Samenvatting

Een corpus is 1) een verzameling documenten of 2) [taalk.] een begrensde verzameling teksten voor linguïstisch onderzoek (Van Dale online woordenboek). Het onderzoek in dit proefschrift richt zich op die tweede betekenis van corpora. In tegenstelling tot een corpus is taal zelf onbegrensd: met een eindig aantal bouwblokken kunnen oneindig veel taaluitingen gemaakt worden. Sommige taalkundigen hebben dan ook hun twijfels over het nut van corpora voor taalkundig onderzoek. Dit proefschrift laat zien dat corpora, ondanks de beperkingen die zij hebben, toch veel bij kunnen dragen aan allerlei soorten taalkundig onderzoek, variërend van theoretische taalkunde tot het automatisch leren van de lexicale eigenschappen van woorden.

Corpora zijn tegenwoordig veelal elektronisch. Dat maakt het mogelijk om met behulp van computerprogramma's interessante taalkundige informatie uit de bestandenverzameling te halen. En hoewel corpora per definitie begrensd zijn, neemt de omvang van de beschikbare elektronische corpora nog steeds toe. Eén jaargang krantentekst is al snel zo'n 17 miljoen woorden. En dan is er nog het web, met een geschatte 11,5 miljard pagina's eind januari 2005 (Gulli and Signorini, 2005) het ultieme corpus.

Het nut van digitale corpora voor taalkundig onderzoek kan nog vergroot worden door de tekst te verrijken met taalkundige meta-informatie. Zo kunnen de woorden voorzien worden van woordsoortlabels (*part-of-speech-tags*), kunnen de grenzen van woordgroepen toegevoegd worden (*chunken*) en kunnen de grammaticale relaties tussen die verschillende woordgroepen aangeduid worden (*parsing*). Deze verrijking van een corpus maakt het mogelijk om naar abstracte taalkundige patronen te zoeken. Zo kunnen bijvoorbeeld passieve zinnen uit een corpus gehaald worden door te zoeken naar zinnen waarin het onderwerp van *worden* overeenkomt met het lijdend voorwerp van het hoofdwerkwoord. Helaas is het handmatig annoteren van tekst heel tijdrovend. Handgeannoteerde corpora zijn dan ook beperkt in omvang. Maar voor onderzoek naar sommige (infrequente) constructies zijn juist heel grote corpora nodig. In dat geval kan een automatisch geannoteerd corpus uitkomst bieden. Hoewel automatische annotatie fouten bevat, blijkt

de meta-informatie toch van nut: in dit proefschrift worden vier uiteenlopende onderwerpen uit de grammatica van het Nederlands behandeld, waarbij automatisch verrijkte corpora telkens een andere rol spelen.

Het eerste onderwerp betreft de gekloofde zin. Dit zijn zinnen zoals in (1)-(2), die gebruikt worden om een bepaald zinsdeel te benadrukken (het meest benadrukte woord is in hoofdletters gedrukt). Ze bestaan uit het voornaamwoord *het* (soms *dit* of *dat*), het werkwoord *zijn*, de benadrukte woordgroep en een ondergeschikte bijzin. De meeste analyses van dit type zinnen—dat overigens in heel veel talen voorkomt—gaan ervan uit dat de zinnen (1) en (2) voorbeelden van een en dezelfde constructie zijn. Aangetoond wordt dat dit in ieder geval voor het Nederlands niet het geval is.

(1) Het is immers niet de TRAINER die kansen voor open doel verknalt.

(2) Het was op ZIJN aandringen, dat ik de redactie van de adviesaanvrage [...] zo heb veranderd.

Enkele argumenten voor het onderscheid tussen deze twee typen: in zinnen zoals (1) is de bijzin altijd een relatieve bijzin, terwijl het in zinnen zoals (2) een onderschikkende bijzin is met het voegwoord *dat*. Het benadrukte zinsdeel verschilt ook: de eerste constructie benadrukt alleen NP's, terwijl de tweede constructie allerlei woordgroepen kan benadrukken. Verder is er een verschil in het voornaamwoord in de beide zinnen. In het zinnen van het eerste type is het niet expletief, en in zinnen van het tweede type wel. Dat blijkt onder meer uit het feit dat de eerste in het corpus ook voorkomt met een demonstratief voornaamwoord *dat* of *dit* in plaats van *het*, de tweede type niet (en geconstrueerde voorbeelden bleken ongrammaticaal). Tenslotte kunnen in zinnen zoals (1) ook andere koppelwerkwoorden dan *zijn* voorkomen, maar in het type (2) niet.

Gekloofde zinnen van het type (1) worden geanalyseerd als koppelwerkwoordzinnen. Het onderwerp *het* en de relatieve bijzin vormen het onderwerp, en de benadrukte woordgroep is het predikaat. Het type (2) daarentegen heeft slechts één syntaktisch argument: het onderwerp, bestaande uit de bijzin plus de benadrukte woorgroep (*het* is in dit geval semantisch leeg). Deze analyse verklaart de verschillen tussen de beide typen en de schijnbare incongruentie tussen het onderwerp en de persoonsvorm én is conform de algemeen veronderstelde regels voor woordvolgorde in het Nederlands, in tegenstelling tot sommige eerdere analyses. De analyse wordt geformaliseerd binnen het theoretisch kader van *Lexical Functional Grammar* (LFG).

In dit hoofdstuk leveren corpora voorbeelden (om de eigen analyse te onderbouwen en de taalkundige eigenschappen van de constructie te illustreren) en tegenvoorbeelden (om de tekortkomingen van alternatieve analyses aan

te tonen). Omdat de constructie laagfrequent is, moet een groot corpus gebruikt worden, en omdat de constructie alleen aan de grammaticale rollen van de woordgroepen te herkennen is, moet een syntactisch geannoteerd corpus gebruikt worden. Om deze redenen zijn automatisch geannoteerde corpora gebruikt in aanvulling op handmatig geannoteerde corpora.

Het tweede onderwerp is de meewerkend-voorwerpconstructie. Het meewerkend voorwerp kan in het Nederlands, net als in veel andere talen, gerealiseerd worden als een zelfstandig-naamwoordgroep (NP) (3) of als een voorzetselgroep (PP) (4).

(3) Heeft hij je dat niet verteld?

(4) Als de speaker die treffer abusievelijk aan Amokachi toekent, grijpt hulptrainer Jo Bonfrère in.

Er zijn 2 typen analyses van deze constructie in het Engels. Het eerste formuleert voorkeuren voor bepaalde woordgroepsoorten ("werkwoorden die een manier van communiceren uitdrukken krijgen een prepositioneel meewerkend voorwerp"), het tweede maakt gebruik van algemene orderingsprincipes ("korte zinsdelen voor lange"). In combinatie met de strikte Engelse woordvolgorde, die dicteert dat een naamwoordelijk meewerkend voorwerp vóór het lijdend voorwerp komt, maar een PP erná, leiden die algemene principes tot de keuze voor een NP (korte meewerkende voorwerpen) of een PP (lange meewerkende voorwerpen). In het Nederlands is de woordvolgorde minder strikt dan in het Engels, en hoeft de PP niet altijd achteraan te staan. De hypothese is dan ook dat algemene orderingsprincipes geen invloed hebben op de keuze voor NP of PP, maar alleen op de volgorde van de beide complementen. De corpusdata bevestigen deze hypothese op sommige punten, maar laten een niet voorspelde invloed van pronominaliteit op de woordgroepsoort zien. Bovendien wordt aangetoond dat gewicht, tegen de verwachting in, de volgorde van de argumenten in het middenveld niet beïnvloedt. Wél hebben zware voorwerpen een voorkeur voor extrapositie.

Hoewel de regels voor de meewerkend-voorwerpconstructie vaak als categoriaal worden gezien, blijken veel contrasten niet zwart-wit te zijn. Zo hebben veel werkwoorden een voorkeur voor een PP of een NP, maar zelden is het alternatief echt onmogelijk. Voor onderzoek naar dergelijke verschijnselen is corpusmateriaal onontbeerlijk. Met behulp van corpora kunnen verschillen in de frequenties van twee constructies gemeten worden, en kunnen bovendien contextfactoren geïdentificeerd worden die op deze frequentie van invloed zijn. Om deze invloeden vervolgens te modeleren, zijn de gebruikelijke taalmodellen op basis van absolute regels of constraints niet toereikend. Optimality Theory, dat een stochastische implementatie kent, is hiervoor beter

geschikt. Met een vaste ordening van constraints voorspelt het model de meest frequente varianten. Bovendien kunnen door middel van herschikking van de constraints ook de minder frequente realisaties voorspeld worden. Verder onderzoek moet aantonen of een stochastische implementatie dezelfde frequenties voorspelt als aangetroffen in het corpus. Duidelijk is in elk geval dat frequentie-informatie onontbeerlijk is in de analyse van de meewerkend-voorwerpconstructie.

Een derde verschijnsel dat onder de loep genomen wordt is de voorzetselgroep zonder determinator (PP-D). Enkelvoudige telbare woorden komen in het algemeen niet voor zonder determinator (bijv. een lidwoord of een telwoord): **Ik koop huis* is ongrammaticaal, net als **huis is mooi*. Opvallend genoeg kan dat vaak wél binnen een voorzetselgroep: *ik ga naar huis*. Geïllustreerd wordt hoe de eigenschappen van verschillende typen PP-D's in grammatica-regels (LFG) gedefinieerd kunnen worden. Maar om de goede PP's (*naar huis*) van de slechte (**naar auto*) te kunnen onderscheiden moet ook bekend zijn welke zelfstandige naamwoorden en welke voorzetsels in zo'n constructie kunnen voorkomen. Bovendien moet bekend zijn wat de grammaticale eigenschappen van die specifieke combinatie zijn, bijvoorbeeld of een bijvoeglijk naamwoord is toegestaan (**naar hoog huis* vs. *op hoge leeftijd*). Met behulp van grote corpora en eenvoudige statistische toetsen is het mogelijk om PP-D's (semi-)automatisch te identificeren en te classificeren naar hun grammaticale eigenschappen.

We onderscheiden 3 basistypen PP-D met elk hun eigen grammaticale eigenschappen. Vaste verbindingen, zoals *in zwang* of *van lieverlee* hebben een betekenis die niet volgt uit de betekenis van de afzonderlijke delen. Ze zijn in corpora te herkennen doordat het zelfstandig naamwoord niet (meer) voorkomt buiten deze constructie. Zo komt het woord *lieverlee* alleen maar voor in combinatie met *van*. Een tweede type is de compositionele PP-D. De betekenis van de PP is wél regelmatig af te leiden en vaak kunnen er ook (bepaalde) bijvoeglijk naamwoorden in voorkomen, bijvoorbeeld *in (wankel) evenwicht*. Kenmerkend is dat de combinatie van voorzetsel en zelfstandig naamwoord vaker zonder determinator voorkomt dan op basis van kans verwacht zou worden. Dit wordt met behulp van de statistische toets *log-likelihood ratio* gemeten. Het derde basistype PP-D wordt gevormd met een voorzetsel uit een kleine groep voorzetsels die verplicht (*per*) of optioneel (bijv. *zonder*) combineren met een zelfstandig-naamwoordgroep zonder determinator. Hoe meer verschillende PP-D's een prepositie vormt, hoe sterker de voorkeur voor deze combinatie. Naast deze driedeling moet ook nog onderscheid gemaakt worden tussen 'zelfstandige' PP-D's en voorzetselgroepen die alleen zonder determinator voorkomen in combinatie met een bepaald werkwoord, bijv. *in toom houden* of *van auto veranderen*. Deze twee meta-

categorieën kunnen onderscheiden worden door een minimum aan variatie in werkwoorden vast te stellen, gemeten door middel van de statistische toets *entropy*.

We kunnen nu voorzetselgroepen classificeren door in een corpus na te gaan of het de karakteristieke eigenschappen vertoont van een bepaald type PP-D. De meeste van de kenmerken zijn echter alleen te herkennen in een syntactisch geannoteerd corpus. Bovendien zijn de statistische toetsen alleen betrouwbaar bij grote hoeveelheden data. Daarom wordt opnieuw automatisch geannoteerde data gebruikt. Met behulp van deze data wordt automatisch een verzameling PP-D's samengesteld. Handmatige evaluatie toont aan dat 20-50% van de geëxtraheerde PP-D's niet syntactisch gemarkeerd is, omdat het zelfstandig naamwoord ontelbaar is en dus geen determinator behoeft. Betere informatie over telbaarheid zou dan ook leiden tot een hogere precisie. Zolang deze niet beschikbaar is, blijft handmatige evaluatie onmisbaar, maar levert extractie op basis van automatisch geannoteerde data een goede kandidatenlijst.

Voor een nauwkeurige extractie van syntactisch gemarkeerde PP-D's is het essentieel dat nauwkeurige informatie over de telbaarheid van zelfstandige naamwoorden beschikbaar is. En niet alleen daarvoor: een sprekende computer moet weten of het **ik wil tosti* is of *ik wil een tosti*. En wanneer hij de zin *ik heb een glas nodig* hoort of leest, dan moet hij weten dat het gaat om een object waaruit gedronken kan worden, niet om een bouwmateriaal, zoals in *ik heb glas nodig*. Helaas is deze informatie niet in ruime mate beschikbaar. Maar met behulp van—alweer—automatisch geannoteerde corpora is het mogelijk om de telbaarheid van woorden automatisch te achterhalen met een hogere precisie dan voorheen voor handen was.

Telbare woorden verschillen van niet-telbare woorden in de context waarin ze voorkomen: telbare woorden komen voor in meervoud, niet-telbare niet (pluralia tantum buiten beschouwing gelaten); enkelvoudige telbare woorden hebben vrijwel altijd een determinator bij zich, niet-telbare niet noodzakelijkerwijs; telbare woorden kunnen voorkomen met het lidwoord *een*, niet-telbare niet. Door de contexten te bekijken van woorden waarvan we zeker weten dat ze telbaar of niet-telbaar zijn, kan een profiel gemaakt worden van de distributie van telbare en niet-telbare woorden. Wanneer nu de distributie van het testwoord in het corpus genoeg lijkt op het profiel van telbare woorden, wordt het als 'telbaar' geclassificeerd. Wanneer de distributie genoeg lijkt op die van de ontelbare woorden, wordt het woord als 'ontelbaar' geclassificeerd. Het kan voorkomen dat een woord zowel telbaar als ontelbaar is, bijvoorbeeld het woord *vis*: het dier is telbaar, het voedsel is ontelbaar. De classificatie vindt plaats op basis van *Memory-Based Learning*, een techniek voor automatisch leren.

De kwaliteit van de classificatie hangt samen met de hoeveelheid trainingsdata (woorden waarvan de telbaarheid bekend is) en voor het Nederlands is er maar weining van die data beschikbaar. Maar omdat het Nederlands en het Engels wat betreft telbaarheid erg op elkaar lijken (de effecten van telbaarheid in corpora zijn ongeveer hetzelfde), gebruiken we niet alleen Nederlandse data, maar ook Engelse: in dat geval wordt het profiel bepaald op basis van Engelse trainingswoorden, en worden daarmee Nederlandse testwoorden geclassificeerd. De resultaten van deze automatische classificatie komen voor 85,7% overeen met die van de handmatige classificatie van een testset, wat een verbetering is ten opzichte van eerder werk op dit gebied.

Er zijn ook andere manieren om de telbaarheid van een zelfstandig naamwoord te bepalen. Zo is betoogd dat de telbaarheid van een woord geen arbitraire lexicale eigenschap is, maar direct samenhangt met de betekenis van een woord. In dat geval zouden het Engelse en het Nederlandse woord voor een bepaald begrip dezelfde telbaarheid moeten hebben, net als het Franse en het Spaanse. De praktijk leert dat dit in veel gevallen ook het geval is. Op basis van deze observatie is een tweede classificatiemethode ontwikkeld. Deze methode maakt gebruik van EuroWordNet. In dit semantische netwerk zijn woorden die een bepaald concept verwoorden (synoniemen, bijvoorbeeld *meel* en *bloem*) gegroepeerd in *synsets*. Deze synsets hebben verbindingen met woorden uit verschillende talen. De synset van *meel* is dus ook verbonden met het Engelse *flour*. Daarnaast is het net hierarchisch geordend, zodat de hypernymen *meel* en *bloem* direct boven de hyponym *tarwebloem* staan. Wanneer nu een trainingswoord in dezelfde synset voorkomt als een testwoord, kunnen we ervan uitgaan dat ze dezelfde telbaarheid hebben. We geven dus de categorie van het trainingswoord door aan het testwoord. De telbaarheid kan niet alleen doorgegeven worden aan synonymen, maar ook aan hypo- of hyper- of cohyponymen ('zusjes' in de hierarchie), en zowel aan Nederlandse als anderstalige (Engelse) trainingswoorden kunnen worden gebruikt. Hoewel de resultaten van dit classificatiesysteem beter zijn dan van een simpel baselinesysteem, halen ze het niet bij de resultaten op basis van corpusdata.

De vier hier samengevatte hoofdstukken illustreren heel verschillende toepassingen van corpusdata: het vinden van voorbeelden en tegenvoorbeelden, het verkrijgen van kwantitatieve data, de extractie van syntactisch gemarkeerde constructies en het automatisch leren van lexicale eigenschappen. In al deze gevallen waren grote hoeveelheden data nodig, die bovendien voorzien moesten zijn van taalkundige annotatie. Hoewel er ruimte blijft voor verbetering, leidde het gebruik van automatisch geparseerde data—ondanks de ruis door mogelijke parseerfouten—in alle gevallen tot goed bruikbare resultaten.