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).

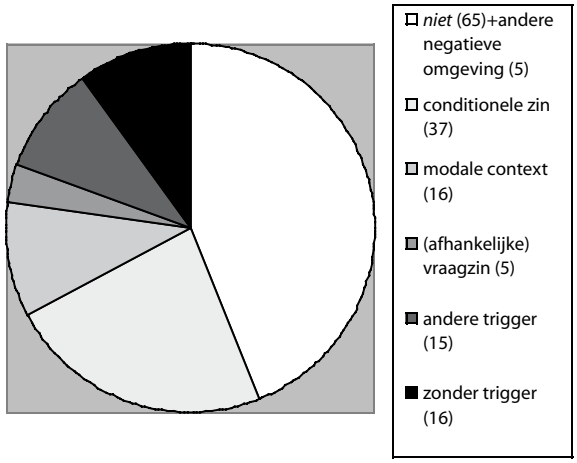
10 In (i) is een voorbeeld gegeven van een contrast dat samenhangt met registervariatie. Zinnen als (ia), waarin een definiet enkelvoud gebruikt is in een karakteriserende zin (i.e. een zin die een generalisatie uitdrukt die van toepassing is op tijgers in het algemeen), zijn gebruikelijker in formele 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]

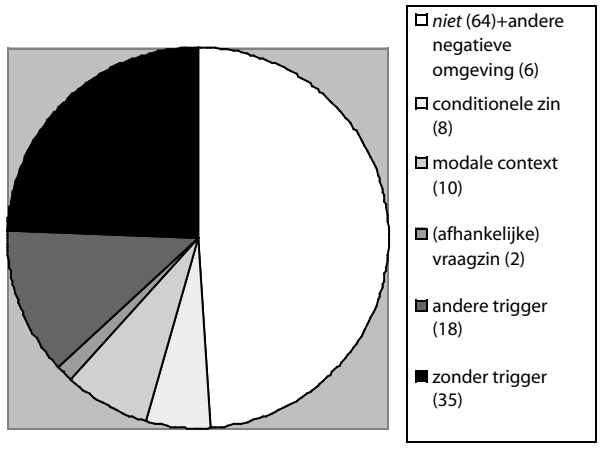
(Voorbeeld (ib) is afkomstig uit een spontane conversatie uit 2000 (Nederland).)

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]

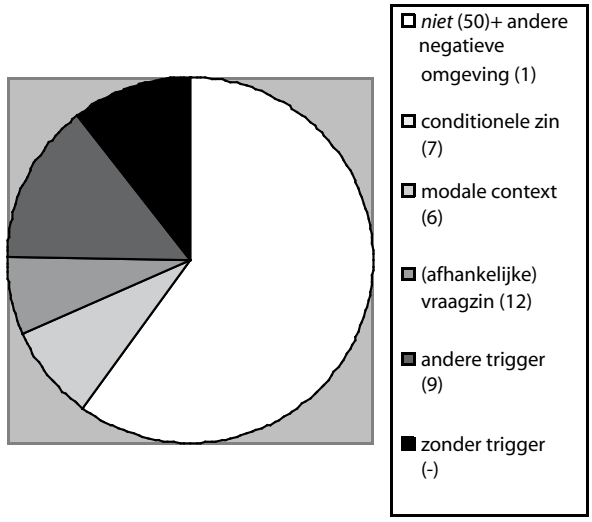
(7) A: 'k zal eerst eens informeren voor uh die langlauftocht. als die doorgaat
B: uhu
A: uhm… klopt dat doorgaan? plaats… ik mis altijd. als ie doorgaat. als ie plaatsvindt. plaatsheeft. [fv400320]

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).

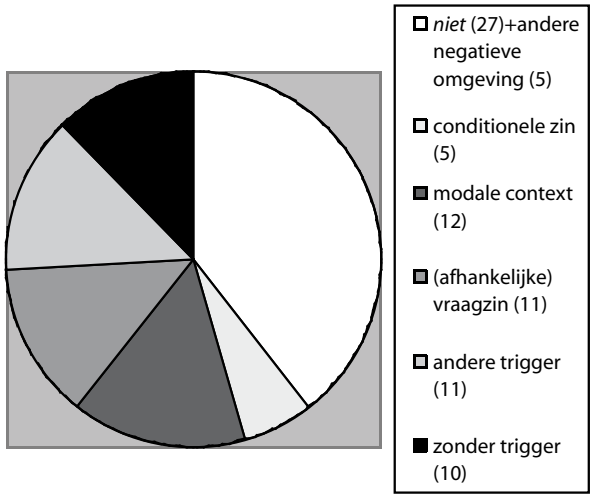
12 Het voorbeeld komt uit een spontane conversatie uit 2001 (Vlaanderen).

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.

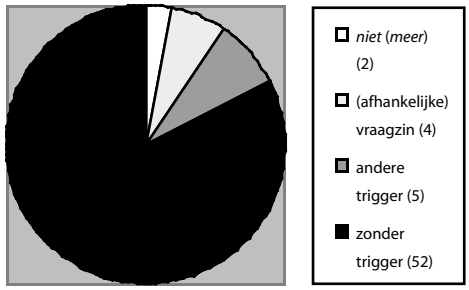
	internet (met Google)	chatmateriaal CONDIV	CGN	N
voedsel	36	3	1	40
personen	46	14	3	63
totaal	82	17	4	103

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).



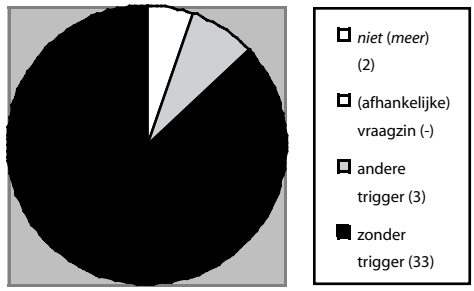
Figuur 6: Omgevingen van zot van, *waarbij het complement verwijst naar personen (N=63).*

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]

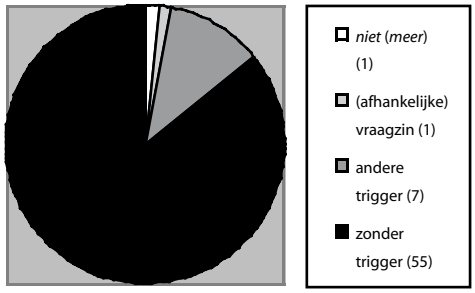
In de figuren 7 en 8 worden de resultaten van een (kleinschalig) corpusonderzoek naar de omgevingen van *gek zijn op* gepresenteerd.



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 enkelvoud		definiet meervoud		kaal meervoud	
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.