

## Bouwstenen van het Nederlands en het CGN

**Ton van der Wouden\***

*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)*

### **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.

## **1 Inleiding: spreektaal**

Spreeken 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.

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 belletrie 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 spreesituaties: 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 te 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 onbepaalde 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.<sup>1</sup>

## **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, 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 vijftienduizend vaste uitdrukkingen kent (vergelijk ook Weinreich 1969), terwijl Mel’čuk 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 collocatieve 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 collocatieve 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*. Booijs (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. [ZIN *Onderwerp* [GEZEGDE *Werkwoord* Lijdend Voorwerp ]]  
 b. [ZIN *Onderwerp* [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.<sup>2</sup> 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.<sup>3</sup> In het VIDI-project “Dutch as a construction language”, gefinancierd door NWO en UL, 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 ELUs. Eén voorbeeld: een programma zoals WordSmith Tools vindt in een mum van tijd klassieke collocatie-types zoals vaste voorzetsels bij adjectieven.<sup>4</sup> Een standaardvoorbeeld is *trots*, dat gewoonlijk vergezeld gaat van *op*. De tabel hieronder geeft de meest frequente tweewoordclusters met *trots* in het Nederlandse deel van het CGN. De collocatie waarnaar we op zoek waren staat bovenaan:<sup>5</sup>

(3)

| Wordsmith-clusters met <b>trots</b> |                   |            |
|-------------------------------------|-------------------|------------|
| N                                   | Cluster           | Frequentie |
| 1                                   | <i>trots op</i>   | 75         |
| 2                                   | <i>heel trots</i> | 15         |
| 3                                   | <i>trots en</i>   | 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*:<sup>6</sup>

- 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 tèt 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* (fn000562)). (Bij 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.<sup>7</sup>

### 3 Resultaten

In het vervolg presenteren we een paar min of meer opvallende resultaten van het onderzoek. We beperken ons daarbij tot een minieme selectie. Eerst bespreken we een relatief nieuwe vaste combinatie van een bijvoeglijk - en een zelfstandig naamwoord, 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.<sup>8</sup>

(8)

| Wordsmith-collocaten met <i>zinloos</i> |  |            |
|---|--|------------|
| N                                       | Cluster                                  | Frequentie |
| 1                                       | <i>zinloos</i> (totaal)                  | 63         |
| 2                                       | <i>is</i> in de buurt van <i>zinloos</i> | 40         |

|   |  |    |
|---|--|----|
| 3 | <i>geweld</i> in de buurt van <i>zinloos</i> | 38 |
| 4 | <i>dat</i> in de buurt van <i>zinloos</i>    | 20 |
| 5 | <i>en</i> in de buurt van <i>zinloos</i>     | 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 staat 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.<sup>9</sup> 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 spreektaalig. 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 <i>zeg</i> |                 |            |
|------------------------------------|-----------------|------------|
| N                                  | Cluster         | Frequentie |
| 1                                  | <i>zeg maar</i> | 3901       |
| 2                                  | <i>ik zeg</i>   | 2230       |
| 3                                  | <i>zeg ik</i>   | 816        |
| 4                                  | <i>'k zeg</i>   | 807        |
| 5                                  | <i>zeg je</i>   | 650        |
| 6                                  | <i>zeg ja</i>   | 471        |
| 7                                  | <i>dan zeg</i>  | 449        |

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) (zie hierboven), dan moet de volle vokaal 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.<sup>10</sup> 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 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.

*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)*.<sup>11</sup>

- (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 aanbidding 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 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.<sup>12</sup> 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.

### 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) (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*.<sup>13</sup>

### 3.3 Derde resultaat: *weten*

*Weten* is een van de frequentere inhoudswerkwoorden van het Nederlands.<sup>14</sup> *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:

(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.<sup>15</sup> 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 illustreert dit: 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* 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:<sup>16</sup>

- (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: 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 predikatieve 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<br>Nederlandse deel van CGN (N = 5863159)<br>(tellingen WordSmith tools) |            |            |       |
|---|------------|------------|-------|
| N   | Woord      | Frequentie | %     |
| 1   | <i>ja</i>  | 192456     | 3,28  |
| 2   | <i>de</i>  | 159545     | 2,72  |
| 3   | <i>dat</i> | 156005     | 2,66  |
| 4   | <i>uh</i>  | 154252     | 2,63  |
| 5   | <i>en</i>  | 141063     | 2,41  |
| 6   | <i>ik</i>  | 125162     | 2,13  |
| 7   | <i>een</i> | 111628     | 1,90  |
| 8   | <i>'t</i>  | 97895      | 1,67  |
| 9   | <i>je</i>  | 90155      | 1,54  |
| 10  | <i>die</i> | 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):<sup>17</sup>

(24)

| subcorpus                           | # <i>ja</i> | %    | frequentst | %    |
|-------------------------------------|-------------|------|------------|------|
| dialoog (N=1815735)                 | 1           | 4,52 | <i>ja</i>  | 4,52 |
| telefoon (N=1286962)                | 1           | 5,96 | <i>ja</i>  | 5,96 |
| radio (N=1001366)                   | 13          | 1,16 | <i>de</i>  | 4,57 |
| voorgelezen tekst (N=558543)        | 135         | 0,09 | <i>de</i>  | 4,78 |
| onderwijs (N=307876)                | 8           | 2,00 | <i>dat</i> | 2,91 |
| interview (N=264621)                | 4           | 2,86 | <i>uh</i>  | 3,98 |
| (politiek) debat (N=220094)         | 35          | 0,47 | <i>de</i>  | 5,02 |
| televisie (N=196865)                | 17          | 0,92 | <i>de</i>  | 4,02 |
| gesimuleerd zakengesprek (N=140349) | 3           | 3,56 | <i>uh</i>  | 4,33 |
| presentatie (N=63492)               | 97          | 0,13 | <i>uh</i>  | 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.<sup>18</sup> 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 aarzelpauze 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  
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)

| <b>Frequente clusters met <i>ja</i></b> |                 |           |
|---|-----------------|-----------|
| <b>cluster</b>                          | <b>rangorde</b> | <b>#N</b> |
| <i>ja ja</i>                            | 1               | 181347    |
| <i>ja ja ja</i>                         | 2               | 84696     |
| <i>ja ja ja ja</i>                      | 3               | 34430     |
| <i>oh ja</i>                            | 5               | 18512     |
| <i>ja dat</i>                           | 6               | 17721     |
| <i>ja maar</i>                          | 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,

bijvoorbeeld 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 intonationale 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*.<sup>19</sup> 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 gespreksthema 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 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 dat 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.<sup>20</sup> De voorbeelden zijn 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)

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 waar 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 dat wel degelijk zijn vruchten af te werpen. Zonder het Corpus Gesproken Nederlands was dat onmogelijk geweest.

- Altenberg, B. (1998).** On the phraseology of spoken English: The evidence of recurrent word-combinations. In: *Phraseology. Theory, Analysis, and Applications*, red. A.P. Cowie, 101–22. Oxford: Clarendon Press.
- 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.
- Den Boon, T., 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.
- van Esch, C. (2007).** Ik zeg oh ik zeg van ja ach tuurlijk. Over de vorm en de betekenis van *ik zeg* in gesproken taal. Bachelorcriptie, 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* 137–149. Thema’s en trends in de Sociolinguïstiek 5.
- Goemaat, M. (2006).** *Niet eens een focusartikel. Een eerste verkenning van de vaste woordcombinatie niet eens in het Nederlands*. Doctoraalscriptie, 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.
- Den Hertog, C.H. (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 *Particles. Belgian Journal of Linguistics* 16, red. Ton van der Wouden, Ad Foolen, & Piet Van de Craen, 53–64. Amsterdam: John Benjamins.
- 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.
- van der Horst, J. & K. van der Horst (1999).** *Geschiedenis van het Nederlands in de twintigste eeuw*. Den Haag/Antwerpen: Sdu/Standdaard.
- 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.
- van der Linden, E.-J. (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 interactieve 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.
- van Oostendorp, M. & 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 *Language and Communication*, red. J.C. Richards & R.W. Schmidt, 191–226. London: Longman.
- Pollard, C., & I. Sag (1994).** *Head-driven phrase structure grammar*. Center for the Study of Language and Information Stanford.
- 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 *Computational Linguistics in the Netherlands 2004: Selected papers from the fifteenth CLIN meeting*, red. T. van der Wouden, M. Poss, H. Reckmann, & C. Cremers, 187-202. Utrecht: LOT.
- 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, pp. 1-15.
- de Saussure, F. (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 *Nauwe betrekkingen*. Voor Theo Janssen bij zijn vijftigste verjaardag, red. R. Boogaart & J. Noordegraaf, 273–282. Amsterdam & Münster: Stichting Neerlandistiek VU & Nodus Publikationen.

- Vismans, R. (1994).** *Modal particles in Dutch directives: a study in functional grammar*. Diss. Vrije Universiteit Amsterdam.
- de Vries, J. (2001).** *Onze Nederlandse Spreektaal*. Den Haag: SDU Uitgevers.
- de Vries, W. (1910).** *Dysmelie. Opmerkingen over syntaxis*. Verhandeling behorende bij het programma van het gymnasium der gemeente Groningen voor het jaar 1910–1911.
- de Vries, W. (1911).** *Dysmelie. Opmerkingen over syntaxis (vervolg)*. Verhandeling behorende bij het programma van het gymnasium der gemeente Groningen voor het jaar 1911–1912.
- de Vries, W. (1914).** *De typen der mededeeling. Opmerkingen over syntaxis*. Verhandeling behorende bij het programma van het gymnasium der gemeente Groningen voor het jaar 1914–1915.
- Weinreich, U. (1969).** Problems in the analysis of idioms. In *Substance and structure of language*, red. Joan Puhvel, 23–81. Berkeley and Los Angeles: University of California Press. (reprinted in *On Semantics*, 1980).
- Woordenlijst (2005).** *Woordenlijst Nederlandse Taal*. Tiel, Den Haag: Lannoo, Sdu. samengest. door het Instituut voor Nederlandse Lexicologie; met een leidraad door Ludo Permentier.
- Wotjak, B. (1992).** *Verbale Phraseolexeme in System und Text*. Tübingen: Niemeyer.
- van der Wouden, T. (1992).** Beperkingen op het optreden van lexicale elementen. *De Nieuwe Taalgids* 85, 513–38.
- van der Wouden, T. (1995).** Moeilijk is (not) difficult. In *Linguistics in the Netherlands 1995*, ed. by Marcel den Dikken & Kees Hengeveld, 203–14. Amsterdam: Benjamins.
- van der Wouden, T. (1998).** Dat had niet zo gehoeven: Modaliteit en negatie in de nieuwe ANS. *Nederlandse Taalkunde* 3, 237–252.
- van der Wouden, T. (2002a).** Particle research meets corpus linguistics: on the collocational behavior of particles. In *Particles. Belgian Journal of Linguistics* 16, red. T. van der Wouden, A. Foolen, & Piet Van de Craen, 151–174. Amsterdam: John Benjamins.
- van der Wouden, T. (2002b).** Partikels: naar een partikelwoordenboek voor het Nederlands. *Nederlandse Taalkunde* 7, 20–43.
- van der Wouden, T. (2007a).** On the phraseology of stop words. In P. Hendriks, F. Landsbergen, M. & J. van der Wal (red.), *Leiden Papers in Linguistics* 4.1 (2007), 56–67 (<http://www.lucl.nl>).
- van der Wouden, T. (2007b).** Bouwstenen van het Nederlands. Een pleidooi voor (nog) meer aandacht voor de idiomaticiteit van het Nederlands. Te verschijnen in *Neerlandistiek in contrast. Bijdragen aan het Zestiende Colloquium Neerlandicum*, Amsterdam, Rozenberg, 287–298.
- van der Wouden, T. (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.

## NOTEN

\* 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 *NT* 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).

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. 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>
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 collocatieve 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 *moelijk* (van der Wouden 1995). In de literatuur vindt men overigens verschillende opvattingen over de precieze definitie van dynamische modaliteit (vergelijk Nuyts et al. 2007).
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.
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.
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.
11. Volgens [www.jeugdsentimenten.net/2005/11/26/jeugdsentimentendag-wat-gaan-we-doen/](http://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 *NT* 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 bereidt aan de Leidse opleiding Nederlandse taal en cultuur een Bachelorscriptie voor 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.
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 (*praeceptique 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.
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  $5663528 \cdot (38892501812165017084584 / 56635285)$ , 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.
16. [www.onzetaal.nl](http://www.onzetaal.nl)
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.
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).
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*).