Particle research meets corpus linguistics: on the collocational behavior of particles

Ton van der Wouden

UiL-OTS, Trans 10, 3512 JK Utrecht, The Netherlands[*]

vdwouden@let.uu.nl

22nd February 2002

Abstract

Modal particles may show strong tendencies to cluster. In this paper, we investigate the possibilities of giving a sound, quantitative, corpus-based foundation for this generalization. We discuss some standard techniques from collocation research and apply them to Dutch particles of various types. We moreover show that the methodologies deployed can help us find other collocational properties of particles.

1   Introduction: particle clusters

It is a classical observation that modal particles, especially those found in the mainland Germanic languages, may exhibit strong tendencies to cluster.[1] The Dutch author Hoogvliet discusses the spectacular six member cluster *dan nu toch maar eens even* in the seminal section on particles of his (1903) book *Lingua*. An almost equally astonishing German cluster *doch nur ruhig auch mal* is offered by Thurmair (1991).[2]

(1) a. *Geef de   boeken dan   nu   toch   maar   eens   even   hier* (Hoogvliet 1903)
      Give the books   PART PART PART PART PART PART here
      'Now give us the books'

   b. *Kombinieren Sie   doch   nur   ruhig   auch   mal   Modalpartikeln!* (Thurmair 1991)
      Combine   you PART PART PART PART PART modal particles
      'Feel free to combine modal particles'

Native speakers of Dutch and German, respectively, will agree that the sentences in (1) are perfectly grammatical, although (it goes without saying) such long clusters are found in real life very seldom. Shorter clusters, however, are far from rare.

[1]Parts of the material in this paper were presented earlier in Marburg (van der Wouden 2000a), Groningen (van der Wouden 2000b) and Toulouse (van der Wouden 2001a). Cf. also Lemnitzer (2001) for comparable approaches to particle clustering.

[2]For reasons of untranslatability we usually refrain from glossing combinations of modal particles: they are represented as PART. We did, however, do our best to come up with more or less idiomatic translations.

In this paper, we investigate the possibilities of giving a quantitative, corpus-based foundation for the aforementioned native speaker's intuitions. Standard methodologies from the field of computational collocation research, usually deployed to find relations between content words, will be shown to be able, in principle, to trace both expected and unexpected clustering properties of particles. Moreover, standard collocation search strategies can help us to find other collocational properties of particles as well.

2   Collocations

It is often claimed (e.g. Matthews (1981)) that the British linguist Firth introduced the term COLLOCATION to refer to the habitual accompaniment of one word by another. Although one finds earlier instances of the term in the linguistic literature (cf. van der Wouden (1997:7–8)), it was Firth who popularized both the term and the study of the subject. According to Palmer (1968:6–7), 'Firth [himself] seems [. . .] to have restricted himself to specialized collocations – to *silly* with *ass*, to *cow* with *milk*. [. . .] But he does not seem to have extended his theory to comprehend the whole of the problems of lexical compatibility or to have seen that the kind of formal grammatical analysis which he recommended is dependent upon the recognition of mutually collocable classes of lexical items.'

In the recent literature, one finds two types of definitions for collocation: restricted ones and general ones. An example of a restricted definition of the concept is given in (2):

(2) On appelera 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)

Note that the restriction to certain structures remains unmotivated.

An unrestricted, very general definition of collocation is given in (3):

(3) The term COLLOCATION refers to the idiosyncratic syntagmatic combination of lexical items and is independent of word class or syntactic structure (Fontenelle 1992)

In computational linguistics, a common way of approaching the linguistic concept of collocation is quantificational. The simplest way of course is to just look at frequently occurring combinations.

We will try to demonstrate this with the Dutch (as opposed to Belgian) subpart of the second release of the Spoken Dutch Corpus, which is currently under development.[3] The size of this Dutch subcorpus is

---

[3]The aim of the Spoken Dutch Corpus project (abbreviated as CGN, from the Dutch name *Corpus Gesproken Nederlands*) is to build an annotated corpus of about one thousand hours of continuous speech, which amounts to 10 million words. The project started in June 1998, and runs for five years. It is a collaborative effort of several Dutch and Flemish universities (Oostdijk 2000; Hoekstra *et al.* 2001b; Hoekstra *et al.* 2001a).

The corpus is intended as a major resource both for linguistic research and for language and speech technology. To serve this dual purpose, it contains materials recorded in a variety of communicative settings: spontaneous face-to-face and telephone dialogues, interviews, discussions, debates, lectures, news broadcasts and book passages read aloud. Two-thirds of the material is collected in

currently somewhat over two million (2397289) words, which will be shown to be large enough for our main argument.[4]

As an efficient means to investigate this corpus we use, among other things, the "Bigram Statistics Package" (BSP) developed by Ted Pedersen and Satanjeev Banerjee of the University of Minnesota, Duluth. It is a library of routines written in Perl allowing the researcher to efficiently extract bigrams from a corpus and apply various statistical metrics on these bigrams, among other things.[5]

The usefulness of the BSP tools can be demonstrated with a textbook example of collocation, the fixed preposition that comes with the Dutch adjective *dol* 'fond', which is *op*.[6] When only bigrams that occur at least five times are taken into account, the program comes up with exactly one bigram with the word *dol* (text frequency 22), and that is *dol op*, as expected:

(4)

bigrams with *dol* ($N = 22$)	
bigram	N
dol op 'fond of'	18

The results become considerably more messy if we repeat this exercise with *luisteren* 'listen' (which may be both the infinitive form of the verb and first, second or third person plural present tense). We expect strong collocational effects with the preposition *naar*. The actual numbers are given in the table in (5).

(5)

bigrams with *luisteren* ($N = 215$)	
bigram	N
luisteren . 'listen .'	75
je luisteren 'you (SG) listen'	46
te luisteren 'to listen'	44
eens luisteren 'PART listen'	19
luisteren en 'listen and'	17
luisteren ik 'listen I'	12
luisteren naar 'listen to'	10
en luisteren 'and listen'	9
luisteren of 'listen whether'	7
gaan luisteren 'go listen'	6
goed luisteren 'well listen'	5
luisteren dat 'listen that'	5

To say the least, this is not completely what we expected. *Luisteren naar* is in the list of bigrams, but certainly not in first position. The 'collocation' *luisteren .* (in first place) shows that the program interprets punctuation marks such as full stops as words too. The combination's top position in the table may be taken as an indication that the string *luisteren* is rather often found in sentence final position. A few examples may help to illustrate this usage:

---

the Netherlands, one third in the Dutch speaking part of Belgium. Upon completion, the corpus will be the largest and most diverse database of spoken Dutch collected so far. Cf. the project's web site `http://lands.let.kun.nl/cgn`.

Since there are considerable differences in particle usage between the Netherlands and Belgium (van der Wouden 2001b), we only take the material from the Netherlands into account in this paper.

[4]Example sentences in this paper without a source given are all from this corpus.

[5]These BSP programs are free software under the terms of the GNU General Public License; the code can be found at the internet address `http://www.d.umn.edu/~tpederse/code.html`.

Comparable program packages are WordSmith tools by Mike Scott (`http://www.oup.co.uk`) and Michael Barlow's Mono-Conc (`http://www.ruf.rice.edu/~barlow/mono.html`).

[6]For simplicity of presentation, we restrict ourselves to adjacent two-word collocations ('bigrams') for the time being.

(6) a. *ik ben een lerares   en jullie   moeten naar mij luisteren.*
     I   am   a   female teacher and you (PL) must   to   me   listen
     'I am a teacher and you must listen to me'

   b. *terwijl mijn moeder buiten   wachtte en   stond te luisteren.*
      while   my   mother outside waited   and stood to listen
      'While my mother waited outside, listening'

Je luisteren, the second combination in the table, is unexpected as well. Inspection of the data shows that nearly all (45 out of 46) cases involve the string *moet je luisteren* 'must you listen' which is a weak directive asking for attention:[7]

(7) a. *moet je   luisteren ik heb nog een voorstel hè.*
        must you listen   I   have yet a   proposal PART
        'hey listen I have another proposal'

   b. *er   zijn ook jongens die zeggen van moet je   luisteren waarom zal ik lezen?*
      there are   also boys   that say   of   must you listen   why   will I   read?
      'there are also boys that say: 'hey listen why would I read?'

The next combination in the table in (5) involves the complementizer *te*, which is not unexpected given that *luisteren* can function as an infinitive form of the verb (cf. example (6b)). The next one, however, *eens luisteren*, with the particle *eens* (etymologically related to English *once*; cf. Zwarts & van der Wouden (2000)) is noteworthy again. The data show that no less than 16 of the 19 corpus cases involve the string *moet je eens luisteren*, i.e., an elaborate variant of the combination *moet je luisteren* discussed above:

(8) a. *moet je   eens luisteren dan   bel ik dus   aan.*
        must you PART listen   then call I   PART   on
        'hey listen: so I ring the bell'

   b. *ik bel die uitgever   op moet jij eens luisteren dat slot van dat stuk dat*
      I   call that publisher up must you PART listen   that end of   that piece that
      ken   ik helemaal niet
      know I   totally   not
      'I call that publisher: 'hey listen: I don't know the end of that paper''

   c. *kom   nou   eens luisteren.*
      come PART PART listen
      'now come and listen'

The other bigrams in (5) are hardly worth discussion. The frequencies of the bigrams *luisteren en* 'listen and' and *en luisteren* show that *luisteren* is rather often part of a coordination structure (e.g. 9a, 9b).

(9) a. *voorzitter   ik probeer goed te luisteren en   ondertussen ook te lezen.*
        chairperson I   try   good to listen   and meanwhile   also to read
        'chairperson I try to listen well and to read at the same time'

---

[7]To put it more strongly, the combination appears to function often as a kind of discourse marker in the sense of Schiffrin (1987) and Fraser (1999): in this case, the speaker expresses the wish to make a statement and does not want an intervention. That is, (s)he is taking the floor and tries to keep it.

b. *dan   ben   je   dus   wel   bezig   met   vaardigheden   als   spreken   en   kijken   en*
  then   are   you   PART   PART   busy   with   skills   as   speak   and   look   and
  luisteren
  listen
  'then you are working on skills such as speaking, looking and listening'

In most instances of the combination *luisteren ik*, *luisteren* is part of the combination *moet je luisteren* discussed above: it turns out that it is often used to introduce a statement in which the speaker is involved: (7a) above is a typical example. The frequency of *luisteren of* is caused by the combination of two different kind of usages of *of*: it can either be a coordinator ('or') or a complementizer ('whether, if'). Finally, most cases of *luisteren dat* involve cases of *moet je luisteren* again.

It will be clear from the discussion of the bigrams of *luisteren* in (5) above that just looking for high frequency bigrams may yield more junk than collocations in the intuitive, linguistic sense. Therefore, various heuristics have been proposed for getting better results in finding collocations automatically (Manning & Schütze 1999:chapter 5). The following methods will be discussed briefly now.

- look beyond adjacent bigrams

- use so-called phrase filters or stop word lists

- use more sophisticated statistics

2.1 Beyond adjacency

If we look at example (6a) once again, we see that there is a word (*mij* 'me') between the fixed preposition *naar* we were looking for and *luisteren*, the head word (or 'node' in the terminology of Sinclair (1991)). This illustrates that not all collocational relationships are strictly local in the sense that the collocants are necessarily adjacent. One way to find such non-adjacent pairs is to take more of the context into account. The WordSmith tools, for example, implement this by taking a 5 word window to the left and to the right of the head word.[8] The output for the string *luisteren* is then as in (10).[9]

(10)

n	word	total	left	right	l5	l4	l3	l2	l1	*	r1	r2	r3	r4	r5
Collocates of *luisteren* according to WordSmith
1	luisteren	215	0	0	0	0	0	0	0	215	0	0	0	0	0
2	moet 'must'	85	76	9	3	3	22	48	0	0	1	2	4	0	2
3	naar 'to'	50	37	13	4	7	12	9	5	0	8	2	0	3	0
4	het 'the, it'	42	11	31	2	2	3	2	2	0	0	8	9	7	7
5	een 'a, one'	40	15	25	6	2	3	4	0	0	0	3	4	7	11
6	dat 'that'	34	4	30	0	2	2	0	0	0	3	9	7	7	4
7	nou PART	32	17	15	1	9	5	2	0	0	0	5	4	5	1
8	maar PART	29	21	8	3	4	10	1	3	0	1	1	1	4	1
9	eens PART	26	24	2	0	1	1	2	20	0	0	0	0	1	1
10	van 'of', CMP	25	12	13	4	7	1	0	0	0	1	0	2	6	4
11	die 'that'	24	11	13	2	5	1	3	0	0	1	4	3	3	2
12	niet 'not'	24	19	5	6	1	4	6	2	0	0	0	1	2	2
13	dan PART	23	7	16	4	1	2	0	0	0	4	4	4	2	2
14	goed 'well'	16	12	4	0	0	4	3	5	0	0	0	1	2	1

---

[8]This default value of 5 is empirically (rather than theoretically) motivated: cf. Manning & Schütze (1999:chapter 5). This value of 5 may turn out to too small for a verb second language such as Dutch, but that is beyond the scope of this paper.

[9]For expository purposes, only the top of the table is given. lx and rx refer to positions x to the left and right, respectively, of the head word, so l1 is the position immediately to the left of *luisteren*.

This table should be read as follows. Collocates are ordered according to importance. '2' tells us that *moet* is found 85 times in the environment (i.e., at most five words tot the left or to the right) of *luisteren*: 76 times to the left, 9 times to the right. In 3 cases where *moet* was to the left of *luisteren*, there were 4 words in between; there were also three cases with 3 words in between; in 22 cases there were 2 words in between, and 48 cases with one word; finally, it never happened that *moet* was immediately to the left of *luisteren*.

The next line in the table in (10) shows that this method yields the preposition *naar* as a very important collocate of *luisteren* as well, which is of course what we wanted to demonstrate in the first place.

2.2 Phrase filters and stop word lists

When we continue our explorations of the table in (10), the lines below '3' do not look very interesting again. According to many collocations researchers, there is a multitude of such combinations of content words plus function words that keep on popping up (because of the mere frequency of the functions words) time and again. For example, most combinations of determiner plus noun, or of preposition and determiners, are uninteresting for most collocation research purposes. Therefore, researchers have proposed mechanisms to separate the interesting bigrams from the uninteresting ones.

A modification which has been claimed to be very effective is to filter the collocations and remove those that have parts of speech or words that are rarely associated with interesting collocations. There simply are no interesting collocations that have a preposition as the first word and an article as the second word. On the other hand, many interesting two word collocations in English take the form 'adjective noun' and 'noun noun' (Manning & Schütze 1999:31).

A very simple heuristic that has been proposed and that supposedly improves the results considerably is to pass the candidate phrases through a part of speech filter which only lets through those patterns that are likely to be 'phrases' (Ross & Tukey 1975; Justeson & Katz 1995). For example, one may choose to accept only combinations that can be labeled A N (e.g. *linear function*) or N P N (*degrees of freedom*).[10]

Another filtering method proposed (Smadja & McKeown 1990) is a list of so-called "stop words", usually high frequency function words such as determiners and prepositions, which are neglected as parts of higher than chance bigrams and N-grams. Again, this method has proven to be successful for "classical" collocations (Manning & Schütze 1999).

Note, however, that both phrase filters and stop word lists may turn out to be pretty useless in light of the main purpose of this paper, viz., the automatic detection of collocational behavior in particles. In all except the most fine-grained part of speech tagging systems, particles will be assigned an adverbial label. Apart from the clusters of particles as exemplified in (1) which raised the interest of the collocation research community only recently (van der Wouden 2000a), one hardly ever finds mention (let alone discussion) of collocation pairs consisting of two (or more) adverbs (cf. also the definition in (2)). Particle clusters as in (1), be they of collocational nature or something else, will therefore never pass standard phrase filters.

Comparable reasoning holds against (standard) stop words lists, which usually consist mainly of high frequency words, among other things. Particle clusters of the type exemplified above will be filtered out immediately, as most particles are exactly this: high frequency words (van der Wouden 2001a). By way of illustration: the six members of the particle cluster in example (1a) are all in the top 100 of the most frequent strings in our corpus: there are 28450 instances of *dan* (16th in rank), 5377 of *nu* (75th), 7840 of *toch* (57th), 31415 of *maar* (14th), 4091 of *eens* (94th), and 3889 of *even* (96th).[11]

---

[10]This method is, in a sense, an implementation of the kind of narrow definition of collocation discussed earlier.

[11]Counts due to WordSmith.

2.3 More sophisticated statistics

A different way of trying to distinguish interesting bigrams (possible collocations) from uninteresting ones is to use more complex arithmetic tools than pure frequency of the combinations. Three popular metrics, that also take into account factors such as the frequencies of the constituent parts, are chi-square (χ^2), loglikelihood (ll) and mutual information (mi) (cf. Manning & Schütze (1999:chapter 5) for exact definitions and discussion). The table in (11) gives the outcome of the three formulas[12] for the frequent bigrams with *luisteren* in our corpus, ranked according to χ^2.

(11)

bigrams with *luisteren* ($N = 215$)				
bigram	N	χ^2	ll	mi
te luisteren 'to listen'	44	2039	271	5,6
eens luisteren 'PART listen'	19	1763	138	6,6
je luisteren 'you listen'	46	960	216	4,5
luisteren . 'listen .'	75	295	170	2,4
luisteren naar 'listen to'	10	234	46	4,7
gaan luisteren 'go listen'	6	125	26	4,5
goed luisteren 'well listen'	5	79	19	4,2
luisteren en 'listen and'	17	77	36	2,7
luisteren of 'listen or/whether'	7	73	23	3,6
luisteren ik 'listen I'	12	24	14	1,9
en luisteren 'and listen'	9	11	7	1,5
luisteren dat 'listen that'	5	1	1	0,5

It seems that the results are improving. We observe for example that the rather uninteresting (from a collocation point of view) combination of *luisteren* with the full stop, which ranked 1 in the raw frequency table in (5), is in 4th position according to chi-square, 3rd for loglikelihood, and 9th according to mutual information. The combination *te luisteren* on the other hand, which definitely points to an important usage of the string *luisteren* (viz. as an infinitive), ranks high to very high according all tests, and the same holds for the combination with the particle *eens*.

In the rest of the paper we will see whether the methods discussed so far yield interesting results with respect to collocational behavior of particles. We may, however, expect that none of the measures will be perfect in all situations (in terms of corpus size, frequent or infrequent data, etc.). It is an open question whether it is possible, in principle, to prove that some metric is the best (for a certain type of collocation, or in general) (Krenn & Evert 2001).

3 Collocational behavior in particles

In this section, we use the methods described above in order to try and find collocational behavior in various types of particles.

3.1 Example 1: modal particles

Let us first look at modal particles. Most modal particles (or their homographs) have other functions as well, which may blur the picture somewhat. The aforementioned *eens*, however, is almost always a modal

---

[12]Calculations by BSP tools.

particle.[13] Etymologically related to the numeral *een* 'one' and English *once*, it is originally an existential quantifier over time. As a modal particle, it may, among other things, function as a counterpart of the Slavic imperfective aspect (Zwarts et al., this volume). The table in (12) shows the most frequent bigrams.

(12)

bigrams with *eens* ($N = 4051$)				
bigram	N	χ^2	ll	mi
wel eens 'PART PART'	961	27185	4983	4,9
eens een 'PART a'	525	3045	1284	2,9
niet eens 'not PART'	433	3046	1164	3,1
nog eens 'PART PART'	405	6898	1665	4,2
eens . 'PART .'	243	44	49	-0,6
eens even 'PART PART'	243	10881	1427	5,5
maar eens 'PART PART'	240	685	367	2,2
eens wat 'PART what/something'	117	472	222	2,5
weer eens 'PART PART'	109	1391	387	3,9
eens in 'PART in'	108	133	89	1,5
nou eens 'PART PART'	105	241	136	2,5
eens kijken 'PART look'	105	4378	598	5,4

We note that almost all frequent collocates of the particle *eens* are other particles. The first combination, *wel eens*, has specialized into something that seems to be counterpart of the Slavic perfective aspect (Zwarts et al., this volume). According to certain prescriptive sources (de Vries & te Winkel et al. 1864–1998; Renkema 1989), it should be written as one word *weleens*. This advice, however, has not been followed consistently in the CGN.

(13) a. *we hebben er   wel   eens   twaalfhonderd gehad.*
          we   have   there   PART   PART   1200   had
          'we once had 1200 of them'

      b. *m'n moeder die   maakt 't   ook   weleens   zelf.*
          my   mother that makes it   also PART-PART   self
          'my mother she makes it herself every once in a while'

The second combination, *eens een*, appears to be an exception to the generalization that *eens* usually collocates with other particles. Closer inspection, however, shows that more than half of the cases involve the combination with *een keer* 'a time', which functions as a modal particle as well:[14]

(14) a. *ik mocht ook   eens   een keer zo'n   vergadering bijwonen*
          I   could also PART   a   time such-a meeting   attend
          '(once) I was allowed to attend one of these meetings too'

---

[13]Cf. Callebaut *et al.* (1998). In the ca. 1 million Eindhoven corpus from the 1970's (the combined sources for Uit den Boogaart (1975), de Jong (1979) and Renkema (1981)) the string is 886 times labeled 500 (for adverbial usage) and 89 times 100 (for adjectival usage, as in *zij zijn het eens* 'they agree'). Only in Flemish, the Belgian variant of Dutch (which is left out of consideration here) *eens* can also function as a complementizer, just like English *once*, as in *de garantie dat ze een flat krijgen eens ze die nodig hebben* 'the guarantee that they will get an apartment once they need it'

[14]Apart from *eens* en *een keer*, there are more modal particles deriving from existential temporal quantifiers, e.g. German *mal* (< einmal 'one time') and Dutch *ereis* (< een reis 'one time').

b. *'t lijkt   me sowieso wel   lekker dat je   gewoon eens   een keer wat*
   it seems me PART   PART nice   that you simple PART a   time somewhat
   meer tijd hebt.
   more time have
   'it would just be very nice to have some extra time'

The third combination, *niet eens*, has developed into a radically different direction, as it usually functions as a negative focus particle along the lines of 'not even' (van der Auwera 1992).[15] The possibility of combining *niet eens* with a predicate such as *weten* 'know' as in (15b) shows that the original function of *eens* as an existential temporal quantifier has disappeared completely: in general, it is impossible to know once.

(15) a. *d'r   zijn ook mensen bij   die niet eens   gealfabetiseerd zijn.*
         there are also people with that not PART alphabetized   are
         'it also includes people who don't even know how to read and write'

      b. *terwijl ze   niet eens precies weten waar ze   over praten*
         while they not PART exactly know where they over talk
         'while they don't even exactly know what they are talking about'

The combination with the temporal particle *nog* can mean 'again', among other things:

(16) a. *ik moet er   nog   eens   over nadenken.*
         I must there PART PART about think.
         'I have to think about it'

      b. *er   wordt vergaderd en   er   wordt nog   eens   een keer vergaderd*
         there is   assembled and there is   PART PART PART   assembled
         'they have meetings and they have more meetings'

Space forbids to pay more attention here to the other combinations in table (12), which involve *even* (literally 'for a short while), polyfunctional *maar*,[16] *wat* ('what' functioning as an indefinite pronoun (Haspelmath 1997)), *weer* ('again'), the preposition *in* 'in', *nou* (originally temporal: 'now'; the combination often expresses impatience) and the verb *kijken* 'look'. I just give some examples:

(17) a. *nou   ik zal eens   even   vragen.*
         PART I wil PART PART ask
         'well I'll ask'

      b. *maak maar eens   een zin   van zes woorden.*
         make PART PART a   sentence of six words
         'try to make a six word sentence'

      c. *is weer eens   wat   anders dan vliegen in een vliegtuig.*
         is PART PART something else   than flying in a   plane
         '[it] makes a change from flying by plane'

---

[15]As far as we know, *eens* functions only as a focus particle when immediately following *niet* and in the fixed combination with a negative quantifier *geeneens* which has the same meaning and function as *niet eens* but belongs to a somewhat more colloquial register (Geerts & den Boon 1999). Given this non-compositional meaning, one would expect the combination *niet eens* to be written as one word, but that occurs very rarely. The only writer we know who does it consistently is Albert Helman. An example: *De ander hoorde het nieteens.* (*Het vergeten gezicht.* Rotterdam: Nijgh & Van Ditmar, 1939)

[16]*Maar* may function, among other things, as a coordinator ('but'), as a scalar focus particle ('not more than'), and as a modal particle: cf. Foolen (1993) for extensive discussion. We will return to the combination *maar eens* later.

d. *gemiddeld bekijken we zo   eens in de twee weken wel   een voorstelling.*
  averaged   observe we PART once in the two weeks PART a performance
  'on the average, we watch a performance every two weeks or so'

e. *ga nou   eens   een keer over inhoud   spreken*
  go PART PART PART   over content talk
  'start talking about content please'

f. *laten we eens   kijken of we met een discussie   d'ruit   kunnen komen.*
  let   we PART see   if we with a discussion there-out can   come
  'let's see whether we can solve it by talking'

3.2 Example 2: restrictive focus particles

Having shown that standard techniques to find collocational behavior in an automated way works quite well in the case of modal particles, we will now see whether we may find interesting results in the case of focus particles as well. We will start by looking at *alleen* 'only', which occurs 2815 times in the corpus.[17] The table in (18) gives the top of the most frequent bigrams according to the BSP program.

(18)

bigrams with *alleen* ($N = 2815$)				
bigram	N	χ^2	ll	mi
alleen maar 'only PART'	739	17467	3632	4,7
. alleen	384	63	56	0,5
niet alleen	355	3749	1187	3,6
alleen de 'only the'	136	64	50	1,0
je alleen 'you only'	96	111	75	1,5
alleen nog 'only PART'	81	371	166	2,7
alleen een 'only a'	76	17	14	0,7
. . .				
hoeft alleen 'need only'	12	325	58	4,9

One seldom finds discussion of clustering behavior of focus particles. Still, the first line in the table is impressive: it shows that in almost one quarter of all cases of *alleen* in our corpus the word is immediately followed by the particle *maar*.[18] Both combinations *alleen maar* and *niet alleen* are already mentioned[19] in the entry *alleen* of the historical dictionary *Woordenboek der Nederlandsche Taal* (de Vries & te Winkel et al. 1864–1998) dating from 1898. In *alleen maar*, the main function of *maar* seems to be a form of rhetorical strengthening (19).[20]

(19)   *'t gaat alleen maar over   treinreizen.*
        it goes PART PART about train-travels
        'it is about nothing but travelling by train'

---

[17]There is also usage of *alleen* as a predicate (*alone*), but that is rare: in the Eindhoven corpus (cf. note 13) some 10 percent of the strings *alleen* carries an adjective label, and over 90 percent an adverbial one.

[18]Cf. note 16.

[19]The first one in a depreciatory way, cf. the next note.

[20]"Ook dient ter versterking, in de volksspraak, het gelijkbeteekenende *maar*, vóór of achter *alleen* geplaatst", i.e. 'in the vernacular the synonymous *maar*, placed before or after *alleen*, is used for strengthening' (*WNT* s.v. *alleen* III).

The second line in the table in (18) shows that *alleen* occurs sentence-initially rather often. However, the various statistical tests uniformly attribute a low value to the combination of full stop and *alleen*, which suggests that the effect is not too interesting.[21]

Niet alleen has two important uses: in cases as exemplified in (20a), the meaning is more or less compositional, whereas in cases such as (20b) the combination is part of a larger coordination construction *niet alleen . . . maar ook* 'not just . . . but also':

(20) a. *ik ga niet alleen les   geven*
         I   go   not   only   lesson give
         'I'll not just teach'

      b. *niet alleen programma's worden opgenomen maar ook discussies   bijvoorbeeld.*
         not   PART   programs   are   recorded   but   also discussions for example
         'next to programs, they also record discussions for example'

The rest of the table in (18) does not look very interesting. The combination *hoeft alleen* in the last line, however, deserves special attention (note that the value for the mutual information metric is extremely high). *Alleen* turns out to function here as the licensor for the negative polarity auxiliary *hoeven* 'need' (21a) (cf. van der Wouden (2001c)).[22] The example in (21b) shows that *alleen maar* can do the same job.

(21) a. *da's   heel saai   werk je   hoeft alleen maar op   knopjes te   duwen*
         that's   very boring work you need   only   PART on   buttons to   press
         'that is very boring work you only need to press keys'

      b. *die   computers staan daar aan dus je   hoeft alleen maar Hotmail in te tikken.*
         those   computers stand there on   so   you need   only   PART Hotmail in to type
         'the computers there are switched on so you only need to type in Hotmail'

The focus particle *slechts*, although slightly more formal (Geerts & den Boon 1999) and less frequent than *alleen*, may replace *alleen* in a great number of cases without change of meaning. However, this suggests that the words are (almost) synonymous. The bigram table in (22), however, shows that the distributional properties of *slechts* are considerably different from those of *alleen*. For example, where the combination with *maar* was the most frequent in the case of *alleen*, *slechts maar* is completely absent from our corpus.[23] This implies that these distributional properties are truly collocational in nature in the sense that is very hard to derive them from syntactic or semantic properties of the items involved.

(22)

bigrams with *slechts* ($N = 124$)				
bigram	N	χ^2	ll	mi
slechts een 'PART a'	24	675	140	4,8
. slechts	14	42	26	2,1
slechts één 'PART one'	8	178	37	4,6
nog slechts 'PART PART'	7	314	42	5,5
er slechts 'there PART'	5	218	29	5,5
slechts op 'PART on'	5	64	18	3,9
slechts de 'PART the'	5	10	6	1,8

---

[21]A comparable effect is found for the additive particle (König 1991) *zelfs* 'even': of all bigrams with *zelfs*, the one with the full stop has the highest frequency in the corpus, but the significance is low according to the three statistical tests.

[22]In van der Wouden (1997) I have tried to argue at great length that the behavior of polarity items is collocational, in a sense. The fact that standard automated collocation search techniques come up with a combination of a polarity item and its licensor is a new, additional argument for this position.

[23]This is corroborated by the fact that for most speakers it sounds awkward or worse.

Another thing from this table that strikes the eye is that *slechts* is found more often right in front of indefinite determiners than definites, whereas the situation is the other way round in the case of *alleen*.

(23)   a. *ze   heeft   ons   slechts   een   notitie   toegestuurd*
           she   has   us   PART   a   note   to-sent
           'she only sent us a note'

       b. *die   garanderen   slechts   een   bepaald   minimum   aan   mensenrechten*
           those   guarantee   PART   a   certain   minimum   to   human rights
           'they only guarantee a minimum of human rights'

We again do not know what causes this difference, but it certainly merits further investigation, which goes beyond the scope of this paper.

3.3   Example 3: collocational behavior of complex focus particle

In this section we will show that clusters around particles may have their own collocational behavior as well. Let us first reconsider the combination *niet eens*, which usually functions as a focus particle (cf. above). If we feed this combination into the WordSmith program, we get interesting results again (see table (31) in the appendix). The first collocate to note is the temporal particle *nog*. Of the 83 times this word is found in the five word window around the 425 occurrences of *niet eens*, it occurs immediately in front of *niet eens* in no less than 75 cases! In the example in (24a) the temporal aspect of *nog* can still be recognized, whereas the particle's main contribution to the sentence in (24b) seems to be a scalar aspect.

(24)   a. *daar   hebben   we   het   dan   hier   nog   niet   eens   over*
           there have   we   it   PART hier   PART   not   PART   over
           'we are not even talking about that yet'

       b. *het   is   nog   niet   eens   een   half   miljard*
           it   is   PART   not   PART   a   half   billion
           'it is not even half a billion'

A comparably skewed distribution is shown by the particle *meer* 'anymore' (which is a negative polarity item, incidentally). There are 57 occurrences of *meer* in the window around *niet eens*, and *meer* is found immediately behind the combination in 45 of the cases.

(25)   a. *ik   kan   me   niet   eens   meer   herinneren   waar   we   het   over   [hadden]*
           I   can   me   not   PART   PART   remember   where   we   it   over   had
           'I can't even remember anymore what we were talking about'

       b. *die   ene   was   echt   zo   vet   die   kon   niet   eens   meer   lopen.*
           that one   was   really so   fat   that could not   PART   PART   walk
           'the one was so fat it couldn't even walk anymore'

Niet eens does not collocate exclusively with particles and other one syllable function words: the first content word in the collocation table is *weet*, a form of the verb *weten*. It is found in 13th position in the table. The second content word in the table is again a form of this verb: it is *weten* in 23rd position. Apparently there is a strong collocational bond between *niet eens* and this verb.

(26) a. *die man weet niet eens wat een manifest voor een ding is*
      that man knows not PART what a manifesto for a thing is
      'that guy doesn't even know what kind of thing a manifesto is'

    b. *ik weet niet eens meer wat ik gekookt heb.*
      I know not PART PART what I cooked have
      'I can't even remember what I cooked'

    c. *hij wil het niet eens meer weten.*
      he want it not PART PART know
      'he doesn't want to know it anymore'

The last two examples show that *meer* can easily enter into this collocation as well.

A final collocate of *niet eens* worth mentioning is the modal auxiliary *kunnen*. The present singular form *kan* is in 17th position; the form *kunnen*, which can both function as present plural and infinitive, is in 22th. The examples in (25) illustrate this usage, as do the following ones:

(27) a. *daar kan ik niet eens bij.*
      there can I not PART by
      'I can't even reach that'

    b. *ze kon Tim niet eens optillen*
      she could Tim not even lift
      'she could not even lift Tim'

3.4 Collocational behavior of a cluster of modal particles

Having shown that the complex focus particle *niet eens* entertains collocational bonds of various types, both with other particles and with content words such as the verb *weten* 'to know', let us now take a closer look at the properties of a modal particle combination: *maar eens*. The combination is part of Hoogvliet's long cluster in (1a), and takes in the 7th position in the table of collocations with *eens* in (12).[24]

The first thing that strikes us in the collocation table (given in (32) in the appendix) is again the considerable amount of particles among the collocates. Especially *nog maar eens*, *nou maar eens* and *eerst maar eens* appear to be relatively frequent combinations.

The contribution of *nog* appears to be compositional, which is reflected by usage of the English word *again* in the translations. *Nou* ('now') rather seems to express an aspect of impatience, whereas *eerst* 'firstly' suggests that the activity alluded to in the sentence has a certain priority according to the speaker.

(28) a. *ik herhaal dat nog maar eens een keer*
      I repeat that PART PART PART PART
      'I repeat that once again'

    b. *kom nog maar eens terug.*
      come PART PART PART back
      'please come again some time'

---

[24]As an aside: *maar eens* ranks only 41st in the list of collocations with *maar*, *alleen maar* (cf. above) is in 13th position. In text book examples of collocations such as Dutch *fond of* and English *dol op*, the content word is usually considered to be the head word and the preposition the collocate. In the case of particle clusters, it is much harder to decide which element is the head word. However, the numbers just given show that the choice of the head word may have consequences for the success of the method.

    c. *ga nou maar eens een keer proberen*
       go PART PART PART PART   try
       'you just might want to try some time'

    d. *ga nou maar eens eventjes zitten*
       gou PART PART PART PART   sit
       'please sit down a moment'

    e. *laten we eerst maar eens uitzoeken wie 'm die dag gehad heeft.*
       let us PART PART PART find-out who him that day had   has
       'let's first try to find out who had it that day'

    f. *als ik jou was zou ik eerst maar eens met dit boekje beginnen.*
       if I you were would I PART PART PART with this booklet start
       'If I were you I'd start with this booklet first'

The next collocate of *maar eens* in table (32) is *moet* 'must'. Other forms of this modal auxiliary occur in the table as well. All in all, 26 (out of 79) sentences with *maar eens* contain a form of *moeten* in the same clause.

(29)   a. *u moet maar eens kijken in de wet*
          you must PART PART look in the law
          'you should take a look into the law'

         b. *daar moeten we maar eens over gaan praten.*
            there must we PART PART over go   talk.
            'we should talk about that some time'

         c. *dat moest maar eens duidelijk zijn.*
            that must (PAST) PART PART clear   be
            'that should be clear'

Now let us return to the examples in (28), and compare them to the ones in (29). It turns out that all sentences (the only exception being (28a)) belong tot the same speech act type: they are all directive sentences, i.e., they all can be seen as "verbal interactions whose object it is to get someone to do something for you" (Vismans 1994). Closer inspection of the data in our corpus demonstrates that this is a general tendency: apart from the 26 sentences containing an instance of the strong modal *moeten*, one finds 24 sentences with a morphological imperative (e.g. 28b, 28c, 28d), several with an adhortative construction involving the auxiliary *laten* (as in 28e), etc. All in all, 59 of the 79, or three quarters of all examples, are directive in one way or another. Moreover, all directives with *maar eens* are weak. The relatively strong directive power of modal *moeten* and of the morphological imperative, which are quite impolite or even rude when used all by themselves, is weakened (or mitigated, to use Vismans' terminology) by *maar eens*. On the other hand, *maar eens* may turn a statement such as (30a) (a main clause variant of (28f)) into a (weak) directive in (30b).

(30)   a. *ik zou met dit boekje beginnen.*
          I would with this booklet start
          'I'd start with this booklet'

         b. *ik zou maar eens met dit boekje beginnen.*
            I would PART PART with this booklet start
            'you'd better start with this booklet'

Note that this is an important difference with the collocational relations between *niet eens* and the modal auxiliary *kunnen* 'can' discussed above: there we find a relationship between a (complex) particle and a certain verb, whereas here we appear to be dealing with a particle combination's preference for a certain speech act type, which can be expressed in various ways, including a modal auxiliary (*moeten* 'must'). In the former case there is no indication that *niet eens* has a general preference for weak modals or for the kind of speech act or semantics *kunnen* may express or contribute.

Perhaps one might want to speculate on the basis of the data discussed in this section that *maar eens* is developing into a specialized marker for (weak) directives.[25] This development, however, if it exists in the first place, has not come to completion, as the existence of examples such as (28a) show in which *maar eens* occurs in an assertion rather than a directive.

4 Summary and conclusion

The main purpose of this paper has been methodological. We have tried to answer the question whether it is possible, in principle, to use standard collocation search techniques to find interesting distributional behavior of particles. The answer has proven to be positive: in a relatively small corpus of Dutch, it turned out to be quite easy to automatically trace various strong collocational bonds between particles that were known from the literature. Moreover, using the same tools and techniques, we also found many unexpected collocational preferences of particles and particle combinations, both towards other particles and members of other word classes. Standard collocation search techniques were thus shown to be able to enrich the particle researcher's tool box, and thus to help further our knowledge with respect to particle behavior.

References

CALLEBAUT, INGE, TON VAN DER WOUDEN, PIET VAN DE CRAEN, & FRANS ZWARTS. 1998. Er was eens: een partikel. lecture TIN-dag, Utrecht, 17 januari 1998.

DE JONG, EVELINE D. (ed.). 1979. *Spreektaal. Woordfrequenties in gesproken Nederlands.* Utrecht: Bohn, Scheltema & Holkema.

DE VRIES, MATTHIAS, & LAMMERT A. TE WINKEL ET AL. (eds.). 1864–1998. *Woordenboek der Nederlandsche taal.* 's-Gravenhage [etc.]: Martinus Nijhoff [etc.].

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.

FRASER, BRUCE. 1999. What are discourse markers? *Journal of Pragmatics* 31, 931–952.

GEERTS, GUIDO, & TON DEN BOON (eds.). 1999. *Van Dale Groot woordenboek der Nederlandse taal.* Utrecht/Antwerpen: Van Dale Lexicografie. 13e, herz. uitg.

---

[25]Data concerning *maar eens* from another source may be constructed as additional evidence for this hypothesis (van der Wouden 2001b). The 1996 novel *De nakomer* by Maarten 't Hart contains 14 instances of *maar eens*, almost always (12 times) in a directive sentence, usually a morphological imperative (9 times) or a form of the strong modal auxiliary *moeten* (3 times). In 10 out of 14 of the cases the combination *maar eens* is part of a larger cluster of particles and particle-like elements. Suppose now that it is true, at least for this author, that *maar eens* has developed into a pure grammatical marker of weak directivity. Then the combination is no longer able to fulfill other functions typical for modal particles, such as mitigation of speech acts, suggesting speaker's opinions, etc. If the speaker still wants to do such things, (s)he has to turn to other means – e.g., other modal particles. Now given that there is only one canonical position in which these elements may occur, the result will be a long(er) cluster, and longer clusters we find.

HASPELMATH, MARTIN. 1997. *Indefinite Pronouns*. Oxford: Clarendon Press.

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).

HOEKSTRA, HELEEN, MICHAEL MOORTGAT, BRAM RENMANS, INEKE SCHUURMAN, & TON VAN DER WOUDEN. 2001a. On certain syntactic properties of spoken Dutch. Paper delivered at Computational Linguistics in the Netherlands, Enschede, November 2001.

——, MICHAEL MOORTGAT, INEKE SCHUURMAN, & TON VAN DER WOUDEN. 2001b. Syntactic Annotation for the Spoken Dutch Corpus Project (CGN). In *Computational Linguistics in the Netherlands 2000*, ed. by W. Daelemans, K. Sima'an, J. Veenstra, & J. Zavrel, 73–87. Amsterdam/New York: Rodopi.

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.

JUSTESON, JOHN S., & SLAVA M. KATZ. 1995. Technical terminology: some linguistic properties and an algorithm for indentification in text. *Natural Language Engineering* 1, 9–27.

KÖNIG, EKKEHARD. 1991. *The meaning of focus particles: a comparative perspective*. London [etc.]: Routledge.

KRENN, BRIGITTE, & STEFAN EVERT. 2001. Can we do better than frequency? A case study on extracting PP-verb collocations. 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 & Geoffrey Williams, 39–46. Toulouse, France: CNRS – Institut de Recherche en Informatique de Toulouse, and Université de Sciences Sociales.

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.

MATTHEWS, P.H. 1981. *Syntax*. Cambridge [etc]: Cambridge University Press.

OOSTDIJK, NELLEKE. 2000. Building a corpus of spoken Dutch. In *Computational Linguistics in the Netherlands 1999. Selected Papers from the Tenth CLIN Meeting*, ed. by Paola Monachesi, 147–157. Utrecht: Utrecht University, Utrecht Institute of Linguistics OTS.

PALMER, F.R. (ed.). 1968. *Selected papers of J.R. Firth, 1952–1959*. Bloomington [etc.]: Indiana University Press.

RENKEMA, JAN. 1981. *De taal van 'Den Haag': een kwantitatief-stilistisch onderzoek naar aanleiding van oordelen over taalgebruik*. 's-Gravenhage: Staatsuitgeverij.

——. 1989. *Schrijfwijzer*. 's-Gravenhage: SDU uitgeverij. Volledig herz. ed. (Oorspr. uitg.: 1979).

ROSS, IAN C., & JOHN W. TUKEY. 1975. Introduction to these volumes. In *Index to Statistics and Probability*, ed. by John W. Tukey, iv–x. Los Altos: R & D Press.

SCHIFFRIN, DEBORAH. 1987. *Discourse markers*. Cambridge: Cambridge University Press.

SINCLAIR, JOHN. 1991. *Corpus, concordance, collocation*. Oxford [etc.]: Oxford University Press.

SMADJA, FRANK A., & KATHLEEN R. MCKEOWN. 1990. Automatically extracting and representing collocations for language generation. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, 252–259.

THURMAIR, MARIA. 1991. Kombinieren Sie doch nur ruhig auch mal Modalpartikeln! Combinatorial regularities for modal particles and their use as an instrument of analysis. *Multilingua* 10, 19–42.

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

van der Auwera, Johan. 1992. Dutch 'once' and related issues. In *Adverbial quantification and other adverbialia*, ed. by J.C. Moreno Cabrera, 49–55. Strasbourg: European Science Foundation. Eurotyp Working Papers 4.

van der Wouden, Ton. 1997. *Negative Contexts. Collocation, polarity, and multiple negation*. London and New York: Routledge.

——. 2000a. Collocational behaviour in the modal realm. Paper presented in the workshop on Collocation, DGFS-meeting, March 2000, Marburg.

——. 2000b. Collocationeel gedrag bij partikels. Lezing TABU-dag, Groningen, 16 juni 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 & Geoffrey 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* 1/2002, 20–43.

——. 2001c. Three modal verbs. In *Zur Verbmorphologie germanischer Sprachen*, ed. by Sheila Watts, Jonathan West, & Hans-Joachim Solms, 189–210. Tübingen: Niemeyer.

Vismans, Roel. 1994. *Modal particles in Dutch directives: a study in functional grammar*. Vrije Universiteit Amsterdam dissertation.

Zwarts, Frans, & Ton van der Wouden. 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.

Appendix

(31)

	Top collocates of *niet eens* ($N = 425$) according to WordSmith														
n	word	total	left	right	l5	l4	l3	l2	l1	*	r1	r2	r3	r4	r5
1	niet eens	427	1	1	1	0	0	0	0	425	0	0	0	0	1
2	dat 'that'	147	78	69	5	20	32	15	6	0	7	16	19	8	19
3	nog 'PART'	83	81	2	2	2	1	1	75	0	0	0	1	1	0
4	die 'that/those'	66	41	25	8	6	10	13	4	0	0	6	3	8	8
5	een 'a/one'	63	17	46	11	5	0	1	0	0	24	4	3	5	10
6	maar 'but/PART'	59	22	37	8	7	7	0	0	0	1	9	11	8	8
7	het 'the/it'	58	37	21	7	2	11	5	12	0	2	2	3	8	6
8	meer 'anymore'	57	3	54	2	0	0	0	1	0	45	1	3	4	1
9	dan 'then/than'	52	31	21	6	10	8	5	2	0	0	3	5	7	6
10	weet 'know (SG)'	47	41	6	0	2	4	18	17	0	1	2	0	0	3
11	nee 'no'	38	14	24	1	5	6	0	2	0	0	7	8	5	4
12	hij 'he'	36	23	13	0	5	4	9	5	0	0	1	4	7	1
13	van 'of'	35	13	22	6	2	4	1	0	0	3	5	4	4	6
14	ook 'PART'	32	25	7	3	3	3	4	12	0	0	0	3	2	2
15	ggg	31	7	24	3	0	3	0	1	0	0	9	5	6	4
16	nou 'PART'	31	19	12	10	3	3	3	0	0	0	2	2	4	4
17	kan 'can (SG)'	30	21	9	4	0	3	8	6	0	1	0	1	3	4
18	was 'was'	30	15	15	1	3	3	5	3	0	1	1	3	5	5
19	heb 'have'	26	16	10	1	2	6	3	4	0	1	0	3	4	2
20	zijn 'are'	26	10	16	2	3	1	3	1	0	3	3	4	4	2
21	daar 'there'	24	18	6	6	4	4	2	2	0	0	4	2	0	0
22	kunnen 'can (PL/INF)'	22	10	12	2	2	2	4	0	0	4	3	1	2	2
23	weten 'know (PL/INF)'	22	10	12	0	1	1	0	8	0	6	3	2	0	1

(32)

	Top collocates of *maar eens* ($N = 79$) according to WordSmith														
n	word	total	left	right	l5	l4	l3	l2	l1	*	r1	r2	r3	r4	r5
1	maar eens	79	0	0	0	0	0	0	0	79	0	0	0	0	0
2	een 'a'	20	4	16	4	0	0	0	0	0	11	3	2	0	0
3	dat 'that'	18	7	11	2	2	1	2	0	0	0	1	2	5	3
4	nog 'PART'	13	10	3	0	0	1	0	9	0	0	0	1	0	2
5	nou 'PART'	13	12	1	1	1	3	2	5	0	0	0	0	0	1
6	moet 'must'	12	11	1	0	2	3	4	2	0	0	0	1	0	0
7	wat 'what'	10	2	8	1	1	0	0	0	0	3	2	1	1	1
8	dan 'PART'	9	7	2	1	1	1	1	3	0	0	1	0	1	0
9	eerst 'PART'	9	9	0	0	0	0	0	9	0	0	0	0	0	0
10	die 'those'	7	2	5	0	0	1	1	0	0	1	0	0	3	1
11	het 'the/it'	7	4	3	1	0	0	1	2	0	0	0	1	2	0
12	van 'of'	7	2	5	0	1	1	0	0	0	0	1	2	1	1
13	zou 'would'	7	3	4	0	0	1	2	0	0	2	0	1	1	0
14	jij 'you (SG)'	6	5	1	0	0	1	2	2	0	0	0	0	0	1
15	keer 'time'	6	0	6	0	0	0	0	0	0	0	6	0	0	0
16	kom 'come'	6	4	2	0	1	0	2	1	0	0	1	1	0	0
17	kijken 'look'	5	0	5	0	0	0	0	0	0	3	2	0	0	0