen". Daar zijn andere taalkundigen het mee eens. Doordat *hoeven* in het Nederlands van België een exogeen woord is, kan het makkelijk hypercorrect gebruikt worden in zinnen zonder negatie of restrictie (Vakgroep Nederlandse Taalkunde 2003:145-6; De Vriendt 1995:24). Wij hebben het idee dat hypercorrectie met *hoeven* in de praktijk zeer zeldzaam is, in elk geval in hedendaags gesproken BN, maar treffers voor hypercorrect *hoeven* zijn niet simpelweg te vinden in het corpus: zoeken op *hoeven* minus negatieve of restrictieve elementen als *niet*, *nauwelijks* of *maar* is niet mogelijk met Corex. Een steekproef van 100 BN en 100 NN voorkomens van *hoeven* levert één geval op:

(10) en of ze wel goed terechtkomen daar waar ze hoeven terechtkomen. (fv600582)

Betrouwbare toetsen zijn met zo'n uitkomst niet mogelijk. Daarom is het item opgenomen in onze enquête: als *moeten* en hypercorrect *hoeven* daar als alternatieven worden aangeboden, kiest geen enkele NN-respondent en maar een enkele BN-respondent (2,03 %) voor *hoeven*, maar het verschil is wel significant ($p = 0{,}025$).

4.3 Taalinterne regionale variatie

Zoals al opgemerkt in de inleiding is ons onderzoek geen onderzoek naar dialecten. Toch zijn gegevens over variatie binnen België en binnen Nederland van belang: uitgesloten moet worden dat een bepaald verschijnsel lokaal is voor we mogen concluderen dat het een NN of BN verschijnsel betreft. In ons corpusonderzoek zijn intratalige regionale verschillen echter buiten beschouwing gebleven. Daar zijn twee redenen voor. De eerste is dat de samenstellers van het CGN zich expliciet ten doel hebben gesteld alleen bovenregionaal taalgebruik op te nemen. We hanteren in ons onderzoek dus alleen de *fn*/*fv*-verdeling in het corpus (zie 2.1), zonder verdere regionale specificatie. De tweede reden is dat het vaststellen van de herkomst van alle individuele sprekers in het corpusmateriaal in principe wel mogelijk is, maar lastig uitvoerbaar. Het is bovendien de vraag of zo'n exercitie veel zou opleveren: gezien de moeilijkheden die we hadden om voor het hele NN en BN taalgebied aan voldoende gegevens te komen (zie 4.1 en 4.2), is het voor de meeste kwesties bijna uitgesloten dat we voor elke Belgische en Nederlandse regio genoeg data zouden vinden om nog een betrouwbare toets uit te kunnen voeren.[12]

De uitkomst van ons aanvullende enquêteonderzoek laat nochtans zien dat we hier wel met een reëel probleem te maken hebben. In dit deel van ons onderzoek konden we de herkomst van de respondenten, gedefinieerd als de regio waar ze basisonderwijs

12 Onderzoek naar regionale variatie binnen België en Nederland is zeer wel mogelijk in het CGN als het gaat om variabelen met een (veel) hogere frequentie dan de hier onderzochte modale constructies, zie bijvoorbeeld Vandekerckhove (2005).

volgden, natuurlijk gemakkelijk meenemen als één van de variabelen in de statistische analyse. Om een representatieve steekproef te krijgen, hebben we Nederland en België in zes regio's gesplitst en hielden we een schriftelijke groepsenquête onder studenten op zes verschillende universiteiten (Antwerpen, Gent, Hasselt, Amsterdam, Groningen, Nijmegen). Voor iedere regio beschikten we over ten minste 100 respondenten. Net zoals de samenstellers van het CGN waren we geïnteresseerd in bovenregionaal taalgebruik en we hebben veel voorzorgsmaatregelen genomen, met name in de mondelinge introductie door de enquêteur, om dialectinterferentie, en overigens vooral ook norminterferentie, zoveel mogelijk te voorkomen (Diepeveen e.a. 2006:88-91). Toch hebben we voor een aantal van de veertien met de enquête onderzochte verschijnselen, naast BN-NN-variatie, significante regionale variatie kunnen vaststellen. In Nederland is er slechts regionale variatie voor één verschijnsel, namelijk het gebruik van *willen* versus 'volitioneel' *moeten* in een aanbod (*willen/moeten jullie nog wat drinken?*) (vgl. 4.1): in de NN regio Noordwest is de voorkeur voor *willen* veel sterker dan in de twee andere NN regio's (Noordoost en Zuid). In België zijn vijf verschijnselen regionaal verschillend gebleken. Zo sluit de BN regio Oost zich bij de keus tussen *hoeven* en *moeten* in een context met restrictief *maar* (vgl. 4.2) meer aan bij NN; in de BN regio's Centrum en West wordt vaker voor *moeten* gekozen. De deontisch-directieve constructie met *kunnen* en *zien/zorgen* (vgl. 4.1) wordt dan weer het meest gekozen in de BN regio West.

Onder de Nederlandse respondenten in de enquête bestaat dus een grotere eensgezindheid over het gebruik van modale uitdrukkingen dan onder de Belgische; de BN variëteit is kennelijk minder 'gevestigd' dan de NN variëteit. Die uitkomst is in het licht van de literatuur over taalvariatie in Nederland en België niet helemaal onverwacht, maar het laat nog eens duidelijk zien dat de inhoud van de notie 'bovenregionaal' taalgebruik, zoals ook de samenstellers van het CGN die hanteren, vooral voor België niet vanzelfsprekend is. We kunnen dus niet uitsluiten dat ook in de kwesties uit paragraaf 3 bij nader onderzoek taalinterne regionale variatie zou blijken te bestaan.

5 Tot besluit

Voor de studie van modale variatie tussen België en Nederland is het CGN een waardevol onderzoeksinstrument gebleken. Een aantal van de modale uitdrukkingen waarin we geïnteresseerd zijn, zijn echter relatief infrequent, met als gevolg dat we in het CGN vaak niet genoeg data vinden voor een betrouwbare statistische toets. Bovendien gaat het bij modale verschillen om (soms vrij subtiele) gebruiksverschillen, die per definitie een kwalitatieve geval-per-geval-analyse van het zoekresultaat noodzakelijk maken. Dat probleem is niet principieel of onoverkomelijk: met een (vaak veel) grotere steekproef, of met een ander zoekprogramma dan Corex, zou een aantal kwesties misschien alsnog met het CGN op te lossen zijn. Wij hebben niet altijd zo'n arbeidsintensieve detailstudie uitgevoerd, omdat het ons doel was binnen het bestek van het onderzoek zoveel mogelijk modale verschillen tussen BN en NN te toetsen. Daarom hebben we gekozen voor een gecombineerde verzameling van spontane spreekdata uit het CGN en testdata uit een enquête-onderzoek (vgl. Cornips 1995). Daarnaast hebben de intuïties van de Nederlandse en Belgische onderzoekers binnen het project over hun eigen taalvariëteit vanzelfsprekend

een grote rol gespeeld, alleen al bij de selectie van te onderzoeken kwesties en bij de formulering van de antwoordmogelijkheden in de enquête, maar natuurlijk evenzeer bij de interpretatie van de data uit het CGN. Hoewel de analyse van authentiek taalmateriaal, zoals dat in het CGN is verzameld, in een onderzoek naar gebruiksverschillen de voorkeur verdient boven geëliciteerde oordelen, is een combinatie van onderzoeksmethoden in ons onderzoek onvermijdelijk gebleken.

Referenties

Algemene Nederlandse spraakkunst. Zie Haeseryn e.a. (1997).

Boogaart, Ronny (2006a). "Het is te zien hoe dat je het ziet": de modale infinitief in Nederland en België. *Voortgang. Jaarboek voor de Neerlandistiek* 24, 37-50.

Boogaart, Ronny (2006b). "Moest ik jong zijn…" Voorwaardelijke bijzinnen in Noord en Zuid. In: *Over Taal* 45, 125-127.

Boogaart, Ronny (2007). Conditionele constructies met *moest(en)* en *mocht(en)* in Belgisch-Nederlands en Nederlands-Nederlands. *Neerlandistiek.nl* 07.05.

Boon, Ton den & Dirk Geeraerts (2005). *Van Dale groot woordenboek der Nederlandse taal.* Veertiende, herziene uitgave. Utrecht: Van Dale Lexicografie.

Cornips, Leonie (1995). De voordelen van testdata in sociolinguïstisch, syntactisch onderzoek. In: Erica Huls & Jetske Klatter-Folmer (red.). *Artikelen van de Tweede Sociolinguïstische Conferentie.* Delft: Eburon, 111-124.

Daems, Frans & Louis Goossens (1983, red.). *Een spyeghel voor G. Jo Steenbergen.* Leuven: Acco.

Daman, Johannes A. (1946). *Het Algemeen Beschaafd in Vlaanderen (Zuid-Nederland).* Gent: Daphne.

Diepeveen, Janneke, Ronny Boogaart, Jenneke Brantjes, Pieter Byloo, Theo Janssen, Jan Nuyts, m.m.v. Hanne Kloots (2006). *Modale uitdrukkingen in Belgisch-Nederlands en Nederlands-Nederlands: corpusonderzoek en enquête*. Amsterdam: Stichting Neerlandistiek VU, Münster: Nodus Publikationen.

Es, Gustaaf A. van (1973). Nederlands uit de pen van Zuidnederlanders. In: COLLECTIEF [PEE Willem] (red.), *Album Willem Pée, de jubilaris aangeboden bij zijn zeventigste verjaardag*. Tongeren: Michiels, 409-416.

Geeraerts, Dirk (2001). Een zondagspak? Het Nederlands in Vlaanderen: gedrag, beleid, attitudes. *Ons Erfdeel* 44, 337-343.

Goossens, Louis (1983). *Can* and *kunnen*: Dutch and English potential compared. In: Daems & Goossens (red.), 147-158.

Goossens, Louis (2000). "Zal de hazelnoot u smaken,... Schuw dan niet, den bast te kraken": Noot over modale werkwoorden en modaliteit in het WNT. In: Gillis, Steven; Jan Nuyts & Johan Taeldeman (samenst.). *Met taal om de tuin geleid. Een bundel opstellen voor Georges De Schutter ter gelegenheid van zijn pre-emeritaat.* Antwerpen: Reprografie UIA, 193-202.

Haeseryn, Walter e.a. (1997). *Algemene Nederlandse spraakkunst.* Groningen: Martinus Nijhoff / Deurne: Wolters Plantyn.

Haeseryn, Walter (1996). Grammaticale verschillen tussen het Nederlands in België en het Nederlands in Nederland: een poging tot inventarisatie. In: Roeland van Hout & Joep Kruijsen (red.). *Taalvariaties. Toonzettingen en modulaties op een thema.* Dordrecht: Foris, 109-126.

Janssen, Theo A.J.M. & Janneke Diepeveen (ms.) *Allicht* en *wellicht* in Belgisch- en Nederlands-Nederlands. [aangeboden aan: *Southern African Linguistics and Applied Language Studies*].

Nuyts, Jan (2001). *Epistemic Modality, Language and Conceptualization. A Cognitive-Pragmatic Perspective*. Amsterdam: John Benjamins.

Peeters, Constant H. (1930). *Nederlandsche taalgids: woordenboek van Belgicismen*. Antwerpen: De Sikkel.

Ostaeyen, Gert Van & Jan Nuyts (2004). De diachronie van *kunnen*. *Antwerp Papers In Linguistics* 109. Antwerpen: Universiteit Antwerpen.

Vakgroep Nederlandse Taalkunde (2003). *De woordsoorten in het Nederlands. Met taalzorgaantekeningen.* Gent: Universiteit Gent.

Van Dale (2005). Zie Den Boon & Geeraerts.

Vandekerckhove, Reinhild (2005). Belgian Dutch versus Netherlandic Dutch: new patterns of divergence? On pronouns of address and diminutives. *Multilingua. Journal of Cross-Cultural and Interlanguage Communication* 24, 379-397

Veering, Jan (1959). Het Zuidnederlandse taaleigen voor het Noordnederlands taalgevoel. *Taal en Tongval* 11, 48-54.

Vriendt, Sera De (1995). *Van geen kleintje vervaard. Essays over Nederlandse taalwetenschap.* Brussel: VUBPress.

Wouden, Ton van der (1998). *Verboden op het werk te komen: klein woordenboek van Vlaamse taal- en andere eigenaardigheden*. Enschede: SIWA.

Bouwstenen van het Nederlands en het CGN

TON VAN DER WOUDEN*

Abstract

The division of labor between grammar and the lexicon is a recurrent point of debate in linguistics. Within the project *Dutch as a construction language*, the question is taken up as an empirical problem. In the current paper, we present results of quantitative and qualitative investigations into Extended Lexical Units in real life Dutch, as exemplified by the Spoken Dutch Corpus CGN.

Algemeen wordt tegenwoordig ingezien, dat men bij de beschrijving van talen, in typus geheel van de Indogermaansche verschillend, er niet komt met de begrippen en namen, waaraan de Latijnsche grammatica ons heeft gewend. Hieruit volgt vanzelf, dat men evengoed een nieuw schema tot aanvulling, ten deele ook tot vervanging, van het oude noodig heeft, als men nieuwe igm. idiomen moet beschrijven, die zich in structuur sterk van de oude onderscheiden. Zelfs mag men verwachten, dat ook van die laatste sommige trekken beter zullen uitkomen, wanneer men ze beschouwt op andere wijs dan de traditioneele. Nu behoort zonder twijfel de ongedwongen spreektaal der Nederlanders in al hare verscheidenheden tot de zeer moderne idiomen, veel meer dan b.v. die der Duitschers, als men de grensstreken buiten beschouwing laat; het schijnt mij zelfs toe, dat zij met die der Scandinaviërs en Engelschen een buitengewoon dankbaar veld is voor den syntacticus, omdat hier de wortels van het oerwoud genoegzaam zijn vergaan om ruimte te maken voor een rijke nieuwe vegetatie, welke voor Engeland het volledigst, voor onze gewesten nog zeer weinig is gedetermineerd. (De Vries 1910:2)

1 Inleiding: spreektaal

Spreken en verstaan komt voor lezen en schrijven. Lang voordat kinderen beginnen met leren lezen en schrijven, kletsen ze je al de oren van je hoofd. Lezen en schrijven moeten je ook aangeleerd worden, maar spreken en luisteren leer je vanzelf. En waar de geschiedenis van taal wordt geschat in termen van honderdduizenden jaren, dateert het eerste bekende schrift hooguit van een aantal duizenden jaren geleden. Er zijn veel taalgemeenschappen zonder schrift, maar alleen dode talen met een schrift als het Latijn, het Sanskriet en het Hettitisch hebben misschien meer schrijvers en lezers dan sprekers.

* LUCL/Opleiding Nederlandse taal en cultuur, VIDI-project *Dutch as a construction language*, gefinancierd door NWO en UL. Contact: t.van.der.wouden@let.leidenuniv.nl. Een deel van het materiaal in dit artikel is aan de orde gekomen in verschillende voordrachten in binnen- en buitenland. Wij danken het aldaar aanwezige publiek voor op- en aanmerkingen, en Mika Poss en de redactie van *Nederlandse Taalkunde* voor commentaar op een eerdere versie van dit stuk. Paragraaf 3.4 vertoont overlap met Van der Wouden (2007b), 3.5 met Van der Wouden (2007a).

Schrijven is ooit uitgevonden en geadopteerd als een representatie van spreektaal, maar dat betekent niet dat de modaliteit het enige verschil tussen schrijf- en spreektaal is; met name in cultuurtalen met een lange schrijftaaltraditie kunnen de twee behoorlijk uit elkaar gaan lopen (Bolinger & Sears 1981). Een voorbeeld: intonatie in gesproken Nederlands correspondeert maar zeer ten dele met interpunctie in geschreven Nederlands. Miller & Weinert (1998) gaan zelfs zo ver te beweren dat spreektaal en schrijftaal (deels) verschillende systemen zijn in vrijwel ieder opzicht: morfologie, syntaxis, woordenschat, zowel als tekstorganisatie. Bovendien: Biber (1988) laat zien dat het mogelijk en zinvol is verschillende soorten schrijftaal en verschillende soorten spreektaal te onderscheiden. Het zal duidelijk zijn dat in bellettrie een ander soort taal wordt gehanteerd dan in dagbladen, en dat in een zitting van de Tweede Kamer anders wordt gesproken dan op een voetbaltribune (vergelijk ook de kwantitatieve gegevens in Uit den Boogaart 1975).

Eén belangrijke oorzaak voor die verschillen tussen schrijftaal en spreektaal is de tijdsdruk die inherent is aan spontane spreeksituaties: die laat minder ruimte en tijd voor complex mentaal rekenwerk dan de acribische compositie van een doorwrochte geschreven tekst vraagt. En één van de strategieën die sprekers gebruiken om dit probleem te omzeilen, is vooral niet proberen origineel te zijn. Kuiper (1996) rapporteert dat de taaluitingen bij geoefende sprekers onder hoge druk (denk bijvoorbeeld aan veilingmeesters of sportverslaggevers) voornamelijk bestaan uit kant-en-klare stukken taal, samengevoegd met behulp van een minimale hoeveelheid grammaticaal rekenwerk: het invullen van een naam (*Huntelaar*) of een ander soort variabele (*Huntelaar passt op* ***rechts***). We kunnen aannemen dat dat het ene uiterste is van een glijdende schaal, met aan het andere eindpunt het Chomskyaanse ideaaltype van een creatieve taalgebruiker met een onfeilbaar geheugen en onbeperkte grammaticale rekenvermogens, die, niet gehinderd door tijdgebrek, nieuwe ideeën en gedachten op een volstrekt originele wijze verwoordt.

Tijdsdruk is overigens beslist niet de enige reden om gebruik te maken van kant-en-klare stukken taal. Gewoonte of ritualisering is een andere (Wray 2002): als iemand in een min of meer formele situatie de aandacht wil van het gehele gezelschap, dan ligt gebruik van de aanhef *dames en heren* voor de hand, ook als hij niet gewend is de aanwezigen individueel met *dame* respectievelijk *heer* aan te spreken. *Heren en dames* daarentegen is veel minder geschikt voor dat doel: niet omdat deze frase ongrammaticaal in strikte zin is, maar omdat je het nu eenmaal zo niet zegt (zie ook Pawley & Syder 1983). Anderzijds kan schrijftaal eigenschappen ontwikkelen of conserveren onder invloed van onderwijs, regels, wetten enzovoort.

Het onderzoek waarover we hier rapporteren gaat ondertussen over spreektaal, en hierboven genoemde 'kant-en-klare stukken taal' staan daarin centraal. Dat heeft namelijk als doel, meer licht te werpen op het aandeel en de typen al of niet compositionele, talige bouwstenen die groter zijn dan woorden in spontane spreektaal. Het Corpus Gesproken Nederlands (in het vervolg: CGN) maakt het voor het eerst mogelijk dit soort onderzoek systematisch te doen.[1]

1 Niet onvermeld mag blijven dat er een zekere traditie van onderzoek bestaat naar eigenaardigheden van de Nederlandse spreektaal. Die traditie begint zo ongeveer bij de Groningse gymnasiumleraar Wobbe de Vries (1910, 1911, 1914), werd voortgezet in de dissertaties van Bertha Uijlings (1956) en Frank Janssen (1981), en vindt zijn voorlopige bekroning in het dikke maar onevenwichtige boek van Jelle de Vries (2001). Al deze studies hebben zeker hun waarde, maar ze zijn gebaseerd op veel minder authentiek taalmateriaal dan wat er nu dankzij het CGN beschikbaar is.

2 Achtergrond en opzet van het onderzoek

2.1 De taakverdeling tussen grammatica en lexicon

In de traditionele visie (bijv. Chomsky 1981, maar ook ver daarvoor al) bestaat het taalsysteem grosso modo uit een Lexicon, dat is een lijst van Woorden, en een Grammatica, een systeem van Regels. Woorden zijn de bouwstenen, dat wil zeggen discrete elementen, tekens in de zin van De Saussure (1916): de kleinste zelfstandige eenheden met een bepaalde vorm en een bepaalde betekenis. De Grammatica specificeert de manier waarop de woorden gecombineerd worden (hoewel de woorden daar zelf ook een bijdrage aan kunnen leveren) en de betekenis van de resulterende structuur. De betekenis van het geheel is een functie van de betekenissen van de samenstellende delen en de manier waarop die zijn gecombineerd, net zoals je van simpele, geometrische LEGO-blokjes zoiets ingewikkelds als een villa, een Amerikaanse president, of een hele marinehaven kunt bouwen.

Tegenwoordig wordt een andere zienswijze (weer) populair, namelijk dat het lexicon meer is dan een oninteressante lijst van woorden. Volgens deze alternatieve visie bevat het lexicon ook tal van 'geprefabriceerde', 'kant-en-klare' grotere eenheden met hun eigen semantiek en hun eigen gebruiksmogelijkheden. Zoals al eerder aangegeven: als iemand de aandacht van het publiek wil, zeg hij *dames en heren*, en niet *heren en dames*, en om die frase te kunnen gebruiken heeft hij geen grammatica nodig, want hij is kant-en-klaar beschikbaar. En één van de vele manieren om beleefd te laten merken dat men een einde wil maken aan een telefoongesprek is *ik ga maar eens ophangen*, hoewel de letterlijke betekenis van die uiting iets heel anders is (en vrijwel niemand zijn telefoon(hoorn) nog ophangt). In de metafoor van de LEGO-blokjes: behalve de algemeen bruikbare klassieke rechthoekige blokjes zijn er ook speciale, onregelmatiger gevormde bouwsteentjes, die je gebruikt als dakpan, als oorlel, als plechtanker of iets dergelijks. Die speciale rol vervullen ze beter, maar voor andere doelen zijn ze veel minder geschikt. Met andere woorden: hoe groter de specialisatie, hoe beperkter de toepasbaarheid, en omgekeerd.

In het vervolg zullen we de grotere lexicale eenheden in taal en taalgebruik pretheoretisch aanduiden met de term 'ELU', voor 'Extended Lexical Unit', omdat we ons (nog) niet willen committeren aan theoretisch beladener noties als lexical phrase (Pawley & Syder 1983), fraseolexeem (Wotjak 1992), constructie (Goldberg 1995), multiword expression (Sag et al. 2001), formule (Wray 2002), idioom (Everaert 2003), enzovoort.

Een verwante notie 'constructie' speelde traditioneel een belangrijke rol in de grammaticale analyse. Fillmore et al. (1988) hebben in een programmatisch artikel over *let alone* (een uitdrukking vergelijkbaar met het Nederlandse *laat staan*, als in *Hij kent niet eens Engels, laat staan Frans* – het voorbeeld is afkomstig uit de ANS (Haeseryn et al. 1997, zie ook Verhagen 1994)) gepleit voor een eerherstel voor constructies. Die suggestie is onder meer uitgewerkt in de monografie *Constructions* van Goldberg (1995), en de term 'constructiegrammatica' verenigt tegenwoordig een actieve groep min of meer gelijkgestemde taalkundigen die ervan uitgaan dat grotere lexicale eenheden een centrale (Goldberg) of op zijn minst belangrijke (Wray 2002) rol spelen in taalproductie, taalperceptie en taalverwerving (Tomasello 2003). Een radicaal standpunt wordt ingenomen door Croft (2001),

die alle verklarende kracht in de grotere lexicale eenheden stopt en geen aparte grammatica meer nodig heeft of meent te hebben.

Daarmee staat de vraag van de taakverdeling tussen grammatica en lexicon weer op de onderzoeksagenda, een vraag overigens die op verschillende manieren opgevat wordt. Enerzijds staat hij centraal in wat een ideologische discussie genoemd zou kunnen worden. Sterk lexicalistische theorieën zoals Head Driven Phrase Structure Grammar (HPSG, Pollard & Sag 1994) hebben, vanuit andere theoretische uitgangspunten, andere antwoorden op deze vraag dan theorieën waarin de autonomie van de syntaxis centraal staat (Chomsky 1981 enz.).

Anderzijds kan de vraag naar de taakverdeling tussen grammatica en lexicon ook als een empirische worden opgevat (vergelijk Jackendoff 1997). Dan gaat het om de taakverdeling in het taalgebruik: welk deel van echte taaluitingen wordt berekend, welk deel wordt kant-en-klaar uit het lexicon opgehaald? De schattingen in de literatuur over het aandeel van ELU's aan feitelijk taalgebruik lopen sterk uiteen: Altenberg (1998) schrijft over het Engels: 'A rough estimation indicates that over 80 per cent of the words in the corpus form part of a recurrent word-combination in one way or another', maar Sprenger (2003) komt voor het Nederlands een orde van grootte lager uit: zo'n tien procent. Dat kan liggen aan een verschil tussen de twee talen, maar het kan ook heel goed een definitiekwestie zijn. Sprenger heeft namelijk alleen naar inhoudswoorden gekeken, en als vaste uitdrukkingen alleen die combinaties geteld die als zodanig in de woordenboeken voorkomen, terwijl voor Altenberg elke woordcombinatie die vaker dan twee keer voorkwam een 'recurrent word-combination' was.

Een verwante vraag is die naar het aantal ELU's dat de taalgebruiker kent, actief dan wel passief. Jackendoff (1997:157) schat dat een gemiddelde spreker van het Engels ten minste vijfentwintigduizend vaste uitdrukkingen kent (vergelijk ook Weinreich 1969), terwijl Mel'ĉuk beargumenteert dat het 'phrasal lexicon' een orde van grootte omvangrijker moet zijn dan het woordenlexicon (Kuiper 2004).

Hoe het ook zij, de conclusie van Jackendoff (1997:157) dat '[t]here are too many idioms and other fixed expressions for us to simply disregard them as phenomena "on the margin of language"' kunnen we alleen maar onderschrijven.

Naast deze kwantitatieve vragen over ELU's is ook een kwalitatieve benadering niet oninteressant. Het is vooralsnog een open vraag wat de configuraties zijn, waar we ELU-achtige fenomenen kunnen aantreffen. Elders is betoogd dat collocationele effecten overal kunnen worden aangetroffen (Van der Wouden 1992). Onder het woordniveau functioneert *spin* als een intensiveerder in *spinnijdig*, maar *spin* kan niet gebruikt worden met de meeste andere bijvoeglijke naamwoorden (vergelijk **spin*(*ne*)*sterk* met *beresterk*). Ook tussen ELU's vinden we collocationele effecten: de partikelcombinatie *maar eens* (Van der Wouden 2002b) heeft een sterke voorkeur voor zwakke directieve constructies (Vismans 1994), als in *laat maar eens wat zien* en *ik zou maar eens naar bed gaan*. Booij (2002) laat zien dat constructies ook een rol kunnen spelen binnen inflectionele paradigma's: sommige werkwoordstijden worden immers uitgedrukt door morfologisch complexe woorden (*stierf, speelde*), andere door constructies (dus ELU's) (*is . . . gestorven, heeft . . . gespeeld*).

Op grond van het bovenstaande lijkt de vraag dan bijna eerder te moeten zijn: zijn er ook constructies en configuraties waarin we géén ELU's vinden? Ook Everaert (2003) stelt die vraag, en hij stelt (p. 27): 'De observatie is dat een idioom [de term die hij hanteert

voor wat wij hier ELU's noemen] altijd een woordgroep is en dat in zo'n woordgroep de kern, het hoofd, altijd deel uitmaakt van het idioom.' Vervolgens postuleert hij onder meer dat de structuren in (1) geen idiomen kunnen zijn (dat wat onderdeel van het idioom zou zijn, is gecursiveerd):

(1) a. [$_{\text{ZIN}}$ *Onderwerp* [$_{\text{GEZEGDE}}$ Werkwoord *Lijdend Voorwerp*]]
b. [$_{\text{ZIN}}$ *Onderwerp* [$_{\text{GEZEGDE}}$ Werkwoord *Lijdend Voorwerp*]]

Een voorbeeld van een ELU met de volgens Everaert onmogelijke structuur in (1a) is evenwel gemakkelijk te vinden:

(2) a. *Joost mag weten* waar hij dat bespottelijke idee vandaan heeft.
b. Waaraan hij dacht dat *mag Joost weten*. (fn001472)

Zo op het oog is in deze structuur ieder soort afhankelijke vraag als complement acceptabel, en een pro-form (als in (2b)) is in bepaalde gevallen ook mogelijk.[2] Dat betekent dat de vraag of er ook constructies en configuraties zijn waarin we géén ELU's vinden, nog niet afdoende beantwoord is.

2.2 Opzet van het onderzoek

Hoewel het primaat dus bij de gesproken taal ligt, houdt de grammaticale traditie zich voornamelijk bezig met de schrijftaal, met op zijn hoogst marginale aandacht voor de eigenaardigheden van de spreektaal (Haeseryn et al. 1997, De Vries 2001). Er is naar verhouding buitengewoon weinig onderzoek verricht naar de eigenschappen van de gesproken variëteiten van het Nederlands.[3] In het VIDI-project *Dutch as a construction language*, gefinancierd door NWO en de Universiteit Leiden, proberen we daar iets aan te doen.

Meer concreet staat in ons onderzoek de vraag naar de taakverdeling tussen grammatica en lexicon centraal. Die vraag operationaliseren we vooralsnog door in een corpus van spontane gesproken taal (het CGN) naar constructen en combinaties te zoeken die significant vaker voorkomen dan op grond van elementaire kansberekening verwacht mocht worden. Daarmee beperken we ons, in elk geval voorlopig, tot ELU's die op een statistische manier gedefinieerd kunnen worden, en die voldoende vaak voorkomen.

Standaard-, computationele technieken voor collocatie-onderzoek (zie bijvoorbeeld Manning & Schütze 1999) zijn zeer bruikbaar voor het vinden van bepaalde soorten ELU's. Eén voorbeeld: een programma zoals Wordsmith Tools vindt in een mum van tijd klassieke collocatie-types zoals vaste voorzetsels bij adjectieven.[4] Een standaardvoorbeeld is *trots*, dat gewoonlijk vergezeld gaat van *op*. De tabel hieronder geeft de meest frequente

2 Van der Linden (1993) geeft al Engelse tegenvoorbeelden voor de generalisatie van Everaert, zoals *A little bird told me that* . . . en *Rumour has it that* . . . .

3 Vergelijk overigens het gestelde in noot 1.

4 http://www.lexically.net/wordsmith/index.html

tweewoordclusters met *trots* in het Nederlandse deel van het CGN. De collocatie waarnaar we op zoek waren staat bovenaan:[5]

(3)

Wordsmith-clusters met *trots*		
N	Cluster	Frequentie
1	*trots op*	75
2	*heel trots*	15
3	*trots en*	7

Deze methoden brengen ons evenwel ook op het spoor van relaties tussen lexicale elementen die niet in de literatuur te vinden zijn. Op precies dezelfde manier konden we namelijk bijvoorbeeld laten zien (Van der Wouden 2002a, Poss & Van der Wouden 2005) dat het complexe focuspartikel *niet eens* (Van der Wouden 1995, en zie Goemaat 2006 voor details en nuances) opvallend vaak wordt aangetroffen in het gezelschap van

(4) andere partikels, met name *nog* en *meer*:
  a. en qua leeftijd uhm maakt het ook nog niet eens zo gek veel uit (fn000082)
  b. de man luistert niet eens meer (fn001546)

(5) het (hoogfrequente) inhoudswerkwoord *weten*:
  a. 'k weet niet eens wie Judith Bosch is (fn000687)
  b. die weten niet eens waar Nederland ligt (fn000094)

(6) modale hulpwerkwoorden, en dan vooral *kunnen*:[6]
  a. dus ik kan niet eens werken als ik zou willen (fn008001)
  b. kunnen nog niet eens hun naam en adres schrijven (fn000209)

Aan de andere kant wordt ook al snel duidelijk dat er heel wat ELU's niet gevonden zullen worden met deze techniek. Een voorbeeld van een té zeldzame constructie is het tamelijk onparlementaire *de klere krijgen*, dat wél gewoon in Van Dale staat (Den Boon et al. 2005 s.v. *klere*) en dat door veel moedertaalgebruikers ongetwijfeld moeiteloos herkend wordt, maar dat slechts één keer in het hele CGN voorkomt (*maar toen dacht ik ja krijg de klere*

5 De CGN-data zijn opgeslagen in duizenden kleine bestandjes. Dat is handig voor de bijgeleverde zoekprogrammatuur (COREX), maar niet zo handig voor het uitvoeren van het soort tellingen dat wij graag willen uitvoeren. Daarom hebben we een paar grote bestanden samengesteld met alle orthografische transcripties bij elkaar. Omdat er aanzienlijke verschillen bestaan tussen het Nederlands van Nederland en het Nederlands van België, en we ons hier niet over die verschillen wil bekommeren, beperken we ons hier en in het vervolg van dit artikel tot het Nederlandse deel van het CGN, dat wil zeggen, alle fn-bestandjes.

6 Dit collocationele effect lijkt bovendien beperkt te zijn tot één lezing of gebruik van *kunnen*, namelijk de dynamische, waar het werkwoord 'in staat zijn tot' of 'beschikken over de vaardigheid van' betekent. Dat *niet eens* gevoelig zou zijn voor het onderscheid tussen dynamische en andere modaliteit mag op het eerste gezicht verrassend zijn, maar precies hetzelfde geldt voor bepaalde lezingen van bijwoorden als *moeilijk* (Van der Wouden 1995). In de literatuur vindt men overigens verschillende opvattingen over de precieze definitie van dynamische modaliteit (vergelijk Nuyts et al. 2007).

(*fn000562*)). Bij de keuze voor een groter corpus, bijvoorbeeld als je het internet als corpus gebruikt (Van Oostendorp & Van der Wouden 1998; Kilgariff & Grefenstette 2003), vind je gemakkelijk en snel voldoende bewijsmateriaal voor dit soort tamelijk zeldzame vaste uitdrukkingen: de zoekmachine Google vond op 26 februari 2007 814 internetpagina's met het rijtje *krijg de klere* en 456 met *de klere krijgen*.

Overigens blijken voor sommige vragen alle bestaande corpora te klein. Hoogvliet (1903) wees er al op dat partikels (hij sprak zelf van 'invoegselwoordjes') regelmatig in groepjes voorkomen waarvan de volgorde zo goed als vastligt (Hulshof 1980; Van der Wouden 2002b) en hij construeerde zelf het volgende spectaculaire voorbeeld:

(7) Geef de boeken **dan nu toch maar 'es even** hier

De meeste moedertaalsprekers van het Nederlands zullen het met Hoogvliet en mij eens zijn dat dit een grammaticale Nederlandse zin is en dat er aan de volgorde van de zes partikels (*dan nu toch maar 'es even*) weinig te variëren valt, maar zelfs het grootste corpus, het internet, is te klein om dat te bewijzen, om de simpele reden dat de combinatie überhaupt niet blijkt voor te komen.[7]

3 Resultaten

In het vervolg presenteren we een paar min of meer opvallende resultaten van ons onderzoek. We beperken ons daarbij tot een minieme selectie. Eerst bespreken we een relatief nieuwe vaste combinatie van een bijvoeglijk en een zelfstandig naamwoord en vervolgens een aantal ELU's met een werkwoordelijk element als kern. We laten een glimp zien van het landschap van ELU's rondom het tussenwerpsel *ja*, en we besluiten onze lijst voorbeelden met een ELU onder het woordniveau.

3.1 Eerste resultaat: zinloos geweld

Zinloos geweld is een relatief nieuwe vaste combinatie. De betekenis ervan ligt niet erg vast, waardoor de uitdrukking uiterst bruikbaar is om van alles te labelen waar je tegen bent, van vandalisme tegen onschuldige bushokjes tot abortus (Burger 2004). Het CGN wijst uit dat de combinatie zeer vast is: *geweld* is, getuige de volgende tabel, het enige inhoudswoord dat met grote regelmaat in de buurt van *zinloos* aan te treffen is.[8]

7 De zoekmachine Google vindt wel bewijsplaatsen voor de partikelreeks *dan nu toch maar 'es even*, maar het gaat dan steeds om directe of indirecte citaten uit Hoogvliet.

8 'In de buurt van' gebruiken we in de zin van de default-instelling van het 'collocates'-menu van de Concordmodule van WordSmith tools, dat is in een venster van vijf woorden ter linker en vijf woorden ter rechter zijde van het gezochte woord.

(8)

Wordsmith-collocaten met *zinloos*		
N	Cluster	Frequentie
1	*zinloos* (totaal)	63
2	*is* in de buurt van *zinloos*	40
3	*geweld* in de buurt van *zinloos*	38
4	*dat* in de buurt van *zinloos*	20
5	*en* in de buurt van *zinloos*	15

Sterker nog, in meer dan de helft van alle gevallen (34 van de 63) wordt *zinloos* onmiddellijk gevolgd door *geweld*; het eerstvolgende bigram, *is zinloos*, komt tegen de verwachting in (de frequentie van *is* is namelijk vele malen hoger dan die van *geweld*) veel minder voor, namelijk slechts negen maal. Omgekeerd is de afhankelijkheid iets minder sterk, want *geweld* komt ook voor in andere contexten dan onmiddellijk na *zinloos*, maar toch staan 30 van de 266 voorkomens van *geweld* in het corpus onmiddellijk na *zinloos*. Bovendien: volgens de statistische tests van Wordsmith zijn *zinloos* en *geweld* wederzijds elkaars meest significante combinatiepartner.[9] Niet alleen semantisch is de combinatie *zinloos geweld* een cliché – erg origineel is de combinatie niet – ook statistisch zou het een cliché genoemd kunnen worden: als je *zinloos* hoort, dan kun je er tamelijk zeker van zijn dat *geweld* eraan komt of in elk geval in de buurt staat.

3.2 Tweede resultaat: zeg

Waar *zinloos geweld* ook volop in schrijftaal aan te treffen is, lijken sommige vaste combinaties met *zeg* vrijwel exclusief spreektalig. Volgens het clusteralgoritme van WordSmith Tools zijn er nogal wat tweewoordcombinaties met *zeg* met een hoge frequentie. In de tabel hieronder geven we de meest frequente:

(9)

WordSmith-bigrammen met *zeg*		
N	Cluster	Frequentie
1	*zeg maar*	3901
2	*ik zeg*	2230
3	*zeg ik*	816
4	*'k zeg*	807
5	*zeg je*	650
6	*zeg ja*	471
7	*dan zeg*	449

9 Wordsmith biedt keus uit onder meer loglikelihood, (een variant van) Mutual Information, en z-score (Manning & Schütze 1999), maar de uitkomsten zijn voor elk van deze toetsen vergelijkbaar.

Bij deze tabel past het terzijde dat de orthografische transcriptie van reductievocalen in het CGN niet erg consequent is geweest. In de tabel staat *ik zeg* op de tweede plaats, en *'k zeg* op de vierde, maar bij beluistering van een aantal gevallen van beide varianten konden we zelf meestal geen verschil horen. Een ander kras staaltje van inconsequente transcriptie vinden we bij *eens*. Als dat een bijwoord is, zoals in (10a) of een bijvoeglijk naamwoord, als in (10b), of *eens* is onderdeel van de vaste combinatie *niet eens* (10c), dan moet de volle vocaal gebruikt worden. In de meeste andere gevallen echter is de klinker sterk gereduceerd. In informele schrijftaal vindt men dat wel weergegeven in spellingen als *'s* of *'ns* of *es*, maar de officiële spellingsgids kent die varianten niet.[10] In het CGN vinden we de varianten *'s*, *eens* en *'ns* zo op het oog volkomen willekeurig naast elkaar (11a-11c).

(10) a. eens overkomt 't elke kring. (fn001326)
b. we zijn het eens dat er helderheid moet komen (fn000170)
c. die weten niet eens waar Nederland ligt. (fn000094)

(11) a. en ook van mijn kant nog 's de nadrukkelijke vraag. (fn000145)
b. later op de avond neemt de arts nog eens de persoonlijke bijzonderheden door. (fn001404)
c. want ach misschien worden ze nog 'ns wakker. (fn000005)

Zo veel varianten als bij het partikel *eens* vinden we niet bij *ik*, maar de cijfers van de tweede en de vierde rij van de tabel bij (9) zouden natuurlijk samengenomen moeten worden. Daarmee zijn *zeg maar* en *ik zeg* verreweg de meest voorkomende tweewoordcombinaties met *zeg*. De volgende paragraafjes gaan daar nader op in.

3.2.1 Zeg maar

De frequente combinatie *zeg maar* heeft bij nadere beschouwing van de data diverse gebruiksmogelijkheden. Allereerst moeten we het compositionele gebruik onderscheiden, waarvan voorbeelden gegeven zijn in (12a-12b).

(12) a. zeg maar wat je d'rvoor geeft. (fn000998)
b. Ik zeg maar wat (fn000561)

Maar fungeert in (12a) als een modaal partikel (Foolen 1993) dat de directieve kracht van de imperatiefvorm *zeg* verzwakt (Vismans 1994): *zeg wat je d'rvoor geeft* is een bevel, *zeg maar wat je d'rvoor geeft* is veeleer een uitnodiging. In (12b) daarentegen vervult *maar* de rol van scalair focuspartikel (Hoeksema & Zwarts 1991) dat uitdrukt dat de focusconstituent, in dit geval *wat*, zich relatief laag op een (contextueel bepaalde) schaal bevindt. De strekking van de uiting is, dat datgene wat gezegd wordt als niet al te belangrijk moet worden opgevat.

10 In het Groene Boekje (Woordenlijst 2005) zochten we vergeefs naar *'s* en *'ns*. Het Witte Boekje (Daniëls 2006) kent wel *'s* maar niet *'ns*.

Zeg maar vinden we ook met enige regelmaat (maar wel zo weinig dat er geen voorbeelden in het CGN te vinden waren) terug als onderdeel van grotere vaste verbindingen (ELU's) als *zeg maar nee, dan krijg je er twee* en *zeg maar dag* (*met je handje, tegen*).[11]

(13) a. Zeg maar nee, dan krijg je er twee (bekeuringen op een dag) (mailgate.dada.net/nl/nl.juridisch/msg534520.html)
b. Zeg maar dag tegen je 1000 euro. (www.geenstijl.nl/mt/archieven/016458.html)
c. Zeg maar dag met je handje, zeg maar dag, dag, dag (www.kinderliedjes.nu/0-2-jaar/zeg-maar-dag/)

In de uitdrukkingen in (13) is nog wel het gebruik uit (12a) te herkennen, maar de frequenties in internet-teksten suggereren dat we hier met zelfstandige ELU's te maken hebben; geraadpleegde moedertaalsprekers bevestigen de uitdrukkingen in (13) te kennen.

Het meest frequente gebruik van *zeg maar* zien we geïllustreerd in (14a-14b):

(14) a. dan zou je nu zeg maar honderd zeventig vierkante meter in de aanbieding hebben. (fn000321)
b. ik zie dat die aspecten zeg maar van het taalgebruik uh eigenlijk steeds uh teruglopen. (fn000097)

Dit gebruik is tamelijk recent: volgens Stroop (2006) wordt de combinatie in het tijdschrift *Onze Taal* voor het eerst gesignaleerd in 2000. Inmiddels heeft *zeg maar* een plekje gevonden in de Van Dale (Den Boon et al. 2005): de combinatie wordt genoemd bij betekenis 4b van *zeggen* en krijgt als betekenisomschrijving 'neem maar aan, laat ons maar aannemen dat het het genoemde is': *de puberteit zeg maar*. Badinerend noemt Stroop *zeg maar* '[h]et meest gebruikte wauwelwoord van heden. *Zeg maar* valt in de categorie stopwoorden. Betekenis heeft het nu meestal niet, maar het heeft die ooit wel gehad.' Even verderop in zijn column is Stroop iets genuanceerder: '*Zeg maar* werd vroeger gebruikt als de spreker het juiste woord of de juiste omschrijving zo gauw niet kon vinden'. Deze beschrijving, die *metalinguïstisch* genoemd zou kunnen worden, lijkt nog steeds van toepassing op een zin als (14b). Het gebruik in (14a) is evenwel anders: *zeg maar* betekent daar ongeveer 'ongeveer'; dit gebruik zouden we *approximatief* kunnen noemen.[12] Interessant genoeg noemt Stroop dit gebruik niet en het wordt ook niet echt gedekt door de betekenisomschrijving in Van Dale, al is het daar wellicht een speciaal geval van.

11 Volgens www.jeugdsentimenten.net/2005/11/26/jeugdsentimentendag-wat-gaan-we-doen/ afkomstig uit een reclame voor een snoepje met de naam Nimm 2, maar een reviewer van *Nederlandse Taalkunde* is de mening toegedaan dat de slogan *Zeg maar nee, dan krijg je er twee* stamt uit het televisiedebuut van Bart de Graaff (later BNN) in een reclame voor chocolade-koekjes.

12 Dianne Landsmeer voltooide in juli 2007 aan de Leidse opleiding Nederlandse taal en cultuur een Bachelorscriptie over *zeg maar*. Ze laat onder andere zien dat approximatief *zeg maar* andere syntactische eigenschappen heeft dan metalinguïstisch *zeg maar*: het metalinguïstische kan op veel meer posities in de uiting voorkomen dan het approximatieve, maar het approximatieve dringt gemakkelijker binnen in een constituent.

3.2.2 Ik zeg

De combinatie *ik zeg* vinden we vooral bij de weergave van dialogen. Het voorbeeld in (15a) laat zien dat *ik zeg* ook een tegenhanger heeft in de derde persoon. Van Esch (2007) stelt voor om *ik zeg* op te vatten als een quotatief-markeerder, dat wil zeggen, een grammaticaal element dat aanduidt dat er iets geciteerd wordt. De presensvorm, zo betoogt ze, geeft het verslag een levendigheid die verleden tijden missen.

(15) a. ik zeg maar wie bent u dan. hij zegt ik zat boven. (fn007577)
b. 'k zei jongens ik zeg ruim alles maar op. (fn007006)
c. ja nee dan brengen we je wel ik zeg maar ja daar gaat 't niet om. (fn000459.89)
d. ik zei ook tegen Jeroen ik zeg moet je even kijken. (fn006995.79)

In het recente verleden is voor het Engelse *like* en het Nederlandse *van* beargumenteerd dat die ook een quotatief gebruik kennen (vergelijk Foolen et al. 2006 en de daar aangehaalde literatuur), zie (16).

(16) de beroemdste exponent is de uh uh Engelse getaltheoreticus uh Hardy die uh zei van nou ik heb nooit iets nuttigs gedaan en daar ben ik eigenlijk nog trots op ook. (fn000063)

Het hoeft geen betoog dat de ene quotatief de andere niet is: *ik zeg* is bijvoorbeeld veel meer gericht op het subject dan *van*, dat wel is opgevat als de orale tegenhanger van de dubbele punt. Waarschijnlijk komt dat verschil in elk geval ten dele doordat *ik zeg* niet alleen bestaat uit een *zeggen*-werkwoord, maar ook een subject-markeerder *ik* bevat.

De voorbeelden (15b) en (15d) laten zien dat *ik zeg* zodanig gegrammaticaliseerd is dat de combinatie zonder bezwaar in dezelfde zin als een compositioneel *ik zei* gebruikt kan worden: kennelijk wordt het presens van *ik zeg* niet meer als zodanig ervaren, zodat er geen conflict gevoeld wordt met het preteritum van *ik zei*.[13]

3.3 Derde resultaat: weten

Weten is een van de frequentere inhoudswerkwoorden van het Nederlands.[14] *Weten* is bovendien een interessant werkwoord vanuit ELU-perspectief. Hierboven kwamen al terloops de vaste uitdrukking *Joost mag (het) weten* (2a) en de voorkeur van de partikelcombinatie *niet eens* voor het werkwoord aan de orde. Hieronder geven we een staatje van de meest prominente pentagrammen met *weten* in het Nederlandse deel van CGN:

13 Vergelijkbaar in de zin van niet-subjectief en afgeleid van een *zeggen*-werkwoord is de quotatief *saying* in de King James-vertaling (*And the LORD God commanded the man, saying, Of every tree of the garden thou mayest freely eat* (Gen 2, 16)) en de bron daarvan, *dicens*, in de Vulgaat (*praecepitque Dominus Deus homini dicens: "Ex omni ligno paradisi comede; de ligno autem scientiae boni et mali ne comedas; in quocumque enim die comederis ex eo, morte morieris"*).

14 Dat neemt niet weg dat *weten* in bepaalde soorten zinnen trekken van een functiewoord vertoont, bijvoorbeeld in *daar wist die man geen informatie over te geven* (fn000750) waar de bijdrage van *weten* vooral modaal is.

(17)

Cluster	Frequentie
ik zou 't niet weten	20
'k zou 't niet weten	12
dat wil je niet weten	10
dat moet je zelf weten	8
zou 't echt niet weten	7
je moet 't zelf weten	6
ik zou 't echt niet	6
zou 't niet weten ja	5
ik zou echt niet weten	5
ja dat moet je zelf	5
ik zou het niet weten	5

Als we even abstraheren van de verschillen tussen *ik* en *'k* en tussen *het* en *'t*, dan komen *ik zou het niet weten* en varianten maar liefst 38 keer voor. Is dit een ELU? Waarschijnlijk wel. Puur kwantitatief is de combinatie veel frequenter dan statistisch zou mogen worden verwacht.[15] Bovendien is de combinatie niet compositioneel: wat is de bijdrage van *zou*, met andere woorden, wat is het verschil in betekenis met *ik weet het niet*?

De volgende interessante combinatie in de tabel in (17) is *dat wil je niet weten*. De uitdrukking heeft inmiddels het woordenboek bereikt: Van Dale (Den Boon et al. 2005 s.v. *weten* 1, 1) beschrijft de betekenis als 'het is onvoorstelbaar', een betekenis die niet af te leiden is uit de betekenissen van de samenstellende delen. Het volgende voorbeeld (18) illustreert dit: de spreker bedoelt dat het geduld van de 'hij' heel groot is.

(18) die heeft een geduld dat wil je niet weten. (fn006903)

De laatste ELU uit de tabel in (17) heeft ook verschillende vormen:

(19) a. dat moet je zelf weten
b. je moet het zelf weten

Dit gebruik vinden we niet terug in het woordenboekartikel over *weten*. Toch zou het daarin zeker op zijn plaats zijn, omdat *weten* hier niet 'kennis hebben van of omtrent, bekend zijn met' betekent, maar, verrassend genoeg, veeleer zoiets als 'beslissen'. Bovendien lijkt deze betekenis alleen mogelijk te zijn als aan twee voorwaarden voldaan is: *weten*

15 Het Nederlandse deel van het CGN bevat met mijn manier van tellen 5663528 woorden; daarvan 146968 keer *ik* of *'k*, 10201 keer *zou*, 97166 + 42717 = 139883 keer *'t* of *het*, 70221 keer *niet*, 2641 keer weten. De kans om de combinatie in dit corpus aan te treffen is dan 5663528 * (38892501812165017084584 / 56635285), dat is ongeveer 1 op 26500. Met andere woorden, de kans op 38 voorkomens van *ik zou het niet weten* is ongeveer 1 op een miljoen.

moet vergezeld worden door één of ander soort modale operator, al hoeft dat niet per se *moeten* te zijn, en het subject moet focus hebben, hetzij door toevoeging van *zelf*, of door een sterke vorm van het pronomen zelf:

(20) a. je mag toch zelf weten wanneer je komt. (fn008006)
b. ze weet hopelijk zelf wel waar ze mee begint. (www.cavaletti.nl)
c. ja dat moeten zij weten maar ik dan vier ik m'n verjaardag dus niet meer voor hun (fn008046)

Deze ELU behoeft zeker nader onderzoek.

3.4 Vierde resultaat: zijnde

In deze paragraaf behandelen we een ELU waarbij de resultaten van taaldescriptie botsen, of kunnen botsen, met de regels zoals die worden geformuleerd door de normatieve taalkunde. Laten we eens kijken naar de omgevingen waarin we het verbogen tegenwoordig deelwoord *zijnde* aantreffen. Het woord dat verreweg het meest in de omgeving van *zijnde* aangetroffen wordt, is *als*. Hieronder staan een paar typische corpuszinnen met *zijnde*.

(21) a. wij zitten ook wel 'ns fout of als docent zijnde goddorie je maakt ook fouten dagelijks nog (fn000108)
b. in deze vakantiemaanden komt het namelijk nogal 'ns voor dat Fransen hun eigen hond als zijnde een langs de weg gevonden exemplaar in het asiel afgeven (fn001159)

De normatieve taalkunde keert zich, zoals bekend, tegen dit soort zinnen: bepalingen van gesteldheid van dit type moeten worden geconstrueerd hetzij met *als* (*als muzikant*), of met *zijnde* (*muzikant zijnde*), maar niet met allebei. Het argument gaat in de formulering van de taaladviesdienst Onze Taal op het internet als volgt:[16]

(22) *Als muzikant zijnde* is een contaminatie (verhaspeling) van twee constructies: *als muzikant* en *muzikant zijnde*. Omdat die laatste formulering in modern Nederlands erg ouderwets overkomt, kun je het best zeggen: 'Als muzikant kan ik dit niet waarderen.'

De formulering is al voorzichtig – er wordt niet met zoveel woorden gezegd dat de uitdrukking *als muzikant zijnde* geheel uit den boze is – maar het argument is in het geheel niet dwingend. Immers, in een advies op dezelfde site gewijd aan *dat klopt als een bus* wordt gesteld dat ook die uitdrukking 'is ontstaan als contaminatie (onjuiste samentrekking)', en wel van *kloppen* en *sluiten als een bus*. Op basis van de autoriteit van Van Dale en het aantal voorkomens van *dat klopt als een bus* op het internet wordt daar echter geconcludeerd 'Er bestaat dan ook geen enkel bezwaar meer tegen het gebruik ervan.'

Op basis van datzelfde soort corpusdata heeft *als … zijnde* echter minstens even goede rechten. Anders gezegd, als we de corpusdata serieus nemen, dan moeten we zeggen:

16 www.onzetaal.nl

tegenwoordig is *als zijnde* een van de manieren om in het Nederlands een bepaling van gesteldheid van de eerste soort (Den Hertog 1974-1979) oftewel een predicatieve toevoeging te markeren. Als vrije bepalingen met een verbogen vorm van het tegenwoordig deelwoord (*alles in aanmerking nemende, ijs en weder dienende*) al ooit iets anders geweest zijn dan een schrijftaalconstructie naar analogie met het Latijn, dan kunnen we in elk geval voor de huidige taalsituatie vaststellen dat die constructie niet meer productief bestaat. Het meest voor de hand liggende alternatief is met *als*. *Als* heeft in het moderne Nederlands echter nogal wat functies, en is daardoor zwaar belast. Het kan dan ook vrij lang duren voor je weet wat de functie is van een zinsdeel dat met *als* begint, wat het voor de hoorder lastig maakt om uit te maken wat de spreker kan bedoelen. We kunnen de opkomst van de constructie *als … zijnde* dan ook opvatten als een geval van specialisering door uitbreiding, zoals we dat bijvoorbeeld ook zien bij de voorzetsels (de opmars van *vanaf, vanuit* enz. die een aantal van de functies van *van* overgenomen hebben (Van der Horst & Van der Horst 1999)) en bij de partikels (Hoeksema 2002).

3.5 Vijfde resultaat: ja

Schrijftaal is anders dan spreektaal. In schrijftaal is *de* het meest frequente woord (Uit den Boogaart 1975), maar in spreektaal, zoals gerepresenteerd door het CGN, is *ja* veruit het meest frequent. De volgende tabel illustreert dat:

(23)

***Meest frequente woorden in het Nederlandse deel van CGN (N = 5863159)* (tellingen WordSmith tools)**			
N	Woord	Frequentie	%
1	*ja*	192456	3,28
2	*de*	159545	2,72
3	*dat*	156005	2,66
4	*uh*	154252	2,63
5	*en*	141063	2,41
6	*ik*	125162	2,13
7	*een*	111628	1,90
8	*'t*	97895	1,67
9	*je*	90155	1,54
10	*die*	88479	1,51
		1162388	19,83

Merk op dat slechts tien woord-'types' bijna 20% van alle 'tokens' dekken. Overigens dienen deze cijfers met omzichtigheid te worden gehanteerd: de frequenties blijken zeer uiteen te lopen tussen de verschillende subcorpora. De tabel in (24) laat zien dat *ja* alleen het meest frequent is in de onderdelen dialoog en telefoongesprek – in de subcorpora radio,

voorgelezen tekst, debat en televisie staat *de*, net als in Uit den Boogaart (1975), bovenaan (de kolom ***#ja*** geeft de rang van *ja* in de desbetreffende frequentielijst):[17]

(24)

subcorpus	***# ja***	**%**	***frequentst***	**%**
dialoog (N=1815735)	1	4,52	*ja*	4,52
telefoon (N=1286962)	1	5,96	*ja*	5,96
radio (N=1001366)	13	1,16	*de*	4,57
voorgelezen tekst (N=558543)	135	0,09	*de*	4,78
onderwijs (N=307876)	8	2,00	*dat*	2,91
interview (N=264621)	4	2,86	*uh*	3,98
(politiek) debat (N=220094)	35	0,47	*de*	5,02
televisie (N=196865)	17	0,92	*de*	4,02
gesimuleerd zakengesprek (N=140349)	3	3,56	*uh*	4,33
presentatie (N=63492)	97	0,13	*uh*	3,85

Uit deze tabel blijkt eerst en vooral dat het feit dat *ja* het meest frequente woord in het CGN is, vooral veroorzaakt wordt doordat het het frequentste woord is in de grootste subcorpora van het corpus.[18] En we kunnen speculeren dat *ja* juist het frequentste woord is in deze subcorpora omdat het een typisch dialoogwoord is. Dat neemt niet weg dat *ja* ook wel degelijk middenin monologische fragmenten kan worden aangetroffen. De volgende tekst van een televisieverslaggever illustreert dit:

(25) maar eigenlijk v valt in 't hele scenario de verdediging van 't binnenland verdediging tegen aanvallen binnen Amerika uh ja valt eigenlijk buiten de taakstelling van bijvoorbeeld uh 't Pentagon of van van 't Amerikaanse leger. (fn007780)

In deze passage wordt *ja* (of misschien *uh ja*) gebruikt om een aarzelingspauze te vullen en om een herstart te maken, waarschijnlijk om een syntactische constructie te herstellen die te complex geworden is en dreigt te ontsporen (vergelijk Levelt 1989). Maar veel vaker blijkt *ja* inderdaad een dialoogwoord te zijn. Twee voorbeelden in het volgende fragment uit hetzelfde televisieprogramma:

(26) a. bedenk ook wel dat Amerika nooit een oorlog op eigen grondgebied wat dat betreft gevochten heeft
b. ja

17 De kleine sub-corpora die 'eucharist', 'meeting' en 'panel' heten, zijn buiten beschouwing gelaten.

18 De vraag of de kwantitatieve verdeling over teksttypes in het CGN op de een of andere manier representatief is voor de verdeling in al het gesproken Nederlands is vermoedelijk niet te beantwoorden, behalve bij stipulatie.

c. dus dat is een enorm verschil met bijvoorbeeld hoe we d'r in Europa over denken en hoe onze eigen defensie is opgebouwd
d. ja (fn007780)

Aan het eind van het eerste deel van zijn bijdrage (26a) lijkt de verslaggever een beetje vast te lopen, maar het *ja* van de presentator (26b) nodigt hem uit om door te gaan. Na het tweede deel (26c) maakt de verslaggever een pauze, waarmee hij aangeeft dat de presentator in de studio de beurt kan overnemen; met diens *ja* (26d) geeft hij als het ware aan de bijdrage van de verslaggever in ontvangst te nemen.

Laten we nu snel overgaan naar ELU's met *ja*. Ruimtegebrek verbiedt me overigens meer dan het topje van de ijsberg te bespreken. Volgens Wordsmith Tools behoren de clusters in de volgende tabel tot de meest frequente die *ja* bevatten (de op dit moment minder relevante zijn weggelaten):

(27)

Frequente clusters met ja		
cluster	*rangorde*	*#N*
ja ja	1	181347
ja ja ja	2	84696
ja ja ja ja	3	34430
oh ja	5	18512
ja dat	6	17721
ja maar	7	16343

De meeste clusters in deze tabel kunnen worden aangemerkt als ELU's. *Ja ja*, bijvoorbeeld, fungeert dikwijls als een hoorderssignaal (een speciaal type *discourse marker*, vergelijk bijvoorbeeld Schiffrin 1987) en kan, afhankelijk van de intonatie, onder meer instemming of twijfel uitdrukken. Het volgende conversatiefragment (lichtelijk aangepast ten behoeve van de presentatie) demonstreert dit:

(28) A: ja die uh die tante Hennie zit een beetje te miepen hè.
B: ja die wilde zich laten euthanaseren of niet?
A: ja ja die wil euthanasie tegen die tijd. (fn008113)

De volgende ELU in de tabel in (27) is *ja ja ja*. Het is niet altijd goed uit te maken wat het verschil is met tweevoudig *ja ja*:

(29) A: zijn zij ook verstoken van radio en TV?
B: ja ja ja
A: oh dat wist ik niet. (fn000274)

In sommige gevallen geeft de intonatie uitsluitsel dat we te maken hebben met een compositionele combinatie van een enkel *ja* plus een dubbel *ja ja*.

(30) A: en neemt ze ook nog wol mee naar Spanje om te breien?
B: ja. ja ja (fn000266)

In dit geval is er zelfs zo'n duidelijke intonationele breuk dat er in de orthografische representatie een punt is toegevoegd, maar we kunnen er niet op vertrouwen dat dat altijd een betrouwbare aanwijzing zal zijn. Hier lijkt het eerste enkelvoudige *ja* bedoeld als eenvoudig positief antwoord op de voorafgaande vraag, terwijl het daaropvolgende dubbele *ja ja* bedoeld lijkt te zijn om eventuele tegenwerpingen in de kiem te smoren – in die zin is het vergelijkbaar met zoiets als *nee echt*.

Als alleen het toeval de distributie van lexicale elementen zou bepalen, dan is het niet heel erg verbazend dat er twee exemplaren van het meest frequente element *ja* naast elkaar worden aangetroffen. Puur op grond van de kansberekening zou *ja ja* de meest voorkomende ELU moeten zijn. Veel verrassender is de hoge positie van de volgende combinatie in de tabel in (27), *oh ja*.[19] Op basis van de bestudering van een aantal voorkomens van *oh ja* zouden we waarschijnlijk ten minste drie verschillende ELU's *oh ja* moeten onderscheiden. In het eerste gebruik geeft het een soort bevestiging:

(31) A: ik heb we uh van die kleine haakjes met vier kleine spijkertjes die gaan maar heel die gaan maar heel uh uh zijn eigenlijk maar heel uh heel ondiep.
B: oh ja dat zei je ja. (fn000253)

In andere gevallen draagt *oh ja* een vraagintonatie, en betekent het ongeveer *echt waar?*:

(32) A: jij hebt hier een keer hele tijden geleden je toilettas laten staan.
B: oh ja? welke dan (fn000719)

En ten derde kan *oh ja* ook gebruikt worden om een nieuw gesprksthema aan te snijden:

(33) uhm oh ja ik heb wat vergeten. die gebakken geitenkaas die kunnen we hier bijdoen warm. (fn008411)

Toch moeten we niet te snel zijn met concluderen dat alle hoogfrequente combinaties met *ja* dan wel opgeslagen zullen zijn in het lexicon. Zo lijkt de combinatie *ja dat* geen interessante ELU te zijn. Veeleer lijkt het grote aantal voorkomens van de combinatie resultaat te zijn van de interactie van een aantal eigenaardigheden van de Nederlandse grammatica enerzijds en van de organisatie van Nederlandse gesprekken anderzijds. Zo is het woordje *ja* buitengewoon geschikt om ter zelfder tijd instemming te betuigen en de beurt over te nemen in het gesprek (Mönnink 1988). Verder is het een goede gewoonte (Onrust et al. 1993) om je beurt te beginnen met te verwijzen naar een onderwerp dat al bekend is in

19 In de orthografische representatie van het CGN is er helaas voor gekozen om geen onderscheid te maken tussen *o* 'uitroep van verrukking, bewondering, verwondering, verbazing; van vreugde of voldoening' enz. (Den Boon et al. 2005:o2) en *oh* 'tussenw. ter uitdrukking van spijt, droefheid (vaak in gerekte vorm)' enz. (Den Boon et al. 2005:*oh*).

de conversatie en om dat dan te doen met een aanwijzend voornaamwoord zoals *dit* of *dat* op de eerste zinsplaats.

(34) a. ja dit moet nog gedaan worden (fn000265)
b. ja dat klopt (o.a. fn000088)

3.6 Zesde en laatste resultaat: kei

In het begin van dit artikel noemden we al *spinnijdig* als een voorbeeld van een ELU onder het woordniveau. De verbinding is vast: *spin* kan (in deze functie) alleen gecombineerd worden met *nijdig,* en niet met al of niet semantisch verwante andere bijvoeglijke naamwoorden, getuige de ongrammaticaliteit van **spinboos* en **spinlief* (Fletcher 1980, Schutz 2007). *Spin* fungeert hier als een versterker van de betekenis: *spinnijdig* is erger dan gewoon *nijdig*. De betekenis van het geheel volgt daarbij niet uit de som van de betekenissen van de delen: *spin* betekent in andere gevallen nooit 'erg'.

Keihard nu is een formatie die vergelijkbaar is met *spinnijdig*: *kei* fungeert net zoals *spin* als een versterkend voorvoegsel. Een verschil is dat hier de metaforische herkomst 'zo hard als een kei' wat duidelijker is dan bij *spinnijdig*: 'hard' is wel een typische eigenschap van keien, maar voor de meeste taalgebruikers is 'nijdig' niet iets wat typisch bij spinnen hoort.

Een min of meer recente ontwikkeling – Van Dale (Den Boon et al. 2005) noemt het 'jeugdtaal' – is dat *kei-* wordt geherinterpreteerd als een versterkend prefix.[20] De voorbeelden liggen in het CGN voor het oprapen:

(35) a. ah ik vind dat **keizielig** voor jou. (fn006921)
b. en da 's uh **keiveel** geld waard. (fn000846)
c. 'k heb wel 'ns bij mensen in de auto gezeten die **keiveel** gezopen hebben. (fn000859)
d. ik zei al dan ben ik **keikapot**. (fn008263)
e. maar ja dat vond 'k echt **keikut** voor d'r. (fn006922)
f. dat valt gewoon **kei-erg** op weet je wel. (fn006800)

Eveneens volgens Van Dale heeft *kei* hier als synoniemen *loei-* en *mega-,* en ook daarvan bevat het CGN voorbeelden:

(36) a. alleen die pannen die ze daar hadden die waren **loeizwaar**. (fn006898)
b. en die jongen was **loei-enthousiast** en wist je overal in mee te slepen in alles wat ie hier las. (fn000139)
(37) a. maar kijk die bedrijven zijn ook zo **megagroot** (n000410)
b. niks en niemand bonkt zo heftig zo **megacool** als de griezelgrungeband van Willem En De Weerwolven. (fn001326)

20 Het is intuïtief niet geheel implausibel dat de eerste stap van de betekenisuitbreiding mogelijk was vanwege de verschillende betekenissen van *hard*: van 'niet of moeilijk samen te drukken' (*keiharde boter*) naar 'luid' (*keiharde geluiden*).

Nadere kwalitatieve en kwantitatieve beschouwing van de data brengt evenwel tenminste drie verschillen tussen de vermeende synoniemen aan het licht:

- Frequentie: In de eerste plaats komen met *kei-* geprefigeerde woorden vaker in het corpus voor dan combinaties met *loei-* of *mega-*. In het Nederlandse deel van het CGN vonden we 247 voorkomens van *kei* als eerste lid (daarvan maar liefst 120 keer *keihard*), tegenover 50 *mega*-afleidingen (waarvan 16 gelexicaliseerde of ontleende woorden als *megapixel, megawatt* en *megaton*) en slechts 7 combinaties met *loei-* als linkerlid.
- Valentie: anders dan *kei-* en *loei-* wordt *mega-* ook aangetroffen voor zelfstandige naamwoorden:

(38) a. Whitney tekent het **megacontract** met platenmaatschappij Arista waarschijnlijk komende maandag. (fn006047)
b. maar v die hele grote keiharde **mega-attracties** ik denk niet dat wij in die race mee willen doen. (fn007693)

- Productiviteit: Booij & Van Santen (1998, 73) bespreken de hypothese dat het aantal hapaxen (woorden die maar één keer voorkomen in een corpus) in een populatie van morfologische vormsels een maat is voor de productiviteit van het achterliggende proces (vergelijk ook Baayen 1989). Zij geven het voorbeeld van *heid*-afleidingen in het corpus-Eindhoven: ze vonden er 466, waarvan er maar liefst 256 slechts één keer voorkwamen. Bij *kei-* vonden we 23 hapaxen, dat is een procent of 10, tegenover 29 bij *mega-* (meer dan 50% van de 50) en 3 bij *loei*. Het corpus is te klein om hier definitieve uitspraken over te doen, maar volgens deze metriek zijn prefigering met *mega-* en *loei-* productiever dan prefigering met *kei-*.

Deze gegevens bij elkaar lijken erop te wijzen dat *kei*-prefigering (nog?) niet bijzonder productief is. *Keihard* is nog steeds verreweg de meest voorkomende vorm met *kei-* als linkerlid. Dat woord fungeert kennelijk als het voorbeeld waarnaar via analogie nieuwvormingen met *kei-* worden gemaakt. Met andere woorden, *keihard* is door die analogiewerking bezig zijn status als ELU (onder het woordniveau) te verliezen: door het ontstaan van een productief proces van *kei*-prefigering wordt *keihard* minder ondoorzichtig. Dat proces is echter nog niet voltooid, maar *loei-* en *mega-* zijn in dit opzicht wel al verder. Een ander verschil is dat de valentie van *mega-* anders is. Het is al met al dus een simplificatie om *kei-*, *loei-* en *mega-* op te vatten als synoniemen.

4 Afsluitende opmerkingen: resultaten en perspectieven

In het project *Dutch as a construction language* doen we empirisch onderzoek naar de taakverdeling tussen grammatica en lexicon in spontaan, gesproken Nederlands. In dit artikel hebben we aandacht besteed aan de achtergronden van het project, en hebben we een paar aansprekende resultaten gepresenteerd. In het begin hebben we een aantal onderzoeksvragen geformuleerd. Op de kwantitatieve – wat is het aandeel van *extended lexical units* (ELU's) in spontane spraak, en hoeveel ELU's heeft het Nederlands? – kunnen we nog geen antwoord geven, al is wel duidelijk geworden dat het om grote getallen gaat.

Zo constateerden we in de paragraaf over *ja* dat dat tussenwerpsel – het meest frequente woord in de grootste subcorpora van het CGN – zeer vaak in vaste combinaties met een eigen betekenis en eigen gebruiksmogelijkheden voorkomt. De voornaamste conclusie uit de andere paragrafen lijkt te moeten zijn, dat je overal ELU's vindt waar je maar kijkt.

Ook op de vraag of er constructies of constructietypen zijn waarin geen ELU's voorkomen, kunnen we nog geen antwoord geven. We hebben met Booij geconstateerd dat ELU's een rol kunnen hebben in morfologische paradigma's, en we hebben tegenvoorbeelden aangedragen voor claims van Everaert met betrekking tot de onmogelijkheid van bepaalde ELU's. De ELU's die uit ons onderzoek naar boven kwamen waren zeer divers: van nominale structuren zoals *zinloos geweld* tot partikelachtige combinaties zoals *ik zeg* en *zeg maar*.

De hier besproken methode is niet zonder problemen, zoals we in het begin al hebben aangeduid: sommige ELU's zijn te zeldzaam om in een corpus als het CGN statistisch op te vallen. Een ander bezwaar is, dat de methode vooralsnog zeer op woorden gericht is. Abstractere ELU's, zoals NPN (*kolom voor kolom werken we even de studiewijzer door* (fn009060), vergelijk Poss 2008), ontsnappen zo aan onze aandacht. En weer andere ELU's zijn zowel te abstract als te zeldzaam: elders (Van der Wouden 2007c) hebben we gerapporteerd over een internetstudie naar de prenominale modale infinitief (*Excelsior is de te kloppen ploeg, niet mis te verstane bewoordingen*). Dat we ook hier te maken hebben met sterk gelexicaliseerde constructies blijkt onder meer uit het feit dat in deze constructie *kloppen* vrijwel altijd 'verslaan' betekent (*dit is de te kloppen slagroom* is heel raar) en dat de negatie verplicht is in *niet mis te verstane bewoordingen*.

Ondanks het feit dat er – zoals bij elke wetenschappelijke methode – problematische kanten zijn aan de in dit artikel beschreven aanpak van onderzoek naar ELU's in gesproken Nederlands, blijkt die wel degelijk zijn vruchten af te werpen. Zonder het Corpus Gesproken Nederlands was dat onmogelijk geweest.

Bibliografie

Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word-combinations. In: A.P. Cowie (red.), *Phraseology. Theory, Analysis, and Applications*. Oxford: Clarendon Press, 101–22.

Baayen, H. (1989). *A corpus-based approach to morphological productivity. Statistical analysis and psycholinguistic interpretation*. Diss. Vrije Universiteit Amsterdam.

Biber, D. (1988). *Variation across speech and writing*. Cambridge [etc.]: Cambridge University Press.

Bolinger, D. & D. Sears (1981). *Aspects of Language*. New York: Harcourt Brace Jovanovitch, 3rd edition. (1st edition 1968).

Booij, G. (2002). Constructional idioms, morphology, and the Dutch lexicon. *Journal of Germanic Linguistics* 14, 301–329.

Booij, G. & A. van Santen (1998). *Morfologie. De woordstructuur van het Nederlands*. Tweede, herziene en verbeterde druk. Amsterdam: Amsterdam University Press.

Boon, T. den, D. Geeraerts & N. van der Sijs (red.) (2005). *Van Dale Groot woordenboek der Nederlandse taal*. Utrecht/Antwerpen: Van Dale Lexicografie. 14e, herz. uitg.

Burger, P. (2004). *Hun hebben gelijk: ieder z'n taal*. Amsterdam: Prometheus.

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford [etc.]: Oxford University Press.

Daniëls, W. (2006). *Het witte boekje: spellinggids van het Nederlands*. Utrecht: Prisma.

Esch, C. van (2007). *Ik zeg oh ik zeg van ja ach tuurlijk. Over de vorm en de betekenis van* ik zeg *in gesproken taal*. Bachelorscriptie, Opleiding Nederlandse taal en cultuur, Leiden.

Everaert, M. (2003). *Wijzen van zeggen*. Rede Katholieke Universiteit Nijmegen.

Fillmore, C., P. Kay & M. C. O'Connor (1988). Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language* 64, 501–38.

Fletcher, W. (1980). 'Blood-hot', 'stone-good': A preliminary report on adjective-specific intensifiers in Dutch. *Leuvense Bijdragen* 69, 445–472.

Foolen, A. (1993). *De betekenis van partikels. Een dokumentatie van de stand van het onderzoek met bijzondere aandacht voor maar*. Diss. Nijmegen.

Foolen, A., I. van Alphen, E. Hoekstra, H. Lammers, H. Mazeland, & E. Pascual (2006). Het quotatieve *van*. Vorm, functie en sociolinguïstische variatie. *Toegepaste Taalwetenschap in Artikelen. Thema's en trends in de Sociolinguïstiek* 5, 137–149.

Goemaat, M. (2006). *Niet eens een fócuspartikel. Een eerste verkenning van de vaste woordcombinatie* niet eens *in het Nederlands*. Doctoraalscripie, Universiteit Leiden, Nederlandse Taal en Cultuur.

Goldberg, A. (1995). *Constructions. A Construction Grammar Approach to Argument structure*. Chicago: University of Chicago Press.

Haeseryn, W., K. Romijn, G. Geerts, J. de Rooij & M.C. van den Toorn (red.) (1997). *Algemene Nederlandse Spraakkunst.* Groningen en Deurne: Martinus Nijhoff en Wolters Plantijn. 2e, geheel herz. dr.

Hertog, C.H. den (1974-1979). *Nederlandsche spraakkunst: handleiding ten dienste van aanstaande (taal)onderwijzers*. Amsterdam: Versluys. 4e bew. dr. ingel. en bew. door H. Hulshof [1e dr. 1892-1895].

Hoeksema, J. (2002). Polarity-sensitive scalar particles in early modern and present-day Dutch: distributional differences and diachronic developments. In: Ton van der Wouden, Ad Foolen, & Piet Van de Craen (red.), *Particles. Belgian Journal of Linguistics* 16. Amsterdam: John Benjamins, 53–64.

Hoeksema, J. & F. Zwarts (1991). Some remarks on focus adverbs. *Journal of Semantics* 8, 51–70.

Hoogvliet, J.M. (1903). *Lingua: een beknopt leer- en handboek van Algemeene en Nederlandsche taalkennis, meer bepaaldelijk bestemd voor leeraren en onderwijzenden in moderne en oude talen*. Amsterdam: S.L. van Looy.

Horst, J. van der & K. van der Horst (1999). *Geschiedenis van het Nederlands in de twintigste eeuw*. Den Haag/Antwerpen: Sdu/Standaard.

Hulshof, H. (1980). Over de zg. invoegselwoordjes van Hoogvliet. *GLOT* 3, 171–181.

Jackendoff, R. (1997). *The Architecture of the Language Faculty*. Cambridge, Mass.: The MIT Press.

Jansen, F. (1981). *Syntaktische konstrukties in gesproken taal*. Diss. Leiden.

Kilgariff, A. & G. Grefenstette (2003). Introduction to the special issue on the web as corpus. *Computational Linguistics* 29, 333–347.

Kuiper, K. (1996). *Smooth talkers: the linguistic performance of auctioneers and sportscasters*. Mahwah, NJ: Lawrence Erlbaum Associates.

Kuiper, K. (2004). [bespreking van] A. Wray: Formulaic language and the lexicon. *Language* 80, 868–872.

Levelt, W.J.M. (1989). *Speaking: from intention to articulation*. Cambridge, Mass.: The MIT Press.

Linden, E.-J. van der (1993). *A categorial, computational theory of idioms*. Diss. Tilburg.

Manning, C. & H. Schütze (1999). *Foundations of statistical natural language processing*. Cambridge, Mass.: The MIT Press.

Miller, J. & R. Weinert (1998). *Spontaneous spoken speech. Syntax and Discourse*. Oxford: Clarendon.

Mönnink, J. (1988). *De organisatie van gesprekken: een pragmatische studie van minimale interaktieve taalvormen*. Diss. Nijmegen.

Nuyts, J., P. Byloo & J. Diepeveen (2007). *Mogen* en *moeten* en de relaties tussen deontische modaliteit en modus. *Nederlandse Taalkunde* 12, 153-174.

Onrust, M., A. Verhagen & R. Doeve (1993). *Formuleren*. Houten: Bohn Stafleu Van Loghum.

Oostendorp, M. van & T. van der Wouden. (1998). Corpus internet. *Nederlandse Taalkunde* 3, 347–361.

Pawley, A., & F.H. Syder (1983). Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In: J.C. Richards & R.W. Schmidt (red.), *Language and Communication*. London: Longman, 191–226.

Pollard, C., & I. Sag (1994). *Head-driven phrase structure grammar*. Stanford: Center for the Study of Language and Information.

Poss, M. (2008). *Cognitive and computational aspects of Extended Lexical Units*. Diss. Leiden, in voorbereiding.

Poss, M., & T. van der Wouden (2005). Extended Lexical Units in Dutch. In: T. van der Wouden, M. Poss, H. Reckmann, & C. Cremers (red.), *Computational Linguistics in the Netherlands 2004: Selected papers from the fifteenth CLIN meeting*. Utrecht: LOT, 187-202.

Sag, I., T. Baldwin, F. Bond, A. Copestake & D. Flickinger (2001). Multiword expressions: A pain in the neck for NLP. *LinGO Working Paper No. 2001-03* (CSLI Linguistic Grammars Online (LinGO) Lab, Stanford University). Ook in *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, Mexico City, Mexico, 1-15.

Saussure, F. de (1916). *Cours de linguistique générale*. Lausanne [etc.]. publ. par Charles Bally et Albert Sechehaye avec la collab. d'Albert Riedlinger.

Schiffrin, D. (1987). *Discourse markers*. Cambridge: Cambridge University Press.

Schutz, Rik (2007). 'Poeslief' of 'brutaal als de beul'? Versterkende vergelijkingen in het Nederlands. *Onze Taal* 76, 134-136

Sprenger, S. (2003). *Fixed expressions and the production of idioms*. Diss. Nijmegen.

Stroop, J. (2006). 'Wauwelwoorden' en wat erop lijkt. *Vaktaal* 19, 5–6.

Tomasello M. (2003). *Constructing a language: a usage-based theory of language acquisition*. Cambridge, MA [etc.]: Harvard University Press.

Uijlings, B.J. (1956). *Praat op heterdaad*. Assen: Van Gorcum [etc.]. [ook als diss. Utrecht onder de titel *Syntactische verschijnselen bij onvoorbereid spreken*.]

Uit den Boogaart, P.C. (red.) (1975). *Woordfrequenties in geschreven en gesproken Nederlands*. Utrecht: Oosthoek, Scheltema & Holkema.

Verhagen, A. (1994). Nauwelijks is niet bijna, laat staan vrijwel helemaal. In: R. Boogaart & J. Noordegraaf (red.), *Nauwe betrekkingen. Voor Theo Janssen bij zijn vijftigste verjaardag*. Amsterdam & Münster: Stichting Neerlandistiek VU & Nodus Publikationen, 273–282.

Vismans, R. (1994). *Modal particles in Dutch directives: a study in functional grammar*. Diss. Vrije Universiteit Amsterdam.

Vries, J. de (2001). *Onze Nederlandse Spreektaal*. Den Haag: SDU Uitgevers.

Vries, W. de (1910). *Dysmelie. Opmerkingen over syntaxis.* Verhandeling behoorende bij het programma van het gymnasium der gemeente Groningen voor het jaar 1910–1911.

Vries, W. de (1911). *Dysmelie. Opmerkingen over syntaxis (vervolg).* Verhandeling behoorende bij het programma van het gymnasium der gemeente Groningen voor het jaar 1911–1912.

Vries, W. de (1914). *De typen der mededeeling. Opmerkingen over syntaxis.* Verhandeling behoorende bij het programma van het gymnasium der gemeente Groningen voor het jaar 1914–1915.

Weinreich, U. (1969). Problems in the analysis of idioms. In: Joan Puhvel (red.), *Substance and structure of language*. Berkeley and Los Angeles: University of California Press, 23–81. (Herdrukt in *On Semantics*, 1980).

Woordenlijst (2005). *Woordenlijst Nederlandse Taal*. Tielt, Den Haag: Lannoo, Sdu. Samengesteld door het Instituut voor Nederlandse Lexicologie, met een leidraad door Ludo Permentier.

Wotjak, B. (1992). *Verbale Phraseolexeme in System und Text*. Tübingen: Niemeyer.

Wouden, T. van der (1992). Beperkingen op het optreden van lexicale elementen. *De Nieuwe Taalgids* 85, 513–38.

Wouden, T. van der (1995). Moeilijk is (not) difficult. In: Marcel den Dikken & Kees Hengeveld (red.), *Linguistics in the Netherlands* 1995. Amsterdam: Benjamins, 203–14.

Wouden, T. van der (1998). Dat had niet zo gehoeven: Modaliteit en negatie in de nieuwe ANS. *Nederlandse Taalkunde* 3, 237–252.

Wouden, T. van der (2002a). Particle research meets corpus linguistics: on the collocational behavior of particles. In: Ton van der Wouden, Ad Foolen, & Piet Van de Craen (red.), *Particles. Belgian Journal of Linguistics* 16. Amsterdam: John Benjamins, 151–174.

Wouden, T. van der (2002b). Partikels: naar een partikelwoordenboek voor het Nederlands. *Nederlandse Taalkunde* 7, 20–43.

Wouden, T. van der (2007a). On the phraseology of stop words. In: P. Hendriks, F. Landsbergen & M.J. van der Wal (red.), *Leiden Papers in Linguistics* 4.1, 56-67 (http://www.lucl.nl).

Wouden, T. van der (2007b). Bouwstenen van het Nederlands. Een pleidooi voor (nog) meer aandacht voor de idiomaticiteit van het Nederlands. Te verschijnen in *Neer-*

landistiek in contrast. Bijdragen aan het Zestiende Colloquium Neerlandicum, Amsterdam, Rozenberg, 287-298.

Wouden, T. van der (2007c). De prenominale modale infinitief: een corpusstudie. Voordracht Taaldag BKL/CLB, Gent, 5 mei 2007.

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

DigiTaal

Digitale bronnen voor taalhistorisch onderzoek van het Nederlands

EVIE COUSSÉ *

Het laatste decennium zijn steeds meer oude teksten digitaal toegankelijk geworden op cd-rom of op het internet. Hoewel die documenten niet altijd gedigitaliseerd zijn uit taalkundige overwegingen, ligt hun bruikbaarheid voor taalhistorisch onderzoek van het Nederlands voor de hand. Zo zijn digitale oude teksten – meer dan papieren tekstuitgaven – snel én gemakkelijk te doorzoeken op trefwoorden. In deze bijdrage wil ik een overzicht geven van oude Nederlandstalige teksten beschikbaar op cd-rom en op het internet anno 2007. Hierbij zal ik niet enkel taalkundige historische corpora bespreken maar heb ik daarnaast ook oog voor het linguïstische potentieel van digitale bibliotheken met literaire werken en digitale archieven met ambtelijke stukken op het internet.

1 Taalkundige corpora

Wie enige interesse heeft voor het Middelnederlands, zal allicht de *Cd-rom Middelnederlands* kennen. Het schijfje bevat naast het gedigitaliseerde Middelnederlandsch Woordenboek (inclusief de bouwstoffen en het citatenboek) ook een schat aan teksten die een prima basis vormen voor de taalkundige analyse van het Middelnederlands. We vinden er de digitale editie van het Corpus Gysseling (alle ambtelijke en literaire teksten uit de dertiende eeuw) en een keur aan middeleeuwse rijm- en prozateksten. De gebruiker kan via een zoekscherm uitgebreid grasduinen in het woordenboek en het vele tekstmateriaal.[1]

De chronologische aanvulling van het Middelnederlandse materiaal is het gedigitaliseerde *Woordenboek der Nederlandsche Taal* (WNT), dat de Nederlandse woordenschat beschrijft vanaf de zestiende eeuw tot in de twintigste eeuw. Hoewel het WNT in de eerste plaats een woordenboek is, vormen de vele citaten bij elk lemma een interessante dataset voor de taalkundige studie van het Nieuwnederlands. Het WNT is niet meer verkrijgbaar op cd-rom maar staat wel sinds kort online (iWNT).[2] In de toekomst zal het woordenboek door het Instituut voor Nederlandse Lexicologie uitgebreid worden met het digitale Oudnederlands Woordenboek (ONW), het Vroegmiddelnederlands Woordenboek (VMNW) en het Middelnederlandsch Woor-

* De auteur is aspirant-onderzoeker bij het Fonds voor Wetenschappelijk Onderzoek – Vlaanderen (FWO). Adres: Universiteit Gent, Vakgroep Nederlandse Taalkunde, Blandijnberg 2, 9000 Gent, België. E-mail: evie.cousse@ugent.be.

1 Voor een bespreking van de zoekmogelijkheden in de cd-rom, zie Van der Sijs (1999).

2 Website Instituut voor Nederlandse Lexicologie, http://www.inl.nl.

denboek (MNW) tot een geïntegreerde taalbank (GTB) op het internet. Dankzij die geïntegreerde digitale woordenboeken zal het voor de neerlandicus mogelijk worden om met één enkele zoekopdracht de hele geschiedenis van een woord te traceren vanaf de vroegste bronnen in het Nederlands tot vandaag.

De bestaande taalkundige corpora vormen een mooie steekproef van het historische Nederlands en hebben hun nut al bewezen bij allerlei diachrone linguïstische kwesties. Voor systematischer taalhistorisch onderzoek naar bijvoorbeeld dialectvariatie in de middeleeuwen voldoen de besproken digitale corpora echter niet. Afgezien van de ambtelijke bescheiden in het Corpus Gysseling zijn de meeste teksten op de cd-rom Middelnederlands immers niet precies te lokaliseren (en te dateren nota bene). Op dit moment loopt er een tweetal corpusprojecten, waarin het ambtelijke tekstmateriaal uit de dertiende eeuw in de tijd wordt aangevuld.

Sinds de jaren tachtig wordt er in de Vrije Universiteit van Amsterdam gewerkt aan de verzameling, transcriptie en annotatie (lemma's en woordsoortinformatie) van een *Corpus Veertiendeeeuwse Middelnederlandse Oorkonden*.[3] De oorkondeverzameling staat ook bekend als het Corpus Van Reenen-Mulder, genoemd naar zijn vroegste samenstellers. Momenteel wordt de bestaande collectie binnen een Vlaams-Nederlands samenwerkingsverband verder vervolledigd met ontbrekend materiaal uit de dialectstreken Vlaanderen en Brabant. Het afgewerkte corpus zal in de toekomst digitaal beschikbaar gesteld worden door het Instituut voor Nederlandse Lexicologie. In afwachting van de definitieve oplevering kunnen geïnteresseerde gebruikers alvast contact opnemen met de samenstellers voor een tussentijdse versie van het corpus.[4]
In de Radboud Universiteit te Nijmegen worden er dan weer binnen het project Variatie en Standaardisering[5] teksten verzameld uit de vijftiende, zestiende en zeventiende eeuw. Tijdens die periode van intensief taalcontact onderging het Nederlands ingrijpende veranderingen, wat de corpussamenstellers ertoe heeft aangezet hun corpus *Dutch in Transition* (DiT) te noemen. Het tekstmateriaal komt uit de verschillende dialectstreken van ons taalgebied met bijzondere aandacht voor de oostelijke regio's. Het corpus bevat in hoofdzaak formele teksten, waar mogelijk aangevuld met informele bronnen (brieven, dagboeken en andere egodocumenten) die traceerbaar zijn in tijd en ruimte. Na afronding van het project zal de databank online komen te staan.

2 Digitale bibliotheken met literaire werken

Los van de taalkundige corpora bevinden er zich verrassend veel oude Nederlandstalige teksten op het internet die interessant zijn voor linguïstisch onderzoek. Het gaat veelal om literaire werken die in virtuele bibliotheken digitaal ontsloten werden voor de internetgemeenschap. Die elektronische boeken kunnen dienen als (een chronologische) aanvulling van de literaire teksten op de cd-rom Middelnederlands. Geen van die teksten is echter taalkundig geannoteerd, wat veel manueel werk met zich mee kan brengen bij de taalkundige analyse van die bronnen.

De grootste collectie Nederlandstalige literatuur op het internet vind je in de *Digitale Bibliotheek voor de Nederlandse Letteren* (dbnl), een initiatief van de Maatschappij der Nederlandse Letterkunde.[6] De webstek bevat naast de grote klassiekers uit de Nederlandse literatuur ook wat minder bekende werken. Het aanbod teksten is

3 Voor een bespreking van het corpus, zie Van Reenen & Mulder (1993).
4 Piet van Reenen (p.van.reenen@let.vu.nl) of Margit Rem (m.rem@let.ru.nl).
5 Website Project Variatie en Standaardisering, http://oase.uci.ru.nl/~varstan.
6 Website Digitale Bibliotheek voor de Nederlandse Letteren, http://www.dbnl.org.

bijzonder rijk en gevarieerd: we vinden er ridderromans, heiligenlevens, kluchten, refreinen, sonnetten, tragedies, lofdichten, politieke pamfletten, wetenschappelijke traktaten, dagboeken, brievenromans, novellen, enz. vanaf de middeleeuwen tot vandaag. De titels zijn gerangschikt per periode, genre en auteur, wat de zoektocht naar geschikt corpusmateriaal een stuk vergemakkelijkt. Ook de beschikbare digitale secundaire literatuur over werk en auteur is erg nuttig bij de selectie van de ideale corpustekst. Voor de echte taalkundige analyse van de teksten is de website minder goed uitgerust: de zoekfunctie is erg algemeen en laat niet toe om een subgroep van documenten (bv. alle middeleeuwse proza) afzonderlijk te doorzoeken. Er zit voor de linguïst niet veel anders op om een selectie van werken van de verschillende webpagina's te halen (manueel door te knippen en plakken of automatisch met behulp van webspider-software) en met een eigen concordantieprogramma te analyseren.

Een alternatieve verzameling literaire teksten is terug te vinden bij het *Project Laurens Jz. Coster*.[7] Die digitale bibliotheek is niet beheerd door een vereniging of instituut zoals de dbnl maar wordt gedragen door een informeel netwerk van internetadepten die vrijwillig hun favoriete literatuur digitaliseren en online ontsluiten. In de loop der jaren heeft het project een aanzienlijk deel van de Nederlandstalige literaire canon vanaf de middeleeuwen tot vandaag online samengebracht.[8] Het project is de laatste jaren niet meer zo actief en het merendeel van de collectie is op dit moment ook al in de dbnl opgenomen. Wie echter goed zoekt, vindt op de website toch nog enkele transcripties die nergens anders op het internet te vinden zijn. Het vergelijkbare internationale *Gutenberg Project* heeft ook nog eens een tweehonderdvijftigtal boeken uit vooral de negentiende en twintigste eeuw gedigitaliseerd.[9]

Naast de Nederlandse canonieke literatuur zijn er online ook enkele kleinere collecties met letterkundige curiosa beschikbaar. Die teksten zijn door hun maniëristische taalgebruik niet altijd even geschikt voor taalkundig onderzoek. Wanneer andere bronnen echter schaars zijn, kunnen ze toch een interessante aanvulling betekenen voor taalkundige corpora. Om te beginnen biedt de website *Ceneton* een overzicht van alle overgeleverde toneelstukken uit de Nederlanden tot 1803. Een selectie van 116 toneelteksten zijn als facsimile beschikbaar en 85 toneelteksten zijn volledig getranscribeerd.[10] Daarnaast bestaat er een online verzameling van 660 *Nederlandse Heldinnenbrieven* van 1415 tot vandaag. Dat dichtgenre was in het Nederlandse taalgebied vanaf de vijftiende eeuw populair maar raakte later wat in de vergetelheid.[11] De webpagina *Ursicula* bevat dan weer een reeks digitale facsimile's van gedichten, toneel en pamfletten uit de zeventiende en achttiende eeuw.[12] Op de webstek van het *Dutch Emblem Project* ten slotte zijn 26 embleemboeken uit de zeventiende eeuw digitaal toegankelijk gemaakt in facsimile en transcriptie. Een embleem (of zinnebeeld) is een kleine afbeelding met daarbij een korte kernachtige spreuk, zoals een spreekwoord of een kort gedicht.[13]

6 Website Digitale Bibliotheek voor de Nederlandse Letteren, http://www.dbnl.org.
7 Website Project Laurens Jz. Coster, http://cf.hum.uva.nl/dsp/ljc.
8 Een selectie van circa 150 teksten zijn in 1999 ook op de *cd-rom Klassieke Literatuur* uitgegeven.
9 Website Gutenberg Project, http://www.gutenberg.org.
10 Website Ceneton, http://www.let.leidenuniv.nl/Dutch/Ceneton.
11 Website Nederlandse heldinnenbrieven, http://www.let.leidenuniv.nl/Dutch/Heroides.
12 Website Ursicula, http://www.let.leidenuniv.nl/Dutch/Renaissance/Ursicula.
13 Website Dutch Emblem Project, http://emblems.let.uu.nl.

3 Digitale archieven met ambtelijke stukken

Naast literaire werken zijn er op het wereldwijde web ook heel wat oude documenten te vinden met een geschiedkundig belang. Die teksten kunnen een waardevolle uitbreiding betekenen van de ambtelijke stukken uit het Corpus Gysseling, in afwachting van de lopende corpusprojecten.

Voor teksten met een regelgevend karakter is de portaalpagina *Regelgeving in de Nederlanden* een uitstekend vertrekpunt.[14] Per provincie wordt hier een overzicht gegeven van de beschikbare regelgevende bronnen op het internet. Je vindt er verwijzingen naar grote compilaties rechtsbronnen die gedigitaliseerd zijn door historici van de Katholieke Universiteit Leuven[15] en de Universiteit Maastricht.[16] Daarnaast zijn er honderden links naar webpagina's van heemkundige genootschappen, musea en bibliotheken die (veelal lokale) ambtelijke bescheiden online hebben geplaatst.[17] Het aanbod is erg divers: oorkonden, stadskeuren, statuten, ordonnanties, plakkaten, wetten, enz. Ook de presentatie van het materiaal is wisselend: soms staat enkel een digitale afbeelding van het originele manuscript online, soms is het document ook getranscribeerd en in tekstformaat ter beschikking. Niet altijd is die transcriptie zomaar geschikt voor elk linguïstisch onderzoek. Zo lossen historici vaak stilzwijgend afkortingen op in oude bronnen om ze toegankelijker te maken voor de moderne lezer. Terwijl die ingreep weinig gevolgen heeft voor bijvoorbeeld woordvolgordestudies, vormt ze voor fonologisch of morfologisch onderzoek dan vaak weer een onoverkomelijk probleem.

Narratieve geschiedkundige tekstsoorten zijn verzameld op de website van het *Instituut voor Nederlandse Geschiedenis*.[18] We vinden er kronieken, de briefwisseling van grote staatsmannen (bv. Willem van Oranje, Constantijn Huygens, Hugo De Groot, Johan de Wit) en ambtelijke bescheiden van de middeleeuwen tot vandaag. Die bronnen zijn beschikbaar als digitale afbeelding die ook als tekst doorzoekbaar is. Daarnaast zijn op de themapagina's rond de *Tachtigjarige Oorlog* nog meer narratieve tekstgenres te vinden zoals brieven en kronieken uit de periode 1550 tot 1650.[19]

4 Toemaatje

Tot slot komt de interesse voor oude teksten soms uit onverwachte hoek. Een kleine greep uit het grote aanbod… Gastronomen komen op *Kookhistorie* niet alleen culinaire geheimen uit grootmoeders keuken te weet, maar vinden er ook enkele kookboeken uit de zestiende en zeventiende eeuw in facsimile en transcriptie.[20] Een oud-leraar natuurwetenschappen heeft dan weer op zijn homepage *Van woorden en wetenschap* wetenschappelijke traktaten van Simon Stevin en Christiaan Huygens geplaatst.[21] In dit

14 Website Regelgeving in de Nederlanden, http://www.geneaknowhow.net/regel/nederlanden.

15 Website Rechtsbronnen Katholieke Universiteit Leuven, http://www.kuleuven-kortrijk.be/facult/rechten/Monballyu/Rechtlagelanden/Homepage.htm.

16 Website Rechtsbronnen Universiteit Maastricht, http://lrg.unimaas.nl/pz/plakkaatzoeker.html.

17 Het is ook uitkijken naar lopende digitaliseringsprojecten van regelgevende teksten. Op http://www.cartago.nl worden de komende jaren alle oorkonden tot 1600 uit Groningen en Drenthe digitaal ontsloten. Op http://www.kantl.be/ctb vind je meer informatie over het project *Talige aspecten van gerecht, onderwijs en politiek in het Verenigd Koninkrijk der Nederlanden* met ambtelijke bronnen uit de negentiende eeuw.

18 Website Instituut voor Nederlandse Geschiedenis, http://www.inghist.nl.

19 Website Tachtigjarige Oorlog, http://dutchrevolt.leidenuniv.nl.

20 Website Kookhistorie, http://www.kookhistorie.com.

21 Website Van woorden en wetenschap, http://www.xs4all.nl/~adcs.

overzicht mogen ten slotte ook niet de digitale edities van de *Statenbijbel* ontbreken.[22]

5 Zelf op zoek...

Het geschetste overzicht biedt verwijzingen naar honderden – misschien zelfs duizenden – digitale oude teksten op het internet. Toch is het overzicht verre van compleet… niet alleen is het internet in omvang nauwelijks te overzien, elke dag komen er nog nieuwe data bij. Niets houdt echter de ondernemende linguïst tegen om met behulp van een internetzoekmachine zelf op schattenjacht te gaan naar andere historische documenten op het wereldwijde web. Niet alleen zijn nieuwe bronnen aan te boren met beschrijvende trefwoorden als *middeleeuwse oorkonde* of *dagboek gouden eeuw*. Je kan ook experimenteren met de afwijkende spelling of woordenschat van het Nederlands in oudere documenten. Zo zal een zoekopdracht naar het trefwoord *ghy* of *wylieden* een hele lijst opleveren met historische teksten in het Nederlands. Meer gerichte zoekvragen als bijvoorbeeld *scepen, voorseit, ghebieden, ghesegelt, lettren* brengen je dan weer op het spoor van echte middeleeuwse oorkonden ... digitaal, wel te verstaan.

Bibliografie

***Cd-rom Klassieke Literatuur. Nederlandse letterkunde van de Middeleeuwen tot en met de Tachtigers* (1999).** Utrecht: Uitgeverij Het Spectrum. [uitverkocht]

***Cd-rom Middelnederlands* (1998).** Samengesteld door het Instituut voor Nederlandse Lexicologie. Den Haag/Antwerpen: Sdu Uitgevers/ Standaard Uitgeverij. [uitverkocht]

***Het Woordenboek der Nederlandsche Taal op CD-Rom* (2003).** Samengesteld door het Instituut voor Nederlandse Lexicologie. (Delen I-XXIX, met Aanvullingen; derde release). Den Haag: Sdu Uitgevers. [uitverkocht]

Van Reenen, P.T. & M. Mulder (1993). Een gegevensbank van 14de-eeuwse Middelnederlandse dialecten op de computer. *Lexikos* 3, 259-281.

Van der Sijs, N. (1999). Bespreking van de cd-rom Middelnederlands. DigiTaal, http://www.niederlandistik.fu-berlin.de/digitaal/digitaal-07.html.

22 Op http://www.statenvertaling.net is de Jongbloed-editie van 1888 digitaal doorzoekbaar. Zie ook de portaalpagina van het Nederlands Bijbelgenootschap http://80.252.86.246 voor de facsimile's van de Delftse Bijbel (1477), de Statenvertaling (1637) en de Mortierbijbel (1700). Recent is door Nicoline van de Sijs een oproep in Neder-L gelanceerd om de originele statenbijbel van 1637 te digitaliseren (zie http://www.neder-l.nl/bulletin/2007/06/070602.html).

Boekbesprekingen

Evers-Vermeul, Jacqueline. *The development of Dutch connectives. Change and acquisition as windows on form-function relations.* Utrecht: LOT, 2005. XII + 283 blz. ISBN 90 76864 77 2. EUR 29,88.

Jacqueline Evers-Vermeul vraagt zich in haar proefschrift af hoe zinsverbindende woordjes (connectieven) zich ontwikkelen. Onder zinsverbindende woordjes rekent ze niet alleen de nevenschikkende en onderschikkende voegwoorden, maar ook bijwoorden. Daarbij heeft ze zowel naar de geschiedenis als naar de verwerving van het Nederlands gekeken. Ze gelooft dat er bij beide sprake zal zijn van een 'venster' op de verhouding tussen (grammaticale) vorm en (taalgebruiks-)functie. Het proefschrift is opgesplitst in vier delen, vijftig pagina's algemene inleiding, een honderd pagina's over *want* en *omdat*, *dus* en *daarom* in de taalgeschiedenis van het Nederlands, en vervolgens 75 pagina's over het verschijnen in de kindertaal van vooral *want* en de additieve connectieven *en*, *maar*, en *toen*. Tenslotte fungeert een kort hoofdstuk als conclusie. Ik zal hieronder eerst een paar van de vele aandachtspunten nagaan en daarna de algemene visie van het proefschrift heroverwegen.

1 Inleiding

Het inleidende deel (deel I) refereert aan Sanders et al. (1993) als uitgangspunt van het proefschrift. Het gaat Sanders et al., - en daarmee ook Evers-Vermeul -, eerder om inhoudelijke, pragmatische, en stilistische functie-karakteristieken van zinsverbindende woordjes dan om grammaticale vorm-distincties. Sanders gebruikt voor connectieven termen als: neutraal additief (*en*), nadrukkelijk temporeel (*toen*), nadrukkelijk redengevend (*omdat*, *want*, *daarom*, *dus*) en polaire waardering (*maar*, *toch*, *heus wel*, *ook al*). Temporeel, redengevend, polair worden daarbij gezien als een set kenmerken toevoegbaar aan het neutrale additief *en*. In plaats van verhalen over X-bar projecties en regeer- en skoopverhoudingen volstaat de grammaticale karakteristiek met 'vooropplaatsing', 'persoonsvorm laat' en 'bijwoord mediaal geplaatst'.

2 Diachronie

Het historisch onderzoek (deel II) is gebaseerd op de constructie van een corpus van 150 tekstfragmenten. Deze fragmenten zijn steeksproefsgewijs geselecteerd uit drie periodes, de 13e eeuw, 16e eeuw en 20e eeuw. Voor elke periode zijn de eerste 50 voorkomens van een connectief genomen. De representativiteit van de geselecteerde teksten wordt in bijlagen ondersteund met een statistische overweging. Die strekt ertoe dat de frequentie van werkwoordplaatsing of bijwoordelijk versus voegwoordelijk gebruik gewaardeerd kan worden als een historische

verschuiving. Het werk van Verdam ('Middelnederlandsch Handwoordenboek', 1911), met zijn massale aantekeningen over zinsconnectieven, wordt nergens genoemd door Evers-Vermeul. Ik vermoed dat de promovenda hiermee bij haar promotoren weggekomen is, omdat de aantekeningen van Verdam, hoe uitvoerig ook, zich niet lenen voor het construeren van ontwikkelingscurves of staafdiagrammen en de selectie van de eerste 50 tekstfragmenten juist wél.

De diachronie van *want* en *omdat* (hoofdstukken 5 en 6) voert tot de best volgehouden poging om vorm en functie van connectieven met elkaar in verband te brengen. In het hedendaagse Nederlands onderscheiden de *want*-zin en de *omdat*-zin zich als nevenschikking en onderschikking blijkens de plaatsing van de persoonsvorm: aan het eind van het predikaat (V-laat) in de *omdat*-zin, en in tweede positie (V-second) in de *want*-zin. Bovendien kan de *want*-zin alleen in nageplaatste positie voorkomen, terwijl de *omdat*-zin ook in vooropgeplaatste en tussenliggende positie kan verschijnen.

De hedendaagse *want*-zin staat daarnaast een wisseling van illocuties (uitroep, vraag, bevel) toe, zie (1), terwijl een ondergeschikte zin dat niet toestaat.

(1) a. want (uitroep) wat een mooi weer was het!
b. want (vraag) was het wel mooi weer?
c. want (bevel) kijk maar eens naar het weer!

Verder kan de *omdat*-zin al dan niet onder het bereik (de skoop) van de negatie vallen, zie de ambiguïteit in (2)a. Zo'n ambiguïteit is bij het hedendaagse *want* uitgesloten, zie (2)b.

(2) a. Niemand kwam omdat de zon scheen
b. Niemand kwam, want/en wel omdat/aangezien de zon scheen.

Evers-Vermeul toont zich overigens niet al te geïnteresseerd in verdere grammaticale of lexicale variaties. Ze onderstreept eerder een pragmatisch onderscheid. Redengevend verband via *omdat* beroept zich op iets wat gewaardeerd wordt als een vanzelfsprekende relatie, terwijl het redengevend verband met *want* meer gezichtspunt van de spreker is. De *want*-zin in (3)b geeft veeleer een argument voor de bewering in de voorafgaande zin. Ze gebruikt voor dit onderscheid de termen 'content domein' en 'epistemisch domein' (Sweetser 1990).

(3) a. We gingen in de tuin zitten, omdat de zon scheen
(oorzaak-gevolg relatie, content domein)
b. De temperatuur zal waarschijnlijk stijgen, want de zon schijnt
(conclusie-argument relatie, epistemisch domein)

De twee soorten redengevend verband vind je in allerlei talen terug (Frans *car*/*parce que*, Engels *for*/*because*, Duits *denn*/*weil*) met toepassing van dezelfde structurele middelen (nevenschikking/onderschikking). Het is aannemelijk dat het Middelnederlands hetzelfde verschil tussen die twee functies, content versus epistemisch, gemaakt heeft. Maar in het Middelnederlands waren *want* en *omdat* grammaticaal minder verschillend dan in het hedendaags Nederlands. Zowel *omdat* als *want* pasten de V-second regel optioneel toe. De vraag die vervolgens gesteld kan worden is de volgende: is V-laat in de historische teksten een indicatie dat de connectiefzin met *want* ook een (content) redengever kon zijn, in de mate die vergelijkbaar is met een *omdat*-zin in 20e-eeuws Standaardnederlands? En is omgekeerd V-second in de historische teksten een indicatie dat de *omdat*-zin ook een epistemische redengever kon zijn, vergelijkbaar met een *want*-zin in 20e-eeuws Standaardnederlands? Die tweede vraag, de functie van *omdat* met V-second, stelt Evers-Vermeul zich niet, omdat haar evidentie-

basis kennelijk niet dergelijke zinnen opleverde (maar zie Verdam 1911).

Van de 100 *want*-zinnen uit de 13e en 16e eeuw, vertonen er 65 een ondubbelzinnige werkwoordplaatsing: 50 V-second en 15 V-laat (tabel 6.4, p.103). Binnen die groep tekent zich af hoe de *want*-zin tendeert naar het 20e eeuwse Nederlands waar de ambiguïteit van de *want*-zin helemaal verdwenen is met een strikte arbeidsverdeling tussen het onderschikkend *omdat* (content relatie) en het nevenschikkend *want* (gezichtspunt). Evers-Vermeul concludeert dat de verdwijning van V-laat in *want*-zinnen plaats heeft gevonden onder druk van het alternatief *omdat* dat hoofdzakelijk content relaties uitdrukte.

Voor het onderzoek naar de diachronische veranderingen van *dus* en *daarom* (hoofdstuk 7) beperk ik mij hier tot de veranderingen in betekenis die *dus* heeft ondergaan, en de manier waarop het hedendaagse *dus* meer of minder bekende conclusies introduceert. Dankzij haar methode van tekstselectie kan Evers-Vermeul de historische verschuivingen van *dus* in beeld brengen.

In de 13e eeuw was *dus* nog overwegend een anafoor met twee anaforische betekenissen: *dus* = *'als volgt'* (cataforisch, refererend aan wat volgt), en *dus* = *'op die manier'* (refererend aan wat voorafgaat). Beide gebruiken zijn bewaard in het Modern Nederlandse *aldus*. In de 16e eeuw wordt de functie van het anaforische *dus* uitgebreid met het gebruik als bijwoord van graad ('dusdanig') *dus deerlic*.

In de 20e eeuw verdwijnt het anaforisch gebruik van *dus* helemaal (p. 116, figuur 7.1), en wordt *dus* overwegend redengevend connectief. Daarnaast verschijnt *dus* als 'discourse markeerder', een elementje dat gebruikt wordt om terug te grijpen naar informatie die gezien wordt als bekend/voorondersteld (p. 117, vb. (19)).

(4) Het ging erom wie het mooiste, het beste produkt maakte. Dat was ik **dus**.

In (4) drukt *dus* geen redengevende relatie uit, maar wordt het specifiek gebruikt om aan te geven dat de informatie (de spreker als gedoodverfde winnaar) al bekend is. Dergelijke discourse markeerders vermijden de zinsinitiële positie.

Het zinsverbindende (connectief) *dus* kan in het hedendaagse taalgebruik zowel zinsinitieel als zinsintern voorkomen. Het vooropgeplaatste *dus* kan met of zonder subjectinversie optreden, zie (5). Wanneer er geen inversie optreedt moet men wel denken aan een voegwoord-status, vergelijk het gebruik van *want*.

(5) Jij hebt de puzzel uit de kast gehaald,
dus jij moet 'm er weer in doen.
dus moet jij 'm er weer in doen.

De verschillende posities van *dus* (zinsintern als discourse markeerder, zinsinitieel of zinsintern als redengevend woordje) zijn door Ariel (1999) gerelateerd aan de mate van 'toegankelijkheid van relevante discourse-informatie'. Volgens Ariel (1999) geven connectieven in zinsinterne positie aan dat de relevante informatie al in de voorafgaande discourse toegankelijk was, zie (6)a. De zinsinitiële positie (6)b daarentegen is bedoeld om een nieuw licht op het verhaal te werpen.

(6) a. P. woonde in die laan, Hendriks kwam er **dus** langs (conclusie als oude koek)
b. P. woonde in die laan, **dus** Hendriks kwam er langs (conclusie als nieuwe vondst)

In het Middelnederlands stond het connectef steevast in zinsinitiële positie. De verschuiving van het connectief *dus* naar zinsinterne positie is een ontwikkeling van de 20e eeuw en evenzo het ontstaan van *dus* als discourse markeerder.

Evers-Vermeul heeft tellingen uitgevoerd om uit te maken of het zinsinitiële *dus* en het zinsinterne *dus* inderdaad de verschillende infor-

matie-functie hadden die Ariel (1999) voorstelt. Het antwoord was 'meestal wel' (p. 124, tabel 7.7). Wat ik eigenlijk niet helemaal volg, is hoe dat onafhankelijk beoordeeld kan worden en dan geteld. Wanneer de schrijver dat *dus* vooropplaatst dan geeft hij daarmee een nieuw perspectief, of de latere lezer (Ariel, Evers-Vermeul) dat nu een zouteloze wending vindt of niet. De intuïtieve beoordeling kan natuurlijk wel juist zijn, maar blijft wel louter een subjectieve indruk. Wat wel geconstateerd kan worden is of een bepaalde auteur, dan wel een bepaalde periode, een opvallende voorkeur gaat krijgen voor *dus* als zinsintern connectief. Daarvoor is de tekstselectie van Evers-Vermeul (de eerste 50 connectieven van een bepaald type) wél ondubbelzinnig geschikt.

3 Verwerving

Het taalverwervingsdeel (deel III) staat tamelijk los van het historische deel II. Dat kon ook moeilijk anders, want de connectieven waar het historische deel zich speciaal op richt komen pas laat in de kindertaal voor. De eenheid van de studie wordt daardoor enigszins benadeeld. Bij de kindertaal is er vooral gekeken naar de 'additieve' connectieven *en*, *maar*, *toen* en naar het redengevende *want*, waarschijnlijk omdat Sanders et al. (1993) over deze elementjes het een en ander te zeggen hadden.

Evers-Vermeul heeft de ontwikkeling van *en*, *maar*, *toen* en *want* bekeken bij 12 kinderen en zowel kwantitatieve als kwalitatieve methodes toegepast. Er is gekeken naar het eerste correcte gebruik (hoofdstuk 10), er worden groeicurves van de ontwikkeling gegeven (hoofdstuk 11). De groeicurves zijn gemaakt op basis van een statistische methode die de waarschijnlijkheid van voorkomen toont. Verder is het gebruik van de connectieven kwalitatief geanalyseerd (hoofdstukken 12, 13).

Prominent in dit deel van de studie (hoofdstukken 10 en 11) staat de notie 'cumulatieve conceptuele complexiteit' (Bloom et al. 1980).

De cumulatieve conceptuele complexiteit die Evers-Vermeul verdedigt is gebaseerd op de vier cognitieve primitieven van Sanders et al. (1993). In de analyse van Sanders et al. is er een neutraal additief *en*. De betekenis van *maar*, *toen* en *want* volgt dan door toevoeging van één kenmerk: *maar* <+negatief>, *toen* <+temporeel>, *want* <+redengevend>. Een stapeling van die distincties voorspelt dan een zeer late verwerving.

Het blijkt dat conceptuele complexiteit, zoals voorspeld door Evers-Vermeul (p. 194 (19)), inderdaad een rol speelt in de volgorde van verwerving voor 11 van de 12 kinderen. *En* is het minst complexe connectief. Het is ongespecificeerd voor negatief, redengevend en temporeel <α negatief, α redengevend, α temporeel>. Het wordt dan ook het eerst verworven, gevolgd door *maar* gebruikt als negatief additief connectief <+negatief, α redengevend, α temporeel>. De verwervingsvolgorde van *want* en *toen* blijkt afhankelijk van de vraag of *want* al of niet een temporele relatie markeert.

Kleine kinderen van net drie jaar kunnen nog geen samengestelde zinnen, bestaande uit twee proposities, vormen. Dergelijke zinnen zijn te complex voor hun verwerkingscapaciteit (Brown & Hanlon 1970, Diessel 2004). Diessel (2004), die zich baseert op de verwerving van het Engels, stelt dat de eerste nevengeschikte zinnen los van elkaar gebruikt worden, in aparte zinnetjes die qua syntaxis en intonatie niet geïntegreerd zijn. Het verbindende woordje wordt daarom als discourse markeerder, eerder dan als connectief, gekarakteriseerd. De integratie van nevengeschikte en ondergeschikte zinnen komt later. Evers-Vermeul volgt dit betoog in hoofdstuk 12 en illustreert hoe kinderen steeds grotere en complexere eenheden kunnen verwerken. Overigens, wanneer een kleuter binnen de conversatie, maar wel in een aparte zin, een aanvulling plaatst van het type *toen ik nog baby was* (p. 225), dan lijkt mij dat een zeer goed geïntegreerd gebruik van grammaticale onderschikking. Dat blijkt ondubbelzinnig uit de zinsfinale werkwoordsplaatsing.

(7) a. (Oh, heb je die van Rosie gekregen?) **Toen** we in Sloten waren. (Josse, 3;1.10)
b. (Waarom wil jij mij niet helpen?) **Omdat** ik niet zin heb. (Hein, 3;0.11)
c. Ik ben al in het ziekenhuis (ge)legen. **Toen** ik nog baby was. (Niek, 3;10.3)

De voorbeelden in (7) tonen dat onderschikking vroeg voorkomt in de Nederlandse kindergrammatica, rond het derde jaar. Niettemin, de complexiteit van *en*, *maar*, *toen*, en *want* blijkt ook in de verwervingsvolgorde.

4 'Verklarend interface' of 'arbitrair verband'?

Het model van het proefschrift heeft in eerste instantie iets aantrekkelijks. Er is een bepaald type constructie, bijvoorbeeld die voor redengevend verband (*want*/*omdat*, *dus*/*daarom*) en men gaat na hoe de varianten zich ontwikkelen in de taalgeschiedenis en in de taalverwerving. Het onderzoek kan vervolgens allerlei richtingen inslaan. De verwachting is dat die verschillende verkenningen elkaar ook gaan verhelderen. Om het beeld van Evers-Vermeul te gebruiken: het ene soort onderzoek zal een venster openen op het andere.

Door het lezen van het proefschrift ben ik eigenlijk tot een andere overtuiging gekomen. Er is wel eens een parallel tussen veranderingen van connectieven in de taalgeschiedenis en de taalverwerving, maar niet dwingend, meer incidenteel. Hetzelfde wil ik ook wel verdedigen voor de bijna heilige koe van vorm en functie. Jacqueline Evers-Vermeul somt in het laatste hoofdstuk (p. 254) zes punten op voor een interactie tussen vorm en functie. Met geen van die punten kan ik het oneens zijn, maar tegelijkertijd zijn er evengoed voorbeelden waar die hele correspondentie weer weg is. Voor de zinsinterne grammaticale vorm is er een belangrijk verschil tussen hoofdzinnen en bijzinnen en ook tussen voegwoorden en bijwoorden, maar pragmatisch, bijvoorbeeld voor de notie redengevend verband, gaan die vormen in elkaar over zonder dat dat voor het redengevend verband verschil hoeft uit te maken. Misschien hoeft er in het algemeen geen systematische samenhang te bestaan tussen het materiaal en de globale functie waarvoor het wordt gebruikt.

Dat lijkt wat vreemd, maar het is niet onbekend. In de beschouwingen van Chomsky zou taal enerzijds een 'interface' hebben met de cognitief/pragmatische interpretatie (LF interface) en anderszijds met fonetisch/fysieke interpretatie (PF interface). Bij het PF interface blijkt dat gesproken taal berust op/gebruik maakt van heel kleine botjes in het strottenhoofd en in het middenoor. De aanstuurbaarheid van die botjes vormt een klein wonder op zich en zonder die apparatuur geen gesproken taal. Dat wil niet zeggen dat het fonetisch systeem en het grammaticale systeem afbeeldingen van elkaar zijn op zo'n manier dat ze vensters op elkaar zouden openen. Dat hoeft in elk geval niet en, lijkt mij, het is ook niet zo. Wanneer de zinsgrammatica echt van de grond komt met woordsoorten, verbuigingen, woordgroepen, regeerverhoudingen, e.d., dan geeft de fonetiek daar geen 'venster' meer op. Dat kan ook voor het andere interface gelden. De cognitief/pragmatische kant van een gesprek of een discourse kan een notie als 'nieuwe topic' of 'redengevend verband' gebruiken zonder dat de realisering in de zinsgrammatica daarmee vastligt. Bij 'interface' moeten we misschien eerder denken aan het oordeel van De Saussure: het verband is arbitrair.

Om terug te keren tot de kindertaal, die zit helemaal vol met middelen om een zin in een situatiecontext te plaatsen. Kinderen van nog geen drie smijten met woordjes als {*ook*, *nou*, *eens*, *even*, *maar*, *toch*}, terwijl wij – vanuit de zinsgrammatica -, nauwelijk weten wat we daarmee aan moeten (zie Foolen 1993, Van der Wouden 2003, 2006). Kinderen hebben met drie jaar niet alleen een kerngrammatica onder de knie, maar

ook een repertoire van cognitief pragmatische conversatievormen (Wittgensteins 'language games'). Het woordje *en* is voor een peuter niet zozeer een 'neutraal additief', maar een claim op het voortzetten van zijn beurt als spreker of een protestmiddel om het onderwerp van een conversatie te veranderen {: *en mijn vader heeft een nieuwe auto*}. Daarmee is niets kwaads gezegd van de karakterisering '<α negatief, α redengevend, α temporeel>' (p.194), maar als oriëntatie van een pragmatische analyse is dat niet het meest gelukkige perspectief. De vraag zou beter kunnen zijn – los van de zinsgrammatica – : wat is het verbale repertoire waarmee een drie- of vierjarige sprekersbeurten en wendingen in de conversatie tot stand brengt? De niveaus van analyse (zinsgrammatica versus situatiepragmatiek) hebben elkaar minder te vertellen dan de schrijfster aanneemt of waar kan maken. De vraagstellingen vallen daardoor wat beperkt en eenzijdig uit. Om een voorbeeld te geven, Evers-Vermeul sluit 'incorrect' gebruikte connectieven uit bij de analyse, zoals het volgende gebruik van *maar* (p. 199, voorbeeld (27).

(8) moeder Loekie: Is de trompet op de slaapkamer?
zoontje Thomas: Ja, **maar** Loekie wil zoeken. (Thomas 2;7.2)

Het lijkt me dat Thomas dat woordje *maar* als wending in de conversatie goed gebruikt. Hij maakt een contrast met een verzwegen [*ik heb de trompet niet*] *maar* [*jij,*] *Loekie wil* [*hem*] [*toch wel*] [*op gaan*] *zoeken,* [*hè?*]. Er ontbreekt veel aan die zin vanuit een volwassen standpunt. Toch lijkt me dat *maar* perfect geplaatst. Woordjes als *hè* (verstandhouding), *toch wel* (beleefde veronderstelling), resultatief *op* met verplicht object *hem* en tijdsaspectueel *gaan*, kun je bij een kind van twee nog niet standaard verwachten. Toch is het *ja*, gevolgd door *maar* voor het doel een cognitief erg goede, om niet te zeggen een indrukwekkende, verbaal manipulatieve reactie. Voor zinsverbindend materiaal in kindertaal heeft de studie niet zoveel opgeleverd.

De historische analyse van zinsverbindend materiaal met behulp van geselecteerde datasets is het meest uitgebreide en, mijns inziens, het meest interessante deel van het proefschrift. Daar is toch een overwinning geboekt op een moeilijk deelproject en dat is een felicitatie meer dan waard.

Jacqueline van Kampen
jacqueline.vankampen@let.uu.nl

Bibliografie

Ariel, M. (1999). Mapping so-called "pragmatic" phenomena according to a "linguistic-extra-linguistic" distinction; the case of propositions marked "accessible". In: M. Darnell, E. Moravesik, F. Newmeyer, M. Noonan & K. Wheatly (red.). *Functionalism and Formalism in Linguistics. Volume II: Case studies.* Amsterdam/Philadelphia: John Benjamins, 11-38.

Bloom, L., M. Lahey, L. Hood, K. Lifter & K. Fiess (1980). Complex sentences: acquisition of syntactic connectives and the semantic relations they encode. *Journal of Child Language* 7, 235-261.

Brown, R. & C. Hanlon (1970). Derivational complexity and order of acquisition in child speech. In: J.R. Hayes (red.) *Cognition and the Development of Language* New York: Wiley, 155-207.

Diessel, H. (2004). *The Acquisition of Complex Sentences*. Cambridge: Cambridge University Press.

Foolen, A. (1993). *De Betekenis van Partikels. Een dokumentatie van de stand van het onderzoek met bijzondere aandacht voor 'maar'.* Proefschrift , Nijmegen.

Sanders, T.J.M., W.P.M. Spooren & L.G.M. Noordman (1993). Coherence relations in a cognitive theory of discourse representation. *Cognitive Linguistics* 4/2, 93-133.

Sweetser, E.E. (1990). *From Etymology to Pragmatics. Metaphorical and cultural aspects of semantic structure.* Cambridge: Cambridge University Press.

Verdam, J. (1911). *Middelnederlandsch Handwoordenboek.* Herdruk 1979-1983, Den Haag: Nijhoff.

Wittgenstein, L. (1953/1975). *Philosophische Untersuchungen.* Frankfurt am Main: Suhrkamp, 3rd edition.

Wouden, T. van der (2003). Particle research meets corpus linguistics. On the collocational behaviour of particles. In: T. van der Wouden, A. Foolen & P. van de Craen (red.). *Particles.* Amsterdam/Philadelphia: John Benjamins. Series: Belgian Journal of Linguistics 16.

Wouden, T. van der (2006). Partikels: woordjes die het Nederlands markeren. In: N. van der Sijs, J. Stroop & F. Weerman (red.). *Wat iedereen van het NEDERLANDS moet weten en waarom.* Amsterdam: Bert Bakker, 119-129.

Česke Niewold. *Spontaan herstel van afasie in en na de acute fase.* Utrecht: LOT, 2006. XII + 209 blz. ISBN 90 76864 95 0. EUR 22,88.

De Klinische Linguïstiek is nog een zeer jong specialisme binnen de Taalwetenschap. De laatste tien jaar verschijnen er echter met enige regelmaat proefschriften die ingaan op de relatie tussen taalstoornissen en taaltheorie. Een van de meest recente proefschriften is dat van Česke Niewold. Hij promoveerde in mei vorig jaar op het gebied van de afasie, een taalstoornis die optreedt als gevolg van verworven hersenletsel. Niewold heeft het zich niet gemakkelijk gemaakt met zijn onderwerpskeuze. Hij heeft twee onderdelen van de afasiologie bij elkaar gebracht waar veel afasiologen hun vingers liever niet aan branden, namelijk het herstel van taalproblemen in de acute fase enerzijds en anderzijds de informatie die spontane-taalanalyse hierbij kan verschaffen.Waarom zijn beide onderwerpen zo lastig?

Taalkundig onderzoek naar afasie maakt vaak gebruik van de data van patiënten die chronisch afatisch zijn. Dat wil zeggen dat het letsel dat tot de afasie heeft geleid minimaal drie maanden geleden moet zijn opgetreden. In het algemeen wordt aangenomen dat vanaf dat moment herstel alleen nog maar kan plaatsvinden door interventie. Belangrijke reden voor deze temporele beperking is dat men bij dergelijk onderzoek spontaan herstel juist wil uitsluiten. Zeker als data van patiënten niet op één moment verkregen kunnen worden, kan er in de beginfase na het hersenletsel een behoorlijke verandering optreden in het taalgedrag van een patiënt.

Niewold heeft echter juist de eerste fase van de afasie, de acute fase, als onderzoeksobject gekozen en dat maakt zijn proefschrift bij voorbaat interessant. De acute fase, die Niewold inperkt tot de eerste drie weken na het letsel, is een periode waarin patiënten een grote mate van herstel kunnen laten zien en het verloop hiervan is niet duidelijk te voorspellen. Op basis van een zeer uitgebreid en gedegen literatuuronderzoek

is Niewold er echter wel in geslaagd om een overzicht te geven van factoren die van invloed kunnen zijn op de mate van herstel. Daarbij doet hij voorspellingen over de percentages patiënten die grotendeels herstellen en zij die afatisch blijven. Zo zijn persoonlijke factoren als geslacht, leeftijd en IQ niet van invloed op het spontane herstel. Bij neurologische en andere medische factoren is het beeld wisselend. Zo is de oorzaak van de afasie van invloed, waarbij er een beter talig herstel optreedt na een trauma dan na een herseninfarct of een tumor. Negatieve invloeden op het herstel hebben bijvoorbeeld een halfzijdige verlamming en atriumfibrileren. Emotionele factoren als depressie en angststoornissen spelen tenslotte ook een negatieve rol bij het herstel.

Op basis van zijn literatuuronderzoek voorspelt Niewold dat van de 21 patiënten die afatisch zijn in de acute fase er 4 overlijden, 4 afatisch blijven en 4 herstellen. Over de overige 9 patiënten kan geen uitspraak worden gedaan. Niewold merkt hierbij op dat veel studies groepsstudies zijn en dat er daarom weinig bekend is over het herstel van individuele patiënten. Verder is er ook weinig bekend over het herstel op specifieke taalniveaus. Een van de beste manieren om de rol van de verschillende taalniveaus bij de taalproductie in kaart te brengen is door gebruik te maken van de spontane taal van patiënten.

Spontane taal is in dit geval de taal die wordt uitgelokt bij patiënten aan de hand van een gesprek over dagelijkse thema's als het werk, de familiesituatie, hobby's en over de ziektegeschiedenis. Bij spontane taal is het de bedoeling dat de patiënt zo veel mogelijk aan het woord is en het wordt daarom gezien als de beste benadering van het alledaagse taalgebruik van de patiënt. Spontane taal wordt echter weinig onderzocht omdat het voeren van een gesprek weliswaar hooguit een kwartier kost, maar de analyse ervan kan enkele uren duren. Bovendien is voor een betrouwbare analyse minimaal een sample van 300 woorden nodig en juist in de acute fase is het vaak moeilijk om dergelijke uitgebreide en lange gesprekken met patiënten te voeren.

Taalkundig gezien is spontane taal niettemin zeer interessant omdat verschillende niveaus van het taalgebruik van de patiënt onderzocht kunnen worden. Op lexicaal-semantisch niveau kunnen semantische fouten, het gebruik van lege woorden (wat duidt op woordvindingsmoeilijkheden) en het voorkomen van lexicale perseveraties (het regelmatig herhalen van het zelfde woord) worden beschouwd. Op fonologisch niveau wordt gelet op het aantal fonologische fouten en fonologische perseveraties (het regelmatig herhalen van dezelfde klankvorm) en op morfo-syntactisch niveau kunnen onder andere de zinsbouw, de woordvolgorde en de vervoegingen van werkwoorden worden onderzocht. Opvallend is echter dat Niewold zich beperkt tot het analyseren van de spontane taal op semantisch en fonologisch niveau en dat hij het morfo-syntactische niveau buiten beschouwing laat, terwijl juist voor dit niveau spontane-taalonderzoek een meerwaarde heeft ten opzichte van bijvoorbeeld benoemscores. Dit is een gemis, vooral ook omdat er sprake kan zijn van interactie tussen de verschillende niveaus. Zo is bijvoorbeeld bekend dat de vervoeging van werkwoorden bij bepaalde patiënten ten koste kan gaan van de variatie in werkwoorden.

Niewold beschrijft in zijn proefschrift de data van 11 patiënten die twee keer zijn onderzocht in de acute fase en zo mogelijk ook twee keer in de chronische fase, waarbij vooral de spontane-taaldata uitgebreid geanalyseerd zijn. Hij laat hierbij statistisch onderbouwd zien dat er verschillende patronen van herstel zijn en dat lang niet alle onderzochte variabelen herstel laten zien. Zo waren er bij de meeste patiënten stoornissen te zien op de variabelen fonologische parafasieën, fonologische perseveraties, lexicale perseveraties en semantische fouten. Deze eerste drie foutencategorieën werden ook nog wel aangetroffen in de chronische fase, maar dat gold veel minder voor semantische fouten en – wanneer aanwezig in de acute fase – voor neologismen. Op de variabele lexicale perserveraties was zelfs vaker sprake van achteruitgang dan van herstel bij de over-

gang naar de chronische fase. In het algemeen concludeert Niewold dat bij de meeste patiënten, in tegenstelling tot de verwachtingen, er meer herstel plaatsvindt in de periode tussen de acute en de chronische fase dan in de acute fase zelf. Daarnaast bleek ook nog herstel plaats te vinden in de chronische fase.

Hoewel deze conclusie een andere kijk kan geven op spontaan herstel bij afasie moeten hierbij wel twee belangrijke kanttekeningen gemaakt worden. Allereerst beschrijft Niewold een aantal analyses waar vraagtekens bij gezet kunnen worden. Van elke patiënt geeft hij namelijk niet alleen zijn eigen analyses, maar ook de scores op de spontane taal van de diagnostische test die in veel gevallen bij afasiepatiënten wordt afgenomen, de Akense Afasie Test (AAT; Graetz, De Bleser en Willmes 1992). Bij de scoring van de AAT wordt rekening gehouden met variabelen als semantische structuur, fonologische structuur en geautomatiseerd taalgebruik (waaronder onder andere perseveraties vallen). De scores van Niewold wijken nogal eens af van deze AAT-scores. Zo is er een patiënt die een goede score haalt op de rubriek fonologische structuur bij de AAT terwijl deze in de analyse van Niewold veel fonematische parafasieën produceert. Een andere patiënt haalt de maximale score voor geautomatiseerd taalgebruik op de AAT, terwijl deze regelmatig perseveraties laat zien. Natuurlijk kan dit duiden op een onjuiste AAT-score, maar Niewold had zijn eigen analyses sterker kunnen maken door een tweede beoordelaar in te schakelen.

Een tweede belangrijke kanttekening betreft de vraag of er sprake is van *spontaan* herstel. Niewold beschrijft alleen zijn meetmomenten en niet wat er in de tussenliggende periodes met de patiënten is gebeurd. Hij merkt alleen in algemene termen op dat er sprake is geweest van logopedische behandeling. Feit is echter dat er bij patiënten al in een zeer vroege fase begonnen wordt met de behandeling en dat in het onderzoek van Niewold dus niet kan worden uitgesloten dat bepaalde vooruitgang het gevolg is van deze behandeling. Andersom is het ook vaak lastig om aan te tonen dat de behandeling wel tot de vooruitgang heeft geleid en dat maakt nu precies het onderzoeken van de taal in de niet-chronische fase zo lastig.

Voor Niewold zal dit laatste niet een zwaar punt zijn, gezien zijn herhaaldelijke kritische opmerkingen in zijn proefschrift over de bijdrage van therapie aan vooruitgang. Helaas blijft hij hierbij continu verwijzen naar Prins, die in 1987 concludeerde dat geen enkel onderzoek had aangetoond dat therapie tot vooruitgang in de spontane taal had geleid bij patiënten. Hoewel Prins op dat moment waarschijnlijk gelijk had, is er op dit moment veel minder reden voor een dergelijke conclusie. In de laatste 10 jaar zijn er regelmatig therapiestudies verschenen die hebben laten zien dat er bij patiënten in de chronische fase vooruitgang kan worden geboekt in de spontane taal als gevolg van therapie (bijvoorbeeld: Bastiaanse, Bosje & Franssen 1996; Pullvermuller, Neiniger, Elbert e.a. 2001; Bastiaanse, Hurkmans & Links 2006).

Concluderend kan ik zeggen dat Niewold onmiskenbaar een nuttige bijdrage heeft geleverd aan de kennis over afasie in de acute fase en het nut van spontane-taalonderzoek hierbij. De titel van zijn proefschrift – *Spontaan herstel van afasie in en na de acute fase* – belooft echter meer dan Niewold kon waarmaken in deze studie.

Roel Jonkers
Afdeling Taalwetenschap, Rijksuniversiteit Groningen
r.jonkers@rug.nl

Bibliografie

Bastiaanse, R., M. Bosje & M. Franssen (1996). Deficit-oriented treatment of word-finding problems: another replication. *Aphasiology* 10-4, 363-383

Bastiaanse, R., J. Hurkmans & P. Links, P. (2006). The training of verb production in Broca's aphasia: A multiple-baseline across-

behaviouors study. *Aphasiology* 20-2/3/4, 298-311

Graetz, P., R. de Bleser K. & Willmes (1992). *Akense Afasie Test*. Lisse: Swets & Zeitlinger.

Prins, R.S. (1987). *Afasie: classificatie, behandeling en herstelverloop*. Academisch proefschrift Universiteit van Amsterdam, Amsterdam.

Pulvermuller, F., B. Neininger, T. Elbert e.a. (2001). Constraint-induced therapy of chronic aphasia after stroke. *Stroke* 32, 1621-1626.

Signalementen

Woordenboeken

In 2007 verscheen bij uitgeverij Van Dale de eerste druk van de volgende tweetalige woordenboeken: *Van Dale pocketwoordenboek Italiaans-Nederlands* en *Van Dale pocketwoordenboek Nederlands-Italiaans.* De boeken bevatten 40.000 actuele en relevante woordbetekenissen. De trefwoorden zijn snel herkenbaar doordat ze rood zijn afgedrukt. Bij Het Spectrum verscheen dit jaar het *Prisma woordenboek Fries*, dat bestaat uit een deel Fries-Nederlands en een deel Nederlands-Fries en ongeveer 15.000 trefwoorden bevat. Het gaat om een belangrijke uitgave, omdat het Fries de taal is van 650.000 inwoners van Friesland en een verplicht onderdeel vormt van het onderwijs op de basischool en in de basisvorming van het voortgezet onderwijs. In het vorige nummer van dit tijdschrift werd melding gemaakt van het dit jaar bij dezelfde uitgeverij gepubliceerde *Van Dale Elftalig woordenboek*, een praktische woordenschat in elf talen (Nederlands, Duits, Engels, Frans, Hongaars, Italiaans, Pools, Portugees, Spaans, Tsjechisch, Zweeds). Dit woordenboek vergemakkelijkt de grensoverschrijdende communicatie van alledag en van de beroepspraktijk. Het geeft in één oogopslag op de linker en rechter pagina de equivalenten van een woord in negen talen. Dit deel bevat 28.000 trefwoorden uit de actuele en praktische basiswoordenschat. Dit jaar verschenen er nog vier delen. Deze bevatten de vertalingen van medische termen, technische begrippen, termen uit de zakenwereld, en computertermen. Elk deel heeft ruim 25.000 trefwoorden.

Van de hand van Wim Daniëls verschenen in 2007 bij uitgeverij Adr. Heinen twee lexicons: *het Lachwoordenboek* en het *Groot dierenwoordenboek*. Het eerste werk pretendeert het eerste lachtwoordenboek ter wereld te zijn. Het bestaat uit vier alfabetisch geordende delen: lachwoorden, lachcombinaties, lachuitdrukkingen en lachcitaten. Het tweede werk is geschreven voor kinderen, die er veel dierennamen en dierenwoorden uit kunnen leren, maar ook iets over andere bijzonderheden die met dieren te maken hebben. Ook op kinderen gericht zijn de dit jaar bij Van Dale verschenen eerste drukken van het *Van Dale junior woordgeschiedenisboek*, geschreven door Wim Daniëls en getekend door Roger Klaassen, en het voorleeswoordenboek *Mijn tweede Van Dale*, geschreven door Martine Letterie en Betty Sluyzer, en getekend door Paula Gerritsen. Beide bijzonder mooi uitgegeven boeken richten zich duidelijk op verschillende leeftijden. Het eerste is bedoeld voor de leeftijdscategorie vanaf 10 jaar, terwijl het tweede werk zich richt op kinderen vanaf 4 jaar. Bijzonder mooi is ook het in 2007 bij uitgeverij BZZTôH gepubliceerde *Woordenboek van platte taal*, geschreven door Heidi Aalbrecht en Pyter Wagenaar en geïllustreerd door Jacques Tange. Het werk bevat woorden en uitdrukkingen over seks, drank, misdaad en dood, scheldwoorden, straattaal en Bargoens. Daarnaast zijn er leeskaders met betekenisverwante woorden.

Dit jaar publiceerde Van Dale de eerste druk van het Van Dale modern gevleugeldewoordenboek *Never a dull moment en 2499 andere gevleugelde woorden*. Het boek werd samengesteld door Ton den Boon.

Ten slotte twee delen uit het systematische *Woordenboek van de Gelderse dialecten*, die in 2006 werden gepubliceerd bij Matrijs. Het deel dat door Harrie Scholtmeijer werd geschreven, bevat de woordenschat van het thema 'de mens' en heeft betrekking op het Veluwegebied. Het ande-

re deel is van de hand van Charlotte Giesbers en heeft hetzelfde thema, maar dan voor het rivierengebied.

Boeken over meertaligheid

Bij de uitgever Multilingual Matters verscheen dit jaar en vorig jaar een groot aantal boeken op het gebied van meertaligheid. Eerst noemen we echter het bij Blackwell verschenen *The handbook of bilingalism*, onder redactie van Tei K. Bhatia en William C. Ritchie. In dit werk van bijna negenhonderd pagina's worden alle mogelijke aspecten van tweetaligheid onder de loep genomen: de methodologische kanten van het onderzoek, neurologische en psychologische aspecten van tweetaligheid, taalverwerving, sociolinguïstische onderwerpen. Daarnaast zijn er honderden pagina's ingeruimd voor een aantal case studies. Een van de bijdragen in deze bundel richt zich op de mixing die voorkomt bij tweetaligen die het Engels en American Sign Language beheersen.

Dan de boeken die bij Multilingual Matters het levenslicht zagen. In *Linguistic Landscape*, onder redactie van Durk Gorter, vinden we een bundeling van zes bijdragen, die zich richten op de meertaligheid in verschillende steden van de wereld. In deze studies is ook aandacht voor het process van globalisering dat optreedt door het gebruik van het Engels.

In *Cross-linguistic influences in the second language lexicon* gaat het over de verschillende aspecten van taaltransfer, die hier worden beschouwd als resultaat van taalcontact bij tweetalige en meertalige verwerving. Het accent in de bijdragen ligt voornamelijk op de lexicale structuren van transfer.

Cross-linguistic similarity in foreign language lerarning van Hakan Ringbom is een boek dat het belang laat zien van cross-linguïstische overeenkomsten bij vreemdetaalverwerving. De nadruk ligt op de essentiële rol die positieve transfer in dit proces speelt.

In studies naar tweedetaalverwerving in natuurlijke settings komt naar voren dat kinderen betere tweedetaalverwervers zijn dan adolescenten en volwassenen. In het boek *Age and the rate of foreign language learning* wordt uitgebreid ingegaan op de effecten van leeftijd op de verwerving van een tweede taal in een gestuurde situatie (school setting).

Studies of fossilization in second language acquisition is een verzameling bijdragen waarin verslag wordt gedaan van het meest recente onderzoek naar fossilisatie, het verschijnsel waarbij een tweedetaalverwerver in een tussentaal is blijven steken.

In *Early trilingualism. A focus on questions* gebruikt de auteur Julia D. Barns als voorbeeld een drietalig kind, dat naast het Baskisch en Spaans ook op hoog niveau het Engels leert beheersen. Belangrijke verschillen met monolinguale en bilinguale verwerving die worden geconstateerd, liggen op het terrein van de cross-linguïstische invloeden vanuit het Baskisch en Spaans op het Engels. Een belangrijk gezichtspunt van de auteur is dat meertaligheid niet moet worden gezien als een verschijnsel waarbij twee of drie talen zijn samengevoegd, maar als een fenomeen waarbij sprake is van een persoonlijke multicompetente kennis, die niet kan worden gemeten in termen van monolinguale maatstaven.

Second language lexical processes is een bundel met studies waarin de psycholinguïstische aspecten van tweedetaalverwerving centaal staan. Daarbij spelen vragen een rol op het gebied van het aantal lexicons van de tweedetaalverwerver, de wijze waarop het lexicon is georganiseerd, verwerving van het lexicon, enzovoort. Daarbij wordt ook beschreven op welke manier gebruik wordt gemaakt van verschillende psycholinguïstische experimenten.

Doofheid en slechthorendheid

In 2006 verscheen bij uitgeverij Bert Bakker een boek van Stuart Blume, getiteld *Grenzen aan genezen*, waarin wetenschap, technologie en de doofheid van een kind centraal staan. Blume,

die als hoogleraar wetenschapsdynamica verbonden is aan de UvA, mag als ervaringsdeskundige spreken. Als Blume en zijn vrouw erachter komen dat hun zoontje doof is, worden zij opeens geconfronteerd met de vragen waarmee hij zich als wetenschapper al jaren bezighoudt: wat zijn de (on)mogelijkheden van de medische wetenschap en hoe ver moet men gaan om onvolkomenheden van de natuur te voorkomen of te genezen? In dit boek geeft Blume een persoonlijk verslag van de moeite die hij heeft moeten doen om de doofheid van zijn zoontje te accepteren. In meer algemene context gaat hij in op de ethische kwesties die de toepassing van de moderne medische technieken met zich brengt.

Dove en slechthorende jongeren zijn regelmatig onderwerp van wetenschappelijk onderzoek, maar zelden zijn het hún ervaringen die in dat onderzoek centraal staan. Er is veel aandacht voor hun spraak- en taalontwikkeling en voor hun geestelijke gezondheid, maar hoe denken dove en slechthorende jongeren over zichzelf, hun ouders, hun school en hun gehoorverlies? In *Hoor hen! Participatieonderzoek: virtueel en in real life*, uitgegeven door Van Tricht, doet Jet Isarin verslag van de ervaringen van jongeren met een auditieve beperking. Om zo dicht mogelijk bij de ervaringen zelf te blijven, liet zij jongeren zelf actief meedoen in alle fasen van haar project en ze communiceerde met hen via MSN, het jongerenmedium bij uitstek. We hebben hier te maken met een zogenaamd omkeerboek: aan de ene kant vinden we het theoretische kader, aan de andere kant zien we het ervaringsdeel, waarin de ervaringen van de jongeren aan bod komen.

Bij Wolters-Noordhoff ten slotte werden de delen 3 en 4 van de taalmethode *Taal op maat* uitgegeven. Deze methode is bedoeld voor het tweetalig onderwijs aan dove en ernstig slechthorende kinderen. In de ene band van elk van beide delen staat het Nederlands centraal, in de andere de Nederlandse Gebarentaal.

Taal en hersenen

In het boek *Introduction to neurolinguistics*, in 2006 uitgegeven bij Benjamins, belicht Elisabeth Ahlsén verschillende kanten van de neurolinguïstiek. In het eerste deel staat de vraag centraal wat neurolinguïstiek is. Het tweede deel brengt de neurolinguïstiek in verband met de verschillende componenten van de grammatica. In het derde deel staat een aantal specifieke onderwerpen centraal, zoals schrijven en lezen vanuit een neurolinguïstisch perspectief, neurolinguïstische aspecten van tweetaligheiod, enzovoort. In het laatste deel wordt aandacht geschonken aan de methoden om het brein te onderzoeken, het ontwerpen van een model van het brein, enzovoort.

In de in 2006 bij LOT gepubliceerde dissertatie *Experiments on the modular nature of word and sentence phonology in Chinese Broca's patients* onderzoekt Liang Ji de effecten van hersenbeschadiging op de verwerking van prosodie (vooral toonhoogteverloop in spraak) door Chineestalige patiënten met Broca's afasie.

In de bundel *Afasie [z]onder* woorden, die in 2006 werd uitgegeven bij Garant onder redactie van Erik Robert en Peter Mariën, worden in veertien bijdragen de recenste ontwikkelingen op het gebied van de afasiediagnostiek en de afasietherapie besproken.

In *Hoe mijn vader zijn woorden terugvond*, in 2006 uitgegeven bij Nieuw Amsterdam, beschrijft Liesbeth Koenen de beangstigende lijdensweg van haar vader, die een beroerte heeft gehad en plotseling niet meer kan praten. Hij verstaat en begrijpt echter nog wel alles wat anderen zeggen. Koenen beschrijft het herstelproces van haar vader met veel gevoel, maar ook vraagt zij zich af wat er taalkundig aan de hand is: waarom haalt hij de t en de r door elkaar?, waarom maakt hij steeds verkleinwoorden?, enzovoort.

Het Oera Linda-boek

In 2006 verscheen bij uitgeverij Verloren een door Goffe Jensma bezorgde facsimile, transcriptie en vertaling van het zogenaamde Oera Linda-boek. Het is een bijzonder mooi uitgegeven werk over een met mysteries omgeven tekst. In 1867 kwam de Friese archivaris Eelco Verwijs door een tip van een Harlingse onderwijzer in contact met Cornelis over de Linden, een scheepstimmerbaas die een oeroud familiestuk in zijn bezit bleek te hebben, dat hij wel graag eens wilde laten onderzoeken. Uiteindelijk werd het handschrift door Johan Winkler bekeken, die het zeer verdacht vond. De classicus Jan Gerhardus Ottema was echter een andere mening toegedaan. Het ging volgens hem om een echt handschrift en de rest van zijn leven zou Ottema aan de studie van dit Oera Linda-boek besteden. Het handschrift beschrijft de geschiedenis van een duizenden jaren oude moreel hoogstaande Friese samenleving, waaruit de gehele West-Europese samenleving zou zijn voortgekomen. Terwijl Ottema ervan overtuigd was dat de Friezen de westerse beschaving hadden gesticht, pijnigden andere onderzoekers hun hersenen om te bedenken wie dit onechte boek zou hebben kunnen schrijven. In deze prachtige uitgave heeft Jensma het ongetwijfeld als zeer intelligente grap bedoelde boek de eer gegeven die het verdient: een mooie heruitgave met een uitvoerige studie erbij van het gebruikte papier, het schrift, de taal, de verhaalstructuur, enzovoort.

Calendarium Nederlandse taal

Het *Calendarium van de Nederlandse taal* van Nicoline Van der Sijs, dat in 2006 door de Sdu werd uitgegeven, is een fraai boekwerk over de geschiedenis van de Nederlandse taal. Het is het eerste calendarium op het gebied van de taal. Calendaria op het terrein van de geschiedenis waren er al wel gepubliceerd. De auteur beschrijft in haar calendarium in chronologische volgorde de externe geschiedenis van de Nederlandse taal. Daarbij gaat het om de belangrijkste taalvoorschriften (grammatica's, spellinggidsen, woordenboeken), de belangrijkse gebeurtenissen in de Lage Landen op cultureel, historisch en politiek gebied, en de gevolgen die deze hadden op de Nederlandse taal, de invloeden die het Nederlands heeft ondergaan door contact met andere talen, veranderingen in het onderwijs, enzovoort. In het boek wordt telkens expliciet verband gelegd tussen een bepaalde gebeurtenis en de gevolgen die deze voor de Nederlandse taal heeft gehad: de Romeinse overheersing, de komst van zigeuners en Joden, de walvisvaart, enzovoort.

Ludieke stellingen

De titel van het in 2006 bij de Sdu verschenen boekje *Science? Fiction!* is de tekst van een stelling waarvan een Utrechtse promovendus ooit zijn proefschrift vergezeld liet gaan. Jan Luitzen publiceerde bij de Sdu een boekje waarin ruim 800 ludieke stellingen zijn samengebracht, onderverdeeld in 22 categorieën. Wat al deze stellingen gemeen hebben, is dat ze grappig bedoeld zijn, maar waar de ene stelling een duidelijk standpunt inneemt, bijvoorbeeld tegen de homeopathie (Een huisarts die homeopatische geneesmiddelen voorschrijft, is als een astronoom die een horoscoop opstelt), zijn andere louter als grap bedoeld (Schaken is een van de meest briljante vormen van verveling; Vampiers kunnen niet veilig vrijen). Er zijn ook stellingen waarvan ik me afvraag waarom ze zijn opgenomen, omdat ik niet zie waarom ze grappig zouden zijn (Het door reklame stimuleren van roken van sigaretten dient wettelijk verboden en streng strafbaar gesteld te worden).

Modale uitdrukkingen

In het in 2006 door de VU uitgegeven boek *Modale uitdrukkingen in Belgisch-Nederlands en Nederlands-Nederlands: corpusonderzoek en enquête*, gaan Janneke Diepeveen e.a. in op het gebruiksverschil

in Nederland en België van de veelgebruikte modale werkwoorden *moeten, hoeven, kunnen, mogen, zullen* en *willen*. Daarnaast laat het boek zien dat er grote verschillen zijn in de gebruiksfrequentie van modale woorden. De resultaten zijn gebaseerd op onderzoek in het *Corpus Gesproken Nederlands* en op een enquête die werd gehouden onder studenten in Antwerpen, Gent, Hasselt, Amsterdam, Groningen en Nijmegen.

Bibliografie

Aalbrecht, Heidi & Pyter Wagenaar. *Woordenboek van platte taal*. 's-Gravenhage: Bzztôh, 2007. 336 blz. ISBN 90 453 0513 4. EUR 29,50.

Ahlsén, Elisabeth. *Introduction to neurolinguistics.* Amsterdam/Philadelphia: Benjamins, 2006. XII + 212 blz. ISBN 90 272 3234 2 EUR 33,00.

Arabski, Janusz (red.). *Cross-linguistic influences in the second language lexicon.* Clevedon enz.: Multilingual Matters, 2006. X + 272 blz. ISBN 1 85359 856 9. BP 49,95.

Barnes, Julia D. *Early trilingualism. A focus on questions.* Clevedon enz.: Multilingual Matters, 2006. VIII + 256 blz. ISBN 1 85359 854 2. BP 49,95.

Bhatia, Tej K. & William C. Ritchie. *The handbook of bilingualism.* Oxford: Blackwell, 2006. XVIII + 884 blz. ISBN 0 631 22735 0. BP 26,99.

Blume, Stuart. *Grenzen aan genezen. Over wetenschap, technologie en de doofheid van een kind*. Amsterdam: Bert Bakker, 2006. 384 blz. ISBN 90 351 3024 3. EUR 25,00.

Boon, Ton den. *Never a dull moment en 2499 andere gevleugelde woorden.* Utrecht/Antwerpen: Van Dale Lexicografie, 2007. 238 blz. ISBN 90 6648 068 1. EUR 21,95.

Daniëls, Wim. *Groot dierenwoordenboek*. 's-Hertogenbosch: Adr. Heinen, 2007. 152 blz. ISBN 9086800092. EUR 14,95.

Daniëls, Wim. *Lachwoordenboek*. 's-Hertogenbosch: Adr. Heinen, 2007. 168 blz. ISBN 9086800469. EUR 9,95.

Daniëls, Wim & Roger Klaassen. Van Dale Junior Woordgeschiedenisboek. Waar komen onze woorden vandaan? Utrecht/Antwerpen: Van Dale Lexicografie, 2007. 448 blz. ISBN 90 6648 092 6. EUR 27,95.

Diepenveen, Janneke e.a. *Modale uitdrukkingen in Belgisch-Nederlands en Nederlands-Nederlands: corpusonderzoek en enquête.* Amsterdam: Stichting Neerlandistiek VU, 2006. X + 145 blz. ISBN 90 72365 98 4. EUR 17,50.

Elftalig woordenboek. Business woordenschat in 11 talen. Utrecht/Antwerpen: Van Dale Lexicografie, 2007. 1135 blz. ISBN 90 6648 034 6. EUR 59,95.

Elftalig woordenboek. Computer woordenschat in 11 talen. Utrecht/Antwerpen: Van Dale Lexicografie, 2007. 1135 blz. ISBN 90 6648 033 9. EUR 59,95.

Elftalig woordenboek. Medische woordenschat in 11 talen. Utrecht/Antwerpen: Van Dale Lexicografie, 2007. 1135 blz. ISBN 90 6648 031 5. EUR 49,95.

Elftalig woordenboek. Technische woordenschat in 11 talen. Utrecht/Antwerpen: Van Dale Lexicografie, 2007. 1135 blz. ISBN 90 6648 032 2. EUR 59,95.

Gerritsen, Paula, Martine Letterie & Betty Sluyzer. *Mijn tweede Van Dale voorleeswoordenboek.* Utrecht/Antwerpen: Van Dale Lexicografie, 2007. 176 blz. ISBN 90 6648 091 9. EUR 21,95.

Giesbers, Charlotte. *Woordenboek van de Gelderse dialecten. Rivierengebied*. De mens. Utrecht: Matrijs, 2006. 248 blz. ISBN 90 5345 304 0. EUR 19,95

Gorter, Durk (red.). *Linguistic landscape: a new approach to multilingualism*. Clevedon enz.: Multilingual Matters, 2006. VI + 89 blz. ISBN 1 85359 916 6. BP 29,95 (Hb).

Han, ZhaoHong & Terence Odlin (red.). *Studies of fossilization in second language acquisition.* Clevedon enz.: Multilingual Matters, 2006. VIII + 214 blz. ISBN 1 85359 835 6. BP 24,95.

Isarin, Jet. *Hoor hen! Participatieonderzoek: virtueel en in real life.* Zwolle/ Twello: Effatha Guyotgroep/Van Tricht, 2006. 132 blz. ISBN 90 77822 14 3. EUR 24,75.

Jensma, Goffe (red.). *Het Oera Linda-boek.* Hilversum: Verloren, 2006. 448 blz. ISBN 90 6550 841 4. EUR 28,00.

Ji, Liang. *Experiments on the modular nature of word and sentence phonology in Chinese Broca's patients.* Utrecht: LOT, 2006. XIV + 189 blz. ISBN 90 78328 00 1. EUR 22,12.

Koenen, Liesbeth. *Hoe mijn vader zijn woorden terugvond.* Amsterdam: Nieuw Amsterdam, 2006. 128 blz. ISBN 90 468 0068 7. Geen prijsvermelding.

Lengyel, Zsolt & Judi Navracsics (red.). *Second language lexical processes. Applied linguistic and psycholinguistic perspectives.* Clevedon enz.: Multilingual Matters, 2007. XIV+ 191 blz. ISBN 1 85359 966 8. BP 19,95 (Pb).

Lo Cascio, Vincenzo & Elisabeth Nijpels (red.). *Van Dale pocketwoordenboek Nederlands-Italiaans.* Utrecht/Antwerpen: Van Dale Lexicografie, 2007. 505 blz. ISBN 90 6648 785 7. EUR 9,95.

Lo Cascio, Vincenzo & Elisabeth Nijpels (red.). *Van Dale pocketwoordenboek Italiaans-Nederlands.* Utrecht/Antwerpen: Van Dale Lexicografie, 2007. 505 blz. ISBN 90 6648 783 3. EUR 9,95.

Luitzen, Jan. *Science? Fiction! En 808 andere ludieke stellingen van wetenschappers in spe.* Den Haag: Sdu, 2006. 82 blz. ISBN 90 12 11680 5. EUR 12,95.

Muñoz, Carmen (red.) *Age and the rate of foreign language learning.* Clevedon enz.: Multilingual Matters, 2006. XII + 283 blz. ISBN 1 85359 891 7. BP 24,95.

Ringbom, Hakan. *Cross-linguistic similarity in foreign language learning.* Clevedon enz.: Multilingual Matters, 2007. VIII + 144 blz. ISBN 1 85359 934 7. BP 19,95.

Robert, Erik & Peter Mariën (red.). *Afasie [z]onder woorden.* Antwerpen/Apeldoorn: Garant, 2006. 329 blz. ISBN 90 441 1865 0. EUR 29,90.

Scholtmeijer, Harrie. *Woordenboek van de Gelderse dialecten. Veluwe.* De mens. Utrecht: Matrijs, 2006. 304 blz. ISBN 90 5345 305 9. EUR 19,95.

Sijs, Nicoline van der. *Calendarium van de Nederlandse taal. De geschiedenis van het Nederlands in jaartallen.* Den Haag: Sdu, 2006. 338 blz. ISBN 90 12 11737 2. EUR 25,95.

Spoelstra, Janneke, Jantsje Post & Arjan Hut. *Prisma woordenboek Fries.* Utrecht: Het Spectrum, 2007. 592 blz. ISBN 90 274 2526 3. EUR 8,50.

Taal op maat. Taalmethode voor het tweetalig onderwijs aan dove en ernstig slechthorende kinderen. Groningen/Houten: Wolters-Noordhoff, 2006. Deel 3 Nederlands (NL). 162 blz. ISBN 90 809512 3 4; Deel 4 Nederlands (NL). 331 blz. ISBN 90 809512 5 0; Deel 3 Nederlandse Gebarentaal (NGT). 124 blz. ISBN 90 809512 4 2; Deel 4 Nederlandse Gebarentaal (NGT). 245 blz. ISBN 90 809512 6 9.

UIT DE TIJDSCHRIFTEN

De rubriek *Uit de tijdschriften* geeft kort weer wat er in andere tijdschriften op het gebied van de taalkunde is verschenen. Momenteel worden in deze rubriek de volgende tijdschriften besproken: *Anéla, Neerlandia-Nederlands van Nu, Neerlandica Extra Muros, Ons Erfdeel, Onze Taal, Over Taal, Taal en Tongval, TaalActief, Taalkundig Bulletin, VDW-berichten*. Uitgevers van niet vermelde taalkundige periodieken die hun tijdschrift besproken willen zien, wordt verzocht contact op te nemen met de redacteur van deze rubriek:

Dr. J. Nijen Twilhaar, Oerdijk 35, 7433 AG Schalkhaar, tel.: 0570-608080, e-mail: jnt@telebyte.nl

Neerlandica Extra Muros
45 (2007), nr. 2

In dit nummer vinden we onder meer een artikel van Rudi Janssens over het Nederlands in het Brusselse onderwijslandschap. Esther Ruigendijk brengt in haar bijdrage verslag uit van een vergelijkend onderzoek naar de problemen met voornaamwoorden en determinatoren van Nederlands- en Duitstalige sprekers met afasie. Matthias Hüning geeft een kroniek van de taalkunde 2005-2006. Alice van Kalsbeek doet hetzelfde voor het Nederlands voor anderstaligen

Verder zijn er boekbesprekingen in de rubriek *Besprekingen en aankondigingen*.

Ons Erfdeel
50 (2007), nr. 2

In de tweede aflevering van deze jaargang staat onder meer een artikel van Siegfried Theissen over de neerlandistiek in Franstalig België en in Frankrijk.

Onze Taal
76 (2007), nr. 5

De vierde aflevering van dit tijdschrift bevat een aantal bijdragen die gewijd zijn aan het thema 'de taal van de politie'. Berthold van Maris gaat in zijn artikel in op de etymologische speculaties van de 16de-eeuwse geleerde Becanus, die net als Willem Hietbrink van mening is dat het Nederlands het dichtst bij de 'oertaal' ligt. Joop van der Horst bespreekt een nieuwe woordvolgorde in de Nederlandse taal ('dat we mensen het lastig maken'). Ton den Boon levert voor het laatst een bijdrage aan zijn rubriek *Merken worden namen* met een stuk over 'borrelnootjes'. Nicoline van der Sijs gaat in haar rubriek in op valse eponiemen. De bijdrage van Rik Schutz gaat over versterkende vergelijkingen in het Nederlands, zoals 'poeslief' en 'brutaal als de beul'.

Daarnaast zijn er de kleinere bijdragen en de vaste rubrieken.

nr. 6

In het juni-nummer van deze jaargang staan enkele bijdragen die 'zorgen over taalvaardigheid studenten' als thema hebben. Koen Jaspaert en Sjaak Kroon voeren discussie met Marc van Oostendorp over streektaalpolitiek. De bijdrage van Ewoud Sanders gaat over het Bargoens. Het artikel van Jaap de Jong is gewijd aan het tijdschrift *Onze Taal* in de Tweede Wereldoorlog.

Daarnaast zijn er de kleinere bijdragen en de vaste rubrieken.

nr. 7/8

Deze dubbele aflevering begint met een artikel van Jan Erik Grezel over de taal van datingsites. René Appel had een interview met de schrijver Adriaan van Dis. Jan Kuitenbrouwer laat in zijn bijdrage zien dat het streven van *Onze Taal* om er voor niet-taalkundigen te zijn, soms botst met het verlangen van dit tijdschrift om deze doelgroep op taalkundig verantwoorde wijze te bedienen. Matthijs de Winter deed onderzoek naar de taal van de kampeerder.

Daarnaast zijn er de kleinere bijdragen en de vaste rubrieken.

Over Taal
46 (2007), nr. 2

Het tweede nummer van deze jaargang begint met een interview dat Veronique de Tier had met Piet van Sterkenburg. Verder is er onder meer een artikel van Bert le Bruyn over verbleekte partitieve constructies.

Daarnaast zijn er bijdragen in de vaste rubrieken.

nr. 3

In deze derde aflevering vinden we een artikel van Gert De Sutter over de groene en rode volgorde. Albert Oosterhof gaat in op de idiomaticiteit van spreekwoorden. Stephanie Nijst en Ils Sterckx hadden een interview met Magda Devos over het dialect van Vlaanderen. De bijdrage van Dirk Caluwé gaat over het verleden en de toekomst van de tussen-n.

Daarnaast zijn er bijdragen in de vaste rubrieken.

Ontvangen boeken

Aalbrecht, Heidi & Pyter Wagenaar. Woordenboek van platte taal. 's-Gravenhage: Bzztôh, 2007. 336 blz. ISBN 90 453 0513 4. EUR 29,50.

Backhaus, Peter. *Linguistic landscapes. A comparative study of urban multilingualism in Tokyo.* Clevedon enz.: Multilingual Matters, 2007. X + 158 blz. ISBN 1 85359 946 0. BP 19,95 (Pb).

Balk-Smit Duyzentkunst, Frida. *Grammatica voor iedereen.* Den Haag: Sdu, 2007. 244 blz. ISBN 90 12 077 7. EUR 14,10.

Boon, Ton den. *Never a dull moment en 2499 andere gevleugelde woorden.* Utrecht/Antwerpen: Van Dale Lexicografie, 2007. 238 blz. ISBN 90 6648 068 1. EUR 21,95.

Braet, Antoine. *Retorische kritiek. Overtuigingskracht van Cicero tot Balkenende.* Den Haag: Sdu, 2007. 203 blz. ISBN 90 12 11956 6. EUR 28,25.

Bree, Elise de. *Dyslexia and phonology. A study of the phonological abilities of Dutch children at-risk of dyslexia.* Utrecht: LOT, 2007. X + 195 blz. ISBN 90 78328 26 1. EUR 21,05.

Capelletti, Matteo. *Parsing with structure-preserving categorial grammars.* Utrecht: LOT, 2007. IV + 167 blz. ISBN 90 78328 33 9. EUR 19,81.

Carnie, Andrew. *Syntax. A generative introduction.* Oxford enz.: Blackwell, 2007. XVIII + 489 blz. ISBN 978 1 4051 3384 5. BP 22,99.

Daniëls, Wim. *Groot dierenwoordenboek.* 's-Hertogenbosch: Adr. Heinen, 2007. 152 blz. ISBN 9086800092. EUR 14,95.

Daniëls, Wim. *Lachwoordenboek.* 's-Hertogenbosch: Adr. Heinen, 2007. 168 blz. ISBN 9086800469. EUR 9,95.

Daniëls, Wim & Roger Klaassen. Van Dale Junior Woordgeschiedenisboek. Waar komen onze woorden vandaan? Utrecht/Antwerpen: Van Dale Lexicografie, 2007. 448 blz. ISBN 90 6648 092 6. EUR 27,95.

Driel, Joost van. *Prikkeling der zinnen. De stilistische diversiteit van de Middelnederlandse epische poëzie.* Zutphen: Walburg Pers, 2007. 256 blz. ISBN 90 5730 477 5. EUR 39,50.

Drop je lyrics 2. Utrecht: Het Spectrum, 2007. 184 blz. ISBN 90 274 5728 8. EUR 5,00.

Feng, Anwei (red.). *Bilingual education in China. Practices, policies and concepts.* Clevedon enz.: Multilingual Matters, 2007. XIV+ 288 blz. ISBN 1 85359 991 0. BP 29,95 (Pb).

Gerritsen, Paula, Martine Letterie & Betty Sluyzer. *Mijn tweede Van Dale voorleeswoordenboek.* Utrecht/Antwerpen: Van Dale Lexicografie, 2007. 176 blz. ISBN 90 6648 091 9. EUR 21,95.

Giesbers, Charlotte. *Woordenboek van de Gelderse dialecten. Rivierengebied.* De mens. Utrecht: Matrijs, 2006. 248 blz. ISBN 90 5345 304 0. EUR 19,95

Gorter, Durk (red.). *Linguistic landscape: a new approach to multilingualism.* Clevedon enz.: Multilingual Matters, 2006. VI + 89 blz. ISBN 1 85359 916 6. BP 29,95 (Hb).

Hüning, Matthias e.a. (red.). *Nederlands tussen Duits en Engels. Handelingen van de workshop op*

30 september en 1 oktober 2005 aan de Freie Universität Berlin. Leiden: SNI, 2006. 291 blz. ISBN 90 78531 02 9. Geen prijsopgave.

Kamalski, Judith. *Coherence marking, comprehension and persuasion. On the processing and representation of discourse.* Utrecht: LOT, 2007. 227 blz. ISBN 90 78328 29 2. EUR 22,03.

Koenen, Mieke. *Inleiding tot de Latijnse syntaxis. Structuur van zin en tekst. Oefenboek.* Amsterdam: AUP, 2007. 128 blz. ISBN 90 5356 930 6. EUR 14,90.

Kratochvíl, František. *A grammar of Abui. A Papuan language of Alor.* 2 delen. Utrecht: LOT, 2007. XX + 541 blz. ISBN 90 78328 28 5. EUR 36,99.

Kroon, Caroline. *Inleiding tot de Latijnse syntaxis. Structuur van zin en tekst. Grammaticaboek.* Amsterdam: AUP, 2007. 176 blz. ISBN 90 5356 950 4. EUR 24,50.

Layton, Bentley. *Coptic in 20 lessons. Introduction to Sahidic Coptic with exercices & vocabularies.* Leuven enz.: Peeters, 2007. 204 blz. ISBN 90 429 1810 8. EUR 27.

Leijten, Mariëlle. *Writing and speech recognition. Observing error correction strategies of professional writers.* Utrecht: LOT, 2007. XII + 301 blz. ISBN 90 78328 31 5. EUR 25,91.

Lengyel, Zsolt & Judi Navracsics (red.). *Second language lexical processes. Applied linguistic and psycholinguistic perspectives*. Clevedon enz.: Multilingual Matters, 2007. XIV+ 191 blz. ISBN 1 85359 966 8. BP 19,95 (Pb).

Lo Cascio, Vincenzo & Elisabeth Nijpels (red.). *Van Dale pocketwoordenboek Italiaans-Nederlands.* Utrecht/Antwerpen: Van Dale Lexicografie, 2007. 505 blz. ISBN 90 6648 785 7. EUR 9,95.

Lo Cascio, Vincenzo & Elisabeth Nijpels (red.). *Van Dale pocketwoordenboek Nederlands-Italiaans.* Utrecht/Antwerpen: Van Dale Lexicografie, 2007. 505 blz. ISBN 90 6648 783 3. EUR 9,95.

Low Countries 2007, The. Rekkem: Ons Erfdeel, 2007. 319 blz. ISBN 90 75862 90 4 EUR 39.

Nyst, Victoria. *A descriptive analysis of Adamorobe Sign Language (Ghana).* Utrecht: LOT, 2007. 243 blz. ISBN 90 78328 22 3. EUR 23,86.

Pluymaekers, Mark. *Affix reduction in spoken Dutch. Probabilistic effects in production and perception.* Nijmegen: MPI Series in Psycholinguistics. 2007. 155 blz. ISBN 90 76203 28 7. Geen prijsvermelding.

Sadeh-Leicht, Oren. *The psychological reality of grammar. The theta principle in parsing performance..* Utrecht: LOT, 2007. X + 207 blz. ISBN 90 78328 23 0. EUR 21,59.

Scholtmeijer, Harrie. *Woordenboek van de Gelderse dialecten. Veluwe*. De mens. Utrecht: Matrijs, 2006. 304 blz. ISBN 90 5345 305 9. EUR 19,95.

Slioussar, Natalia. *Grammar and information structure. A study with reference to Russian.* Utrecht: LOT, 2007. X + 245 blz. ISBN 90 78328 32 2. EUR 23,28.

Snyder, William. *Child language. The parametric approach.* Oxford enz., Oford University press, 2007. XII + 209 blz. ISBN 0 19 929669 9. BP 55.

Van Dale Spellingcorrector Nederlands. Utrecht/Antwerpen: Van Dale Lexicografie, 2007. EUR 62,50.

Spoelstra, Janneke, Jantsje Post & Arjan Hut. *Prisma woordenboek Fries.* Utrecht: Het Spectrum, 2007. 592 blz. ISBN 90 274 2526 3. EUR 8,50.

Tjia, Johnny. *A grammar of Mualang. An Ibanic language of Western Kalimantan, Indonesia.* Utrecht: LOT, 2007. XVIII + 439 blz. ISBN 90 78328 25 4. EUR 32,27.

Van Dale elftalig woordenboek. Business woordenschat in 11 talen. Utrecht/Antwerpen: Van Dale Lexicografie, 2007. 1135 blz. ISBN 90 6648 034 6. EUR 59,95.

Van Dale elftalig woordenboek. Computer woordenschat in 11 talen. Utrecht/Antwerpen: Van Dale Lexicografie, 2007. 1135 blz. ISBN 90 6648 033 9. EUR 59,95.

Van Dale elftalig woordenboek. Medische woordenschat in 11 talen. Utrecht/Antwerpen: Van Dale Lexicografie, 2007. 1135 blz. ISBN 90 6648 031 5. EUR 49,95.

Van Dale elftalig woordenboek. Technische woordenschat in 11 talen. Utrecht/Antwerpen: Van Dale Lexicografie, 2007. 1135 blz. ISBN 90 6648 032 2. EUR 59,95.

Van Dale groot 4-talig beeldwoordenboek. Utrecht/Antwerpen: Van Dale Lexicografie, 2007. VI + 544 blz. ISBN 90 6648 977 6. EUR 44,95.

Vicente, Luis. *The syntax of heads and phrases. A study of verb (phrase) fronting.* Utrecht: LOT, 2007. VIII + 267 blz. ISBN 90 78328 25 4. EUR 24,44.

Vlachou, Evangelia. *Free choice in and out of context: semantics and distribution of French, Greek and English free choice items.* Utrecht: LOT, 2007. XVIII + 383 blz. ISBN 90 78328 26 1. EUR 29,78.

Wang, Hongyan. *English as a lingua franca. Mutual intelligibility of Chinese, Dutch ans American speakers of English.* Utrecht: LOT, 2007. XIV + 259 blz. ISBN 90 78328 20 9. EUR 24,08.

Wijk, Judith van. *The acquisition of the Dutch plural.* Utrecht: LOT, 2007. X + 318 blz. ISBN 90 78328 21 6. EUR 27,37.