

Collocaties en het probleem van de corpusgrootte

Ton van der Wouden

Bijeenkomst STDH: *Het internet als bron*, 16 november 2001, Meertens-instituut

Samenvatting

De collocatieliteratuur gaat doorgaans over inhoudswoorden: het adjectief "dol" gaat vrijwel altijd vergezeld van een voorzetselcomplement met "op" als hoofd, en het geijkte graadadverbium bij het werkwoord "slapen" is "diep". Toch vinden we ook collocatoneel gedrag bij functiewoorden: het focusbijwoord "alleen" wordt vaak gevolgd door "maar", terwijl we in het geval van het partikel "eens" al met een heel klein corpus, of twintig seconden reflectie, op het spoor komen van vaste combinaties als "niet eens" en "wel eens". Voor andere vaste combinaties van partikels blijkt het grootste corpus - het internet dus - daarentegen nog nauwelijks toereikend. Een verslag van lopend onderzoek.

1 Wat is een collocatie?

Ik ga het hier over collocaties van partikels hebben, maar wat een collocatie is, dat weet bijna niemand, ten minste, niet precies. De term wordt meestal teruggevoerd op de Britse taalkundige Firth (Firth 1957),¹ die aandacht vroeg voor vaste combinaties die niet terug te voeren zijn op bekende grammaticale noties als subcategorisatie. Klassieke voorbeelden uit de Britse collocatieschool zijn combinaties als

- (1) vaste adjectieven: *strong tea* / *?powerful tea* (vergelijk: *powerful computer* / *?strong computer*)
- (2) vaste voorzetsels: *decide on*, *operate upon*; *dol op*, *afhankelijk van*
- (3) vaste relaties tussen inhoudswoorden, bijvoorbeeld werkwoorden en nomina: wat doe je met *postzegels*? die *verzamel* je (onder andere)²

Firth geeft zelf geen heldere definitie van collocatie, maar maakt gebruik van slogans als

- (4) You shall know a word by the company it keeps (Firth)

In de literatuur na Firth vind je grofweg twee soorten definities van collocatie: beperkte en algemene. Een voorbeeld van een beperkte definitie is die van Hausmann uit Hausmann (1989-1991), aangehaald in (5):

¹Maar volgens van der Wouden (1997:7, n.5) dateert de oudste bewijsplaats in de OED van *collocation* in kennelijk taalkundige zin uit 1750. Geerts & den Boon (1999) geeft als oudste vindplaats voor *collocatie* 1658, maar daar is niet duidelijk om welke van de vier betekenissen het gaat.

²Dit voorbeeld laat gelijk zijn dat collocatonele afhankelijkheden niet beperkt zijn tot strikte adjacentie, en zelfs zinsgrenzen kunnen overschrijden.

(5) On appellera collocation la combinaison caractéristique de deux mots dans une des structures suivantes :

- a substantif + adjectif (épithète);
- b substantif + verbe;
- c verbe + substantif (objet);
- d verbe + adverbe;
- e adjectif + adverbe;
- f substantif + (prép.) + substantif.

La collocation se distingue de la combinaison libre (*the book is useful / das Buch ist nützlich / le livre est utile*) par la combinabilité restreinte (ou affinité) des mots combinés (*feuilleter un livre* vs. *acheter un livre*). (Hausmann 1989-1991)

De beperking tot bepaalde soorten inhoudswoorden wordt in het artikel in het geheel niet gemotiveerd.

Twee voorbeelden van heel ruime, algemene definities van collocatie staan in (6):

- (6) a. COLLOCATIE: een idiosyncratische restrictie op de verbindbaarheid van woorden (Geeraerts 1986:134)
- b. The term COLLOCATION refers to the idiosyncratic syntagmatic combination of lexical items and is independent of word class or syntactic structure (Fontenelle 1992)

Elders (van der Wouden 1992; van der Wouden 1994; van der Wouden 1997) heb ik aannemelijk proberen te maken dat zelfs deze definitie nog te krap is, maar dat is nu niet zo relevant. In de loop van deze lezing hoop ik u er wel van te kunnen overtuigen dat ongemotiveerde restricties zoals in de eerste definitie hoe dan ook ongewenst zijn.

2 Collocationeel gedrag bij partikels

Zoals de meesten van u wel zullen weten ben ik al geruime tijd bezig met een woordenboek van de partikels van het Nederlands.³ Een van de vele intrigerende eigenschappen van de partikels is hun sociale gedrag – Hoogvliet (1903) wees er al in 1903 op, dat clusters van wel 6 partikels mogelijk zijn:⁴

- (7) Geef de boeken *dan nu toch maar eens even* hier (Hoogvliet)
- (8) Zo is het *toevallig dan toch ook nog wel eens even* (Foolen)
- (9) Ik zie mezelf *best nog wel* over tien jaar *ook nog* voor de klas staan (CGN)

Vooruitlopend op wat er later in deze lezing nog komt: er is geen corpus dat groot genoeg is om positief bewijs te vinden voor mijn intuïtie – een intuïtie die u vermoedelijk deelt – dat (7) en (8) welgevormde zinnen van het Nederlands zijn. Ook het Internet is daarvoor niet groot genoeg.⁵ Maar dit voorlopig ter zijde – we komen daarop terug.

³VNC-project 'Partikelgebruik in Nederland en Vlaanderen'.

⁴Noordegraaf (2000) toont aan dat Hoogvliet in dezen schatplichtig is aan Henry Sweet en Georg von der Gabelentz.

⁵De clusters in (7) en (8) zijn via standaard internetzoekmachines niet te vinden in spontane teksten. Het langste cluster in de eerste vier releases van CGN dat met *dan toch ook nog* begint vinden we in de Vlaamse zin *waar ik dan toch ook nog wel iets uit kan leren*.

De volgorde in clusters van modale partikels is meestal zo goed als vast (Thurmair 1989; Vismans 1994); de lezer kan voor zichzelf nagaan dat volgordevariatie in de gegeven voorbeelden hetzij moeilijk gaat of betekenisverandering oplevert.

De volgordes binnen clusters van modale partikels wordt volgens de Vriendt *et al.* (1991) in grote mate bepaald door de oorspronkelijke betekenissen van de partikels.

Clusters kunnen verstenen en een eigen, ondoorzichtige betekenis ontwikkelen: de combinatie *niet eens* (als in *ze weten niet eens wat voetballen is*) betekent niet meer ‘niet een keer’ maar *zelfs niet* (Rullmann & Hoeksema 1997).

Clusters kunnen verstenen en eigen, onverwachte gebruiksmogelijkheden en gebruiksrestricties ontwikkelen. Vismans (1994) constateert bijvoorbeeld dat de combinatie *maar eens* vaak, maar niet uitsluitend voorkomt in directieve zinnen.

(10) Ga *maar eens* kijken wat er in je schoen zit

(11) Je moet het *maar eens* met een vork proberen

(12) Ik denk dat ik *maar eens* naar huis ga

Bij sommige moderne taalgebruikers lijkt deze tendens zich zover te hebben doorgezet dat *maar eens* uitsluitend nog aangetroffen wordt in zwakke directieven – en dan niets meer lijkt te zijn dan een markeerder van dat soort zinnen.

Een voorbeeld van zo’n moderne *maar eens*-gebruiker is Maarten ‘t Hart in zijn roman *De Nakomer* uit 1996.⁶ Daarin komt de combinatie 14 keer voor, waarvan maar liefst 13 keer in een directieve zin, meestal een morfologische imperatief (9 keer) of een vorm van het werkwoord *moeten* (3 keer).

Interessant genoeg maakt *maar eens* in de meeste gevallen in deze roman deel uit van een groter partikelcluster. Een paar voorbeelden:

(13) Sta *eerst maar eens* op

(14) denkt u er *nog maar eens* goed over na, als ‘t zover is, zult u vanzelf wel van gedachten veranderen

(15) Ik moet *eerst maar eens* kijken of ze bij me thuis zijn geweest

Dit bij wijze van **anecdotische** inleiding op clustergedrag of – een term die ik prefereer – collocatoneel gedrag bij partikels. In het vervolg van deze voordracht wil ik graag ingaan op de mogelijkheden en onmogelijkheden zijn van **systematisch** corpusonderzoek naar zulk soort gedrag (vergelijk ook Lemnitzer (2001)).

3 Het automatisch vinden van collocaties

Iets anders is dat we zo’n definitie willen gebruiken, en daarvoor zijn de hier gegeven definities aan de erg abstracte kant. Een tegenwoordig erg populaire manier om definities van collocaties handen en voeten te geven is ze kwantitatief te interpreteren (Manning & Schütze 1999:Ch. 5). Een praktische definitie van collocatie kan er dan ongeveer als volgt uitzien:⁷

⁶Maarten ‘t Hart, *De nakomer*. Amsterdam; Antwerpen: De Arbeiderspers, 1996.

⁷van der Wouden (1997) beargumenteert dat ook woordcombinaties die *minder vaak* voorkomen dan de kansberekening voorspelt – “negatieve collocaties” bij gebrek aan een betere term – interessant kunnen zijn, maar dat ter zijde.

(16) COLLOCATIES zijn (2- of meer-)woordcombinaties die vaker voorkomen dan de kansberekening voorspelt.

(We gaan nu maar niet in op vragen als: wat verstaan we precies onder een woordcombinatie (adjacent of niet? hoe ver van elkaar?), welke woorden mogen meedoen (Hausmann!), welke statistiek moeten we gebruiken?)

Er zijn tegenwoordig handige programma's op de markt die in elk geval het ingewikkelde rekenwerk voor je doen. Ik noem er, zonder pretentie van volledigheid, drie.

- WordSmith tools van Mike Scott (www.oup.co.uk)
- MonoConc van Steve Barlow (www.ruf.rice.edu/~barlow/mono.html)
- "Bigram Statistics Package" (www.d.umn.edu/~tpederse/code.html)

In dit verhaal gebruik ik de WordSmith tools.⁸

Om te laten zien dat dit soort programma's inderdaad (ongeveer) doet wat je zou willen doen geef ik zo meteen de uitvoer van dat programma voor de vorm *luisterde* in het Nederlandse gedeelte van het Corpus Gesproken Nederlands (de eerste vier releases), totaal een kleine twee en een half miljoen woorden (2483589 tokens in 67241 types). Maar eerst moet ik heel in het kort iets zeggen over Het Corpus Gesproken Nederlands.

- 10 miljoen woorden gesproken Nederlands
- een derde België, twee derde Nederland
- sociografisch gestratificeerd, teksttypologisch gespreid, enz.
- uitgebreid verrijkt:
 - orthografische transcriptie
 - lemmatisering
 - POS-tagging (taalkundige ontleding)
 - syntactische annotatie (redkundige ontleding)
 - prosodische annotatie
 - enz.

Dan geef ik nu een stukje van de KWIC-lijst voor *luisterde* uit het Nederlandse gedeelte van CGN tot en met de vierde release (oktober 2001) (waarbij KWIC staat voor Keyword in Context (zie bijvoorbeeld Sinclair (1991))):

⁸Elders (van der Wouden 2001a) doe ik verslag van vergelijkbare exercities met een ander hulpmiddel, het "Bigram Statistics Package" (BSP) van Ted Pedersen en Satanjeev Banerjee aan de University of Minnesota, Duluth. Het gaat om een bibliotheek Perl-routines waarmee de onderzoeker op handige wijze bigrammen kan extraheren en ze kan onderwerpen aan statistisch onderzoek. (The BSP is free software under the terms of the GNU General Public License; the code can be found at www.d.umn.edu/~tpederse/code.html.)

Word-Smith-KWIC voor <i>luisterde</i> , CGN-Nederland okt. 2001 (2483589 tokens in 67241 types; $N = 40$)			
1	herinneringen ophaalde	luisterde	mevrouw Van D
2	ed bleef hij staan en hij	luisterde	naar de regelm
3	m midden op het plein	luisterde	Surya vaak na
4	en dan haar pas in en	luisterde	naar het feestg
5	e dat hij daar niet naar	luisterde	en brieven over
6	at*x je naar de mensen	luisterde	. aan de basis
7	grote zwarte dikke kat	luisterde	spinnend en hi
8	k de dag d'rna xxx toen	luisterde	ik effe m'n voic
9	de cipier had geopend	luisterde	hij naar het gel
10	it vaker gezegd maar je	luisterde	nooit. je hoeft
11	ompstationsweg op en	luisterde	verrukt naar de
12	us omdat ik heel goed	luisterde	naar de piano.
13	erkdag van veertien uur	luisterde	er in haar auto
14	dansen. en iedere dag	luisterde	de koningin he
15	Gerrit Huizer. ja. ja hij	luisterde	altijd gewoon n
16	cht naar de drukte. dan	luisterde	hij naar de ges
17	et gips op zijn borst en	luisterde	naar zijn snelle
18	monoloog en niemand	luisterde	eigenlijk alleen

Dat ziet er aardig uit: we zien dat het voorzetsel *naar* inderdaad, volgens verwachting, regelmatig in de buurt van de werkwoordsvorm staat. Maar wat is regelmatig? Dat kunnen we laten berekenen (en we hoeven ons niet bezig te houden met de statistiek: die blijft onder de motorkap).⁹

<i>luisterde</i> in het Nederlandse deel van CGN (2483589 tokens in 67241 types)															
N	WORD	TOTAL	LEFT	RIGHT	L5	L4	L3	L2	L1	*	R1	R2	R3	R4	R5
1	LUISTERDE	42	1	1	0	1	0	0	0	40	0	0	0	1	0
2	naar	28	6	22	1	1	1	2	1	0	9	7	2	1	3
3	hij	11	7	4	0	2	2	0	3	0	2	1	1	0	0
4	van	11	3	8	1	0	2	0	0	0	0	1	0	0	7
5	het	10	5	5	3	1	0	1	0	0	0	1	1	3	0
6	niet	9	3	6	0	0	2	1	0	0	4	1	0	0	1
7	een	8	3	5	1	2	0	0	0	0	0	1	2	1	1
8	maar	7	4	3	0	0	0	3	1	0	0	0	0	3	0

Ik wijs erop dat het vaste voorzetsel *naar* inderdaad bovenaan staat, en kennelijk bij voorkeur onmiddellijk rechts van het sleutelwoord ('keyword') *luisterde* staat. Ik wijs er ook op dat *niet* tamelijk vaak in de omgeving van *luisterde* wordt aangetroffen – dat is een type collocatie waar ik zelf niet direct aan gedacht had.

Voor de volledigheid herhalen we de exercitie met *luisterde* voor het Belgische deel van CGN (tot en met de vierde release) : daar hebben we het over 1.775.222 woorden (types) verdeeld over 63678 types:

⁹Onverlet de gegevenheid dat collocationale relaties in theorie grote afstanden kunnen overbruggen (vergelijk noot 2), wijst de ervaring uit dat een 'venster' van 5 woorden links en rechts van het sleutelwoord voldoende is om de belangrijkste collocaties op het spoor te komen.

<i>luisterde</i> in het Belgische deel van CGN (1775222 tokens in 63678 types)															
N	WORD	TOTAL	LEFT	RIGHT	L5	L4	L3	L2	L1	*	R1	R2	R3	R4	R5
1	LUISTERDE	27	1	0	1	0	0	0	0	26	0	0	0	0	0
2	naar	15	5	10	0	0	2	1	2	0	4	3	2	1	0
3	hij	12	9	3	0	3	0	0	6	0	0	1	1	0	1
4	het	5	2	3	0	0	2	0	0	0	0	1	0	2	0
5	niet	5	1	4	0	0	1	0	0	0	1	1	0	0	2

Onze bevindingen van daarnet worden bevestigd: het (voor ons gevoel) vaste voorzetsel *naar* scoort het hoogst, maar ook *niet* vinden we weer regelmatig terug.

4 Collocationeel gedrag bij partikels

Dan zijn we er nu eindelijk aan toe om te kijken of je op deze manier ook collocationeel gedrag van partikels op het spoor kunt komen. Uit de pre-computationele literatuur is het een en ander bekend over clustergedrag bij modale partikels. Daar gaan we het zo meteen over hebben.

4.1 Alleen

Over clustergedrag bij focuspartikels vind je zelden iets in de literatuur, maar ook bij sommige focuspartikels sterke collocationele tendensen (van der Wouden 2000). Kijk bijvoorbeeld maar eens naar *alleen* ($N = 2844$ in het Nederlandse gedeelte van CGN tot en met nu):

<i>alleen</i> in het Nederlandse deel van CGN															
N	WORD	TOTAL	LEFT	RIGHT	L5	L4	L3	L2	L1	*	R1	R2	R3	R4	R5
1	ALLEEN	2950	51	55	13	9	16	10	3	2844	3	3	20	13	16
2	maar	1226	206	1020	58	57	34	25	32	0	738	79	66	69	68
3	dat	687	396	291	90	62	99	105	40	0	48	49	59	66	69
4	niet	680	539	141	47	48	45	32	367	0	33	16	22	27	43
5	die	430	209	221	44	37	48	60	20	0	36	44	51	35	55
6	een	395	135	260	47	33	31	24	0	0	73	63	44	41	39
7	van	344	138	206	36	41	32	21	8	0	28	41	55	43	39
8	het	334	177	157	34	33	36	41	33	0	31	41	30	23	32
9	dan	321	202	119	23	39	69	15	56	0	20	17	25	28	29
10	ook	285	180	105	40	34	37	34	35	0	1	4	22	35	43
11	voor	212	66	146	14	11	18	18	5	0	62	26	22	19	17
12	nog	199	53	146	17	15	9	8	4	0	81	7	17	29	12

Ik wijs op *maar*, dat heel erg hoog staat; ook *niet* is interessant, al is het alleen maar omdat het veel vaker links dan rechts van *alleen* wordt aangetroffen. Een gedeelte van de verklaring voor dit kaatste feit zal wel liggen in het bestaan van de reeksvormer (om een term van Paardekooper ([n.d.]) te gebruiken) *niet alleen ... maar ook*.

Een andere manier om naar de resultaten te kijken is het programma clusters te laten zoeken. In de volgende tabel staat het topje van de ijsberg van hoogfrequente bigrammen met *alleen*.

Hoogfrequente bigrammen met <i>alleen</i> in CGN-NL ($N \approx 2950$)	
cluster	Freq.
alleen maar	746
niet alleen	361
alleen de	140
je alleen	93
alleen nog	85
alleen een	77
alleen in	70
is alleen	69
alleen uh	66
alleen voor	64

De bovenste twee combinaties, *alleen maar* en *niet alleen*, zijn vermoedelijk het interessantst. Ik wijs er in het voorbijgaan op dat het WNT de combinatie *alleen maar* wel opmerkt maar impliciet afkeurt:

(17) “Ook dient ter versterking, in de volksspraak, het gelijkbeteekenende *maar*, vóór of achter *alleen* geplaatst” (WNT s.v. *alleen* III).

(18) *niet alleen ... maar ook*

Ik geef ook nog de meest frequente *trigrammen*:

Hoogfrequente trigrammen met <i>alleen</i> in CGN-NL ($N \approx 2950$)	
cluster	Freq.
alleen nog maar	47
niet alleen de	37
je alleen maar	36
alleen maar een	32
t alleen maar	32
is niet alleen	28
ik heb alleen	25
niet alleen maar	25
is alleen maar	24
dan alleen maar	22

Interessant (in mijn ogen) is dat het meest frequente bigram, *alleen maar*, niet voorkomt in het meest frequente trigram, *alleen nog maar* — al hebben ze natuurlijk vermoedelijk wel met elkaar te maken.

4.2 Eens

Laten we vervolgens eens kijken naar *eens*, het meest sociale modale partikel dat ik ken.¹⁰ In het Nederlandse gedeelte van het CGN tot nu toe komt dat partikel maar liefst 4050 keer voor.

¹⁰Vergelijk onder meer Callebaut *et al.* (1998) en Zwarts & van der Wouden (2000).

<i>eens</i> in het Nederlandse deel van CGN															
N	WORD	TOTAL	LEFT	RIGHT	L5	L4	L3	L2	L1	*	R1	R2	R3	R4	R5
1	EENS	4257	92	115	25	20	21	22	4	4050	19	24	19	23	30
2	wel	1139	1052	87	34	23	13	18	964	0	1	6	13	32	35
3	dat	1116	517	599	118	148	127	93	31	0	111	125	106	138	119
4	nog	785	708	77	17	22	16	246	407	0	3	9	21	20	24
5	niet	719	586	133	29	31	43	50	433	0	9	15	27	36	46
6	maar	690	421	269	61	56	30	34	240	0	53	46	60	48	62
7	dan	611	348	263	78	93	80	59	38	0	36	61	56	67	43
8	ook	560	447	113	19	28	56	284	60	0	2	15	30	36	30
9	die	544	243	301	48	66	76	44	9	0	48	53	62	77	61
10	een	512	93	419	49	23	15	5	1	0	78	88	68	80	105
11	nou	504	330	174	63	61	48	53	105	0	30	35	34	33	42
12	het	413	224	189	34	50	62	57	21	0	23	30	46	42	48
13	van	382	122	260	46	33	28	14	1	0	63	39	46	64	48
14	ggg	342	108	234	34	30	25	19	0	0	55	48	51	44	36
15	heb	336	242	94	31	66	95	41	9	0	7	21	22	23	21
16	wat	323	57	266	25	14	11	6	1	0	73	58	45	53	37
17	daar	304	197	107	33	41	31	55	37	0	14	23	18	24	28
18	keer	302	20	282	5	4	3	2	6	0	261	8	5	4	4

Ook hier kunnen we wel weer eens naar de clusters rond *eens* kijken, omdat dat dikwijls een duidelijker eerste indruk geeft. In de volgende tabel staat het topje van de ijsberg van hoogfrequente bigrammen met *eens*.

Hoogfrequente bigrammen met <i>eens</i> in CGN-NL ($N \approx 4050$)	
cluster	Freq.
wel eens	970 (1037)
eens een	531
niet eens	437
nog eens	411
eens even	247
maar eens	238
eens wat	124
weer eens	113
eens uh	111
eens in	106
nou eens	106
eens kijken	104

Bovenaan staat de combinatie *wel eens* – een combinatie die zo frequent is dat ze volgens bepaalde gezaghebbende bronnen (de Vries & te Winkel et al. 1864–1998; Renkema 1989; Geerts & den Boon 1999) in sommige gevallen aan elkaar geschreven moet worden – maar volgens een andere minstens zo gezaghebbende bron (Woordenlijst 1995) weer niet. In de orthografische transcriptie van het Corpus Gesproken Nederlands treffen we die spelling *wel eens* ook wel eens aan (om precies te zijn, 67 keer in het Nederlandse materiaal en 3 keer in het Vlaamse), maar zo te zien niet consequent – ik kan in elk geval de systematiek niet ontdekken. Hoe het ook zij, de score van de combinatie van *wel* plus *eens* is dus nog hoger: 1037.

En als we toch aan het mopperen zijn over die orthografische transcriptie: het partikel in *eens* komt in twee vormen voor, *eens* en *'ns*, en ook hier kan ik de ratio niet ontdekken, behalve dan dat de variant die met een volle klinker wordt uitgesproken zelden of nooit als *'ns* wordt geschreven. Hieronder een bloemlezing van de varianten.

- (19) en dan zou 't *wel 'ns* uit kunnen gaan.
- (20) want ik ik zit nog *wel 'ns* zo 'ns te piekeren
- (21) ja dat heb ik *weleens* gedaan.
- (22) i*a ik ik zeg *wel 't* volgend jaar zou 't *weleens* heel anders kunnen zijn.
- (23) want je hebt er *wel eens* over gesproken.
- (24) want we zeggen hier *wel eens* ja goed als je puzzelt als je bridget

Voor deze presentatie heb ik alle gevallen van *'ns* in het corpus vervangen door *eens* – onder het motto dat orthografie orthografie is.

De volgende in het rijtje is *eens een* – en waarom die combinatie zo hoog staat, zal zo meteen hopelijk wel duidelijk worden.

De volgende combinatie is met *niet*. Als er nou een combinatie is die aan elkaar geschreven zou moeten worden dan is het wel dit *niet eens*: in het gebruik met een volle klinker heeft de combinatie *niet eens* een volstrekt ondoorzichtige betekenis – het is namelijk een focuspartikel, met een betekenis vergelijkbaar met *zelfs niet*. Toch ken ik maar een schrijver die *niet eens* consequent als een woord schrijft, namelijk Albert Helman. Een paar voorbeelden:

- (25) De ander hoorde het nieteens. (Albert Helman, *Het vergeten gezicht*. Rotterdam: Nijgh & Van Ditmar, 1939.)
- (26) Op de divan lag Macario te slapen; hij had nieteens de poncho over zich getrokken, zich niet uitgekleeed. (ibid.)
- (27) Zo waren mannen. Elizondo. Allen even verachtelijk. Nieteens aan het verachtelijke toe, nieteens om echt en hevig te kunnen haten. (ibid.)

Wat daarna nog komt – *nog eens*, *eens even*, *maar eens*, *eens wat*, *weer eens*, enz. – behoeft denk ik geen toelichting, behalve misschien *eens uh* dat laat zien dat er ook regelmatig gearzeld wordt na *eens*.

Ter zijde: voor het materiaal uit België ziet die lijst er anders uit. Om dat zo duidelijk mogelijk te maken zet ik de cijfers maar even naast elkaar.

Bigrammen met <i>eens</i> , NL-B			
België		Nederland	
cluster	Freq.	cluster	Freq.
nog eens	575	wel eens	970 (1037)
wel eens	463 (465)	eens een	531
eens een	302	niet eens	437
eens even	157	nog eens	411
eens naar	131	eens even	247
niet eens	129	maar eens	238
maar eens	125	eens wat	124
al eens	104	weer eens	113
kijk eens	103	eens uh	111
ook eens	89	eens in	106
eens in	87	nou eens	106
dat eens	84	eens kijken	104

Ik merk op dat de verschillen tussen de twee tabellen aanzienlijk zijn. In van de Poel & van de Walle (1995:330) wordt gesteld dat er “geen noemenswaardige verschillen” in het partikelgebruik van Nederland en België zouden zijn. Het zojuist gedemonstreerde verschil in het collocatieel gedrag van *eens* is maar één van de vele forse verschillen die we in ons partikelproject hebben aangetroffen (van der Wouden 1999; van der Wouden 2001b).

Maar ook dat terzijde. We gaan weer verder met *eens* in Nederland, op zoek naar langere clusters. In de volgende tabel staan de trigrammen met *eens*.

Hoogfrequente trigrammen met <i>eens</i> in CGN-NL ($N \approx 4050$)	
cluster	Freq.
eens een keer	256
ook wel eens	175
nog wel eens	138
wel eens een	128
nog eens een	100
nog niet eens	78
ook nog eens	71
eens even kijken	62
ik wel eens	51
moet je eens	50
niet eens meer	46
je wel eens	42
eens in de	39

Hier zien we waar die hoogfrequente combinatie *eens een* vandaan komt: *eens een keer* staat op de eerste plaats.

Dan de tetragrammen (natuurlijk worden de clusters minder frequent naar mate ze langer worden):

Hoogfrequente tetragrammen met <i>eens</i> in CGN-NL ($N \approx 4050$)	
cluster	Freq.
nog eens een keer	68
wel eens een keer	49
eens een keer een	36
ook wel eens een	31
nog wel eens een	27
ook nog eens een	25
ook nog wel eens	25
heb ik wel eens	23
ik ook wel eens	22
moet je eens luisteren	16
ook eens een keer	16
af en toe eens	15

En er zijn zelfs pentagrammen en hexagrammen die boven komen drijven:

Hoogfrequente pentagrammen met <i>eens</i> in CGN-NL ($N \approx 4050$)	
cluster	Freq.
ook nog eens een keer	20
nog wel eens een keer	10
ook wel eens een keer	10
heb ik ook wel eens	9
nog eens een keer een	9
af en toe wel eens	8
Hoogfrequente hexagrammen met <i>eens</i> in CGN-NL ($N \approx 4050$)	
cluster	Freq.
af en toe wel eens een	5
dat ben ik met u eens	5

5 Tussentijdse conclusie

Hopelijk heb ik u hiermee alvast overtuigd van twee zaken.

1. Kwantitatieve benaderingen kunnen wel eens helpen collocatieel gedrag te vinden
2. Partikels kunnen interessant collocatieel gedrag vertonen

6 Corpusgrootte: *maar eens*

Eerder liet ik al zien dat ook vaste combinaties zelf hun eigen collocatieve eigenschappen kunnen bezitten: ik besprak de combinatie *maar eens* die voor mijn gevoel (en niet alleen het mijne) vooral in directieve contexten (Vismans 1994) voorkomt. Om dat kwantitatief te onderbouwen

moeten we een trucje uithalen, omdat Wordsmith niet ingericht is op het zoeken naar collocaties bij woordcombinaties. Dat trucje is eenvoudig genoeg: vervang, bijvoorbeeld met een editor (neem geen tekstverwerker zoals Word!) *maar eens* door *maareens*.

Dat heb ik 241 keer gedaan, en dan kun je het programma weer laten rekenen. De volgende tabel geeft een heel klein stukje van de KWIC-lijst (waarbij ik van *maareens* maar weer *maar eens* heb gemaakt).

1	ach*x en verzin d'r nog	maar eens	uh vijf vage ar
2	tergrond nou ja. vertel 't	maar eens	. ja. geen pla
3	nee alleen in de u moet	maar eens	kijken in de w
4	. en als ik dat dan ook	maar eens	ten dele krijg
5	ncona nog. ja doe dat	maar eens	dan. 'k weet h
6	ach te spelen moet je	maar eens	proberen. da
7	nken oh ik ga vandaag	maar eens	naar de huisw
8	edoel ik moet hem nog	maar eens	een engeltje v
9	n. mmm. maar tel nou	maar eens	even. ja. tel..
10	ent nu nog in opleiding	maar eens	kan dit allema
11	en haar veranderd. om	maar eens	letterlijk te zij
12	meent de eerste. eerst	maar eens	zien dat we in
13	h papieren*x om*x 't*x	maar eens	zo te zeggen.
14	ncurrenten*n om dat zo	maar eens	te noemen to
15	r d'rover heb d*a laat ik	maar eens	proberen om t
16	nu helderheid uh ma*a	maar eens	over moeten g
17	t daar een paal. moet je	maar eens	... en daar sta
18	arom. ja? maar ik zou	maar eens	even*x kijk*a
19	at dacht je van nou doe	maar eens	een voorstel.
20	g. nou wor*a wordt dan	maar eens	niet boos. ja

Dit is natuurlijk te weinig als wettig en overtuigend kwantitatief bewijs, maar we zien hier toch veel imperatief-vormen (*verzin*, *vertel*, *doe*), flink wat voorkomens vormen van *moeten*, af en toe een infinitief die als imperatief fungeert (*eerst maar eens zien*, enzovoort. Dus we hebben in elk geval een indicatie dat onze indruk (dat *maar eens* vooral in directieve contexten voorkomt) ergens op slaat. Maar er zijn ook tegenvoorbeelden te vinden.

Bij Maarten 't Hart zagen we dat *maar eens* nogal eens vergezeld ging van het partikel *eerst*. De vraag is nu, of die combinatie *eerst maar eens* eenzelfde soort restricties tot directieve omgevingen kent – of juist niet.

Het CGN bevat maar 14 voorkomens van *eerst maar eens*

- (28) nou laten we *eerst maar eens* uitzoeken
- (29) maar ja kom d'r *eerst maar eens* in de buurt hè.
- (30) leer jij *eerst maar eens* die telefoon op te nemen.
- (31) ga *eerst maar eens* naar een
- (32) van wacht er maar effe mee ga *eerst maar eens*
- (33) laat 'r *eerst maar eens* die kosten uh...

- (34) ga *eerst maar eens* squashen
- (35) zou ik *eerst maar eens* met dit boekje beginnen.
- (36) luister je *eerst maar eens*.
- (37) *eerst maar eens* hieruit ontsnappen.
- (38) *eerst maar eens* zien dat we in de buurt

De meeste gevallen zien er inderdaad nogal directief uit, maar (35) zou zowel een suggestie tot bepaald gedrag (dus een directief) alsook een ander soort taalhandeling kunnen zijn.

Dus gaan we verder kijken. Van de drie voorbeelden die Het Corpus Eindhoven biedt zijn er twee onmiskenbaar directief, en de derde juist niet.

- (39) en dat de andere man tussen de stammen 'n geluidstechnicus met zijn recorder voorstelde, moest-ie ook eerst maar eens bewijzen. (Corpus Eindhoven: gezinsbladen)
- (40) ga eerst maar eens je talen bijspijkeren. (Corpus Eindhoven: gezinsbladen)
- (41) ma toen bennen we daar gegaan, maar toen gingen ze hier afbreken, nou, toen zeggen wij tegen mekaar nou zullen we eerst maar eens kijken wat 't wordt. (Corpus Eindhoven: gesproken taal)

Voorlopige conclusie: we kunnen niet echt een conclusie trekken.

Kijken we vervolgens naar een nog langere combinatie, *nou eerst maar eens*. We vragen ons af: is de restrictie tot directieve contexten nog sterker? Als we onze intuïtie raadplegen dan is het in elk geval zo, dat de combinatie zich prima thuis lijkt te voelen in directieven: *ga nou eerst maar eens je handen wassen* en *je moet nou eerst maar eens een stukje gaan slapen* klinken beide onberispelijk.

Als we kwantitatieve ondersteuning willen voor onze intuïtie komen we echter in de problemen: de hele combinatie komt niet voor in CGN tot en met de vierde release. Voor ons partikelproject hebben we in de loop van de tijd een aardige database opgebouwd met onder meer meer dan 10000 voorkomens van *eens*. Daarbij twee voorkomens van de gezochte combinatie – beide van good old Maarten 't Hart:

- (42) 'Kom *nou eerst maar eens* binnen.'
- (Maarten 't Hart, *De nakomer: roman*. Amsterdam; Antwerpen: De Arbeiderspers, 1996.)
- (43) Ga *nou eerst maar eens* naar m'n opvolger, ik zal hem opbellen, dan kan hij alvast een kikker op de schoorsteenmantel klaarzetten. (ibid.)

De veel grotere INL-corpora kunnen ons ook niet echt helpen, want daarin vond ik ook maar drie voorkomens:

- (44) een lach niet onderdrukken en wijst naar de koud geworden koffie op tafel. 'Drink die **NOU EERST MAAR EENS** op...', zegt ze quasi-vermanend. De dominee gunt haar een veelbetekenende knipoog: 'Wat was ik zonder (MCDEC92OVE.SGZ via INL 38M)
- (45) ergste vrezin. Borsboom verklaart erin dat ze 'als vrouw met serieuze vrouwenfilms bezig' is. Begin **NOU EERST MAAR EENS** als menselijk wezen een gooi te doen naar een goede film, dan zien we daarna (NRCFEB1994.KRANT via INL 27M)

- (46) gebrek aan interesse van de bond en over het tekort aan financiële middelen. "Laten we NOU EERST MAAR EENS sport bedrijven. Ik vergelijk het met studenten. Die moeten ook investeren. Zonder dat ze zekerheid (NRCAPR1995.KRANT via INL 27M)

Dan (eindelijk) maar naar het Internet. Volgens Google waren er deze week ca. 158000 Nederlandse pagina's (op het hele web 249000) met voorkomens van *eens* (dus ca. $158 * 10^6$ woorden). De zoekmachine vond ruim 300 hits voor *nou eerst maar eens*. Ik geef een paar voorbeelden van *nou eerst maar eens* in directieve contexten.

- (47) Een dag later spreekt Ha van Da weer wijze woorden. "Laten we nou eerst maar eens dat ding een stukje kopen, dan eens een stukje varen, en dan nog ...
(www.hansvandijk.org/AnitaII/log1.htm)
- (48) het nou leuk vindt of niet... dat pak is van jou, zeikerdje... trek het nou eerst maar eens aan... Het lijkt erop dat ik weinig keus heb. Tsja jongen, ik doe ook ...
(www.deajaxster.nl/mm/mokumman34.pdf)
- (49) ... want het was de hele nacht behoorlijk koud. Nou, eerst maar eens proberen om met stokken het luik los te wrikken. Geen succes. Dan maar met heet water. ...
(www.chello.nl/~n.wolterink/Verhalen.htm)
- (50) ... totaal verlepte geranium glurend. 'Dat dacht ik wel,' zei moeder, 'Eet nou eerst maar eens lekker, en dan heb ik nog een verrassing.' 'O, nee,' kreunde Bastje ...
(members.tripod.com/~dekater/baarlijkenonsens/baarlijkenonsens13.htm)
- (51) enz.

Een nog langere combinatie is *nou eerst maar eens even*. Weer heb ik het gevoel dat die combinatie het natuurlijkst klinkt in directieve zinnen. Google had van de week nog maar vier hits voor deze string:

- (52) "Ga jij nou eerst maar eens even knuffelen want ik denk dat je dat op dit moment wel kunt ... (home.planet.nl/~elvo/draaikolk7p.html)
- (53) in proportie gaan zien en niet meteen alles vergroten. Leg nou eerst maar eens even uit wat je gedaan hebt. We moeten dingen beter uitleggen aan elkaar. Dat ...
(www.nrc.nl/W2/Nieuws/1999/08/06/Vp/01.html)
- (54) Dus weer drie telefoonnummers rijker. ... Nou, eerst maar eens even Entac bellen, misschien weten die een goed nummer. ... Bastiaan
(www.student.utwente.nl/~thibats/actueel/gastenboek.html)
- (55) Wat is icq? Zorg jij nou eerst maar eens even dat je www.beaver-hunter.com webserver up houdt. Wil gelof ik niet helemaal ...
(62.108.19.93/news/article.php/chello.nl.support.network/)

Voorbeeld (54) is alweer twijfelachtig: het kan gelezen als een infinitivus pro imperativo, dus als een directief, maar het kan ook anders: het kan ook zijn dat de spreker uiting geeft aan een voornemen.

Kortom: de weinige voorbeelden laten zien dat ook bij deze lange combinatie de restrictie tot directieven niet absoluut is of hoeft te zijn, en dat het corpus Internet te klein is om gefundeerde uitspraken hierover te doen.

Ik behandel kort *nóg* een voorbeeld van een zeer zeldzame lange combinatie van partikelachtige woordjes: *best nog wel eens*. Naar mijn intuïtie zal die string vrijwel uitsluitend in de omgeving van zwakke modalen voorkomen. In het CGN komt de combinatie niet voor. Het 38 miljoen-corpus van het INL 38 heeft precies één geval van dit rijtje

- (56) dat kan best nog wel eens wat worden met die bond
(Rooie Vrouwen Magazine via INL 38Miljoen)

Dat is zeker geen tegenvoorbeeld, maar ook geen sterke bevestiging.

Het Internet biedt ruim 100 voorbeelden, vrijwel allemaal uit Nederland. Voorbeelden die de hypothese bevestigen zijn in de meerderheid:

- (57) Een leuke reis, die ik best nog wel eens zou willen doen. . .
(www.sindbad.nl/gastenboek.htm)
- (58) Stel je bent 31. Dan kan het best nog wel eens iets worden met een 112-jarige. Sex is er echter niet meer bij.
(wacko.dds.nl/)
- (59) Maar ik zou die meid best nog wel eens een beurt willen geven.
(www.nattenel.nl/winkel.html)

Maar er zijn ook een paar tegenvoorbeelden:

- (60) Internet is soms best nog wel eens traag en dat is vervelend, zeker als je lang op een pagina moet wachten.
(www.wirehub.nl/~ttw/news.htm)
- (61) In sommige gezinnen is deze vraag best nog wel eens een punt van discussie.
(www.ngk.nl/maastricht/studies/gedenk-de-sabbatdag.html)

De beperking van de combinatie *best nog wel eens* tot zwakke modale contexten is dus kennelijk niet absoluut, maar slechts een tendens.

7 De moraal

- Ik hoop (nog een keer) te hebben laten zien dat je het Internet onder meer kunt gebruiken voor het vinden van bewijsmateriaal voor of tegen bepaalde hypothesen over collocatiegedrag van hoogfrequente functiewoorden, zoals partikels.
- Ik hoop ook te hebben laten zien dat je tamelijk gemakkelijk taalkundige vragen kunt stellen waar zelfs het Corpus Internet te klein voor is.

8 Nawoord

- Je kunt bijna niks: 4 hits voor de string *nou eerst maar eens even*, 43000 hits voor de combinatie *nou eerst maar eens even*
- Je kunt via de standaard zoekmachines vrijwel alleen woordgerichte taalkundige vragen stellen

Wat we nodig hebben is een taalkundig verrijkte zoekmachine.

In van Oostendorp & van der Wouden (1998) schreven we al

Het valt echter niet te verwachten dat deze zoekmachines ooit zullen voldoen aan alle eisen die een taalkundige zou kunnen stellen. Het zou daarom wenselijk zijn als instellingen zoals het INL, voor zover ze ambiëren dienstverlenend werk te doen voor taalkundig onderzoekers, zich oriënteerden op de ontwikkeling van eigen, op de taalkunde gerichte, zoekprogramma's voor het Internet-corpus.

Tot mijn verdriet heeft het INL er tot op heden geen blijk van gegeven deze uitnodiging te aanvaarden. Daarom herhaal ik haar hier nog maar eens, en zal ik nog eens proberen uit te leggen wat ik bedoel.

Wat zou zo'n taalkundig verrijkte zoekmachine zoal moeten kunnen?

- retrograde zoeken mogelijk maken (geef me voorkomens van woorden op eindigend op *-erij*)
- morfologische analyse bieden (zodat ik kan vragen naar *-baar*-afleidingen die derivatie hebben ondergaan (zoals *dankbaarheid*).
- zinsafbakening bieden (zodat ik kan zoeken naar zinnen die de partikels *weer nog maar eens even* bevatten, maar niet noodzakelijk in die volgorde)
- POS-tagging bieden (oftewel taalkundige ontleding) (zodat ik kan zoeken naar het focuspartikel *louter* (en niet de eigen naam of de werkwoordsvorm)) en het focuspartikel *enkel* (en niet het lichaamsdeel)
- (oppervlakkige) syntactische analyse (zodat ik kan zoeken naar getopicaliseerde focuspartikels, enz.

...

Dat kan allemaal automatisch voor het Nederlands tegenwoordig – niet 100% correct, maar wel met heel acceptabele resultaten.

Hier ligt een taak! Voor het INL, voor het NIWI, voor het TST-platform ...

References

- CALLEBAUT, INGE, TON VAN DER WOUDE, PIET VAN DE CRAEN, & FRANS ZWARTS. 1998. Er was eens: een partikel. lecture TIN-dag, Utrecht, 17 januari 1998.
- FIRTH, J.R. 1957. *Papers in Linguistics 1934–1951*. London: Oxford University Press.
- FONTENELLE, THIERRY. 1992. Collocation acquisition from a corpus or from a dictionary: a comparison. In *EURALEX '92 Proceedings I–II. Papers submitted to the 5th EURALEX international congress on lexicography in Tampere, Finland*, ed. by Hannu Tommola, Krista Varantola, Tarja Salmi-Tolonen, & Jürgen Schopp, 221–228. Tampere: Department of translation studies, University of Tampere.
- FOOLEN, AD. 1993. *De betekenis van partikels. Een dokumentatie van de stand van het onderzoek met bijzondere aandacht voor maar*. Nijmegen dissertation.
- GEERAERTS, DIRK. 1986. *Woordbetekenis. Een overzicht van de lexicale semantiek*. Leuven/Amersfoort: Acco.
- GEERTS, GUIDO, & TON DEN BOON (eds.). 1999. *Van Dale Groot woordenboek der Nederlandse taal*. Utrecht/Antwerpen: Van Dale Lexicografie. 13e, herz. uitg.
- HAUSMANN, FRANZ JOSEF. 1989-1991. Collocations. In *Wörterbuecher: ein internationales Handbuch zur Lexikographie / Dictionaries: an international encyclopedia of lexicography / Dictionnaires: encyclopedie internationale de lexicographie*, ed. by Franz Josef Hausmann et al., I: 1010–19. Berlin [etc.]: De Gruyter. (Handbücher zur Sprach- und Kommunikationswissenschaft; Bd. 5).
- 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.
- LEMNITZER, LOTHAR. 2001. Wann kommt er denn nun endlich zur Sache? Modalpartikel-Kombinationen – Eine korpusbasierte Untersuchung. In *Sprache im Alltag. Beiträge zu neuen Perspektiven in der Linguistik. Herbert Ernst Wiegand zum 65. Geburtstag gewidmet*, ed. by Andrea Lehr, Matthias Kammerer, Klaus-Peter Konerding, Angelika Storrer, Caja Thimm, & Werner Wolski, 349–371. Berlin and New York: Walter de Gruyter.
- MANNING, CHRISTOPHER D., & HINRICH SCHÜTZE. 1999. *Foundations of statistical natural language processing*. Cambridge, Mass.: The MIT Press.
- NOORDEGRAAF, JAN. 2000. *Van Hemsterhuis tot Stutterheim: over wetenschapsgeschiedenis*. Münster: Nodus.
- VAN OOSTENDORP, MARC, & TON VAN DER WOUDE. 1998. Corpus internet. *Nederlandse Taalkunde* 3, 347–361.
- OED. *The Oxford English dictionary: being a corrected re-issue with an introduction, supplement, and bibliography of "A new English dictionary on historical principles"*; founded mainly on the materials coll. by the Philological Society; ed. by James A.H. Murray [et al.]. Oxford: Clarendon Press, 1970.

- PAARDEKOOPEL, P.C. [n.d.]. *Beknopte ABN-syntaksis*. Eindhoven: Uitgave in eigen beheer. Zevende druk.
- VAN DE POEL, KRIS, & L. VAN DE WALLE. 1995. Nederlandse partikels en andere kleine woorden in het kader van beleefdheidsstrategieën. In *Nederlands in culturele context. Handelingen Twaalfde Colloquium Neerlandicum, Universitaire Faculteiten Sint-Ignatius Antwerpen, 28 augustus – 3 september 1994*, ed. by Th.A.J.M. Janssen, P.G.M. Kleijn, & A.M. Musschoot, 325–343. Antwerpen: Universitaire Faculteiten Sint-Ignatius Antwerpen.
- RENKEMA, JAN. 1989. *Schrijfwijzer*. 's-Gravenhage: SDU uitgeverij. Volledig herz. ed. (Oorspr. uitg.: 1979).
- RULLMANN, HOTZE, & JACK HOEKSEMA. 1997. De distributie van *ook maar* en *zelfs maar*. een corpusstudie. *Nederlandse taalkunde* 2, 281–317.
- SINCLAIR, JOHN. 1991. *Corpus, concordance, collocation*. Oxford [etc.]: Oxford University Press.
- THURMAIR, MARIA. 1989. *Modalpartikeln und ihre Kombinationen*. Tübingen: Niemeyer.
- VISMANS, ROEL. 1994. *Modal particles in Dutch directives: a study in functional grammar*. Vrije Universiteit Amsterdam dissertation.
- DE VRIENDT, SERA, WILLY VANDEWEGHE, & PIET VAN DE CRAEN. 1991. Combinatorial aspects of modal particles in Dutch. *Multilingua* 10, 43–59.
- DE VRIES, MATTHIAS, & LAMMERT A. TE WINKEL ET AL. (eds.). 1864–1998. *Woordenboek der Nederlandsche taal*. 's-Gravenhage [etc.]: Martinus Nijhoff [etc.].
- Woordenlijst. 1995. *Woordenlijst Nederlandse Taal*. Den Haag, Antwerpen: Sdu, Standaard. samengest. door het Instituut voor Nederlandse Lexicologie; met een leidraad door Jan Renkema.
- VAN DER WOUDE, TON. 1992. Beperkingen op het optreden van lexicale elementen. *De Nieuwe Taalgids* 85, 513–38.
- . 1994. *Negative Contexts*. Groningen dissertation.
- . 1997. *Negative Contexts. Collocation, polarity, and multiple negation*. London and New York: Routledge.
- . 1999. Partikelvariation, Textvariation, Sprachvariation. Vortrag FU Berlin 15.10.1999.
- . 2000. Prototypicality vs. variation: Restrictive focus particles in Dutch. Paper delivered at Discourse Particles, Modal And Focal Particles And All That Stuff . . . , An international conference on Particles, Brussels, 8–9 December 2000.
- . 2001a. Collocational behaviour in non content words. In *COLLOCATION: Computational Extraction, Analysis and Exploitation. Proceedings of a Workshop during the 39th Annual Meeting of the Association for Computational Linguistics and the 10th Conference of the European Chapter, Toulouse, France, July 7th*, ed. by Béatrice Daille & Geoffery Williams, 16–23. Toulouse, France: CNRS – Institut de Recherche en Informatique de Toulouse, and Université de Sciences Sociales.

—. 2001b. Partikels: naar een partikelwoordenboek voor het Nederlands. lezing TIN-dag, Utrecht, 3 februari 2001, te verschijnen in *Nederlandse Taalkunde*.

ZWARTS, FRANS, & TON VAN DER WOUDE. 2000. The various faces of Belgian and Dutch 'eens'. Paper delivered at "Making Sense. From Lexeme to Discourse. A Conference in Honour of Werner Abraham on the Occasion of his Retirement", Groningen, November 6–8, 2000.