On the phraseology of stop words*

Ton van der Wouden

In this corpus-based study, ample evidence is provided for the position that high frequency non-content words (also known in the literature as stop words) can have collocational properties, or may be part of larger lexical units, just like content words. For researchers in Natural Language Processing, this result may be a reason to rethink their strategies and algorithms for extracting and implementing fixed phrases. Language teachers might want to reconsider the kind of idiomatic expressions they choose to teach their pupils, whereas our findings may be important for linguistic theory as well, as they offer an empirical argument in the ongoing discussion about the division of labor between grammar and the lexicon.

1. Introduction: phraseolexemes

When producing or interpreting language, all language users heavily deploy ready-made linguistic building blocks. These are known in the literature under an impressive variety of names, such as constructions, formulaic units, idioms, phraseologisms and stock phrases (Pawley & Syder 1983, Jackendoff 1997, Wray 2002); in the remainder of this paper, I will refer to this kind of prefabricated language pieces with the term **phraseolexemes** (the term is not new, see e.g. Wotjak 1992, but her use of the term does not necessarily match mine completely). It is only relatively recently that linguistics is beginning to appreciate the importance of phraseolexemes in everyday language usage (cf. references given above): the contribution of these phraseolexemes to actual language use, or the amount of them within the average speaker's language competence, can hardly be overestimated. But as definition issues have not been settled yet, every researcher has his or her own estimate of the amount of phraseolexemes an average language user knows and uses.[1]

As phraseolexemes form a major stumbling block in all serious Natural Language Processing applications, from spelling correction to Automatic Translation, there is a considerable tradition of trying to extract, in an automated way, collocations and other phraseological units from large text corpora. However, definitional issues comparable to the ones hinted at above apply. Under the most general interpretation of the notion phraseolexeme, any two lexical elements occurring more often in each other's proximity than chance predicts, should be considered as having an interesting relationship. From this

[*] In Pepijn Hendriks, Frank Landsbergen, Michaela Poss & Jenneke van der Wal (eds.), *Leiden Papers in Linguistics* 4.1 (2007), 56-67
http://www.lucl.nl ISSN 1574-4728

[1] I will refrain here from speculating on the ramifications for linguistic theory of assuming that part of linguistic knowledge is stored in the form of combinations of lexical items with idiosyncratic properties; cf. for example Goldberg (1995) and Croft (2001).

perspective, (Firthian) textbook combinations such as *collect stamps*, *proud of* and *wait for* qualify as phraseolexemes just like *George W. Bush* and *the United Nations* do. But then, the same holds for *have to* and *an apple*, since these combinations occur more often than chance predicts as well (van der Wouden 1997). Traditionally, therefore, both theoreticians (such as Haussmann 1989-91) and NLP practitioners (e.g. Manning & Schütze 1999) often restrict themselves to combinations consisting of at least one content word, and exclude so-called stop words, i.e. high-frequency function words, from their research and/or their extraction tools. However, as I have argued in van der Wouden (2001) and elsewhere, strategies of excluding stop words from participating in potential phraseolexemes, although highly effective in some cases, are not without their dangers in others. More specifically, I have demonstrated there that certain high-frequency function words may show all kinds of collocational effects. One example: the Dutch particle *eens/'ns* (cf. Eng. *once*) occurs in many fixed, non-compositional combinations, e.g. *niet eens* 'not even', *nog eens* 'once again'. The combination *niet eens* has lexicalized into a negative focus particle meaning 'not even', which in turn shows strong collocational bonds with, among others:

- other particles, such as *nog* 'yet, again, ...';
- the modal verb *kunnen* 'can', especially in the dynamic reading 'be able to';
- the lexical verb *weten* 'know'.

This paper will continue this line of research by investigating the phraseological properties of the most frequent word in current day spoken Dutch – which is, of course, a function word.

2. Spoken language

Considerable differences exist between spoken and written variants of one and the same language. Miller & Weinert (1998), for example, argue that there is nothing in spoken language that corresponds directly to the sentence in written language. Others too have shown that the grammars of spoken and written variants of a language are not identical, although they, of course, overlap to a considerable extent. There are, however, construction types that occur felicitously in spoken language while being (considered) ungrammatical in the written language, and vice versa. A Dutch example of the first type is given in (1) (after van der Wouden et al. 2003, cf. also Jansen 1981):

(1) *Ik ben eigenlijk ben ik docente Frans*
I am actually am I teacher-FEM French
'Actually, I teach French'

Just as in English, a standard Dutch main clause contains one and only one inflected verb. But in this construction, which is far from infrequent in spoken language, we see the verb is doubled, so to speak, together with the subject, probably for discourse reasons (Huesken 2001).

Another relevant difference between spoken and written language may be the usage of phraseolexemes. It has been conjectured (cf. Wray 2002) that, due to time and memory constraints of the speaker, spontaneous spoken language usually employs these ready-made linguistic building blocks at a larger scale than consciously, scrupulously composed written language does. Kuiper (1996) shows that in certain situations, such as at auctions and in sports reports, professional speakers utilize nothing more than a very limited set of (open-slot) phraseolexemes, thus saving the computational power of their minds, so to speak, to fill in the open slots with the contextually dependent information like prices, amounts, players' names and their positions in the field.

So far, most of our knowledge about the usage and types of phraseolexemes in spoken discourse is based on anecdotal or domain-specific evidence, at best. But since larger and better corpora of spoken variants of languages are becoming available, new perspectives are opening to investigate these issues in a more systematic way.

Here, I will report on the phraseological properties of the most frequent word in the Dutch part of the Spoken Dutch Corpus (CGN) (Oostdijk et al. 2002). It will be shown that this word – a function word of course – occurs in fixed combinations far more often than chance predicts. Various of these combinations, be they frequently occurring or not, have developed special properties, either syntactic (such as specialization or grammaticalization), phonological (special stress patterns (van der Wouden 2002), fossilized reduced pronunciations (Strik et al. 2005)), semantic (see Hoeksema and Rullman 2001 for an example) and/or pragmatic (for instance, development into discourse markers).

Many of these special, unpredictable properties have so far remained unnoticed (or at least undocumented) in the descriptive literature (grammars, dictionaries). Moreover, they are also important for language learners as well as for NLP applications. This makes our research potentially relevant for grammarians, lexicographers and language teachers alike.

3. The most frequent word in spoken Dutch

The recently finished Spoken Dutch Corpus (CGN) (Oostdijk et al. 2002) aimed at collecting a representative overview of current spoken Dutch. Currently (2007), it contains approximately 9,000,000 words (900 hours) of speech of adult speakers from various backgrounds and from various regions in the Netherlands and Flanders, in a variety of speech situations.

Although the Spoken Dutch Corpus is not the first corpus of spoken Dutch, it is much larger and much more balanced than earlier attempts of this kind. The corpus opens new horizons to further our knowledge about spoken varieties of the Dutch language.

It is a well-known fact that considerable differences exist between the variants of Dutch spoken in the Netherlands and in Belgium, and as these variants are not of immediate concern for the point made in this paper, the Belgian part of the corpus will be left out of consideration.

Table 1 offers an overview of the most frequent words in the Dutch part of the corpus.

N	word	frequency	%
1	*ja* ‘yes’	192,456	3.22
2	*de* ‘the’	159,545	2.72
3	*dat* ‘that’	156,005	2.66
4	*uh* ‘uh’	154,252	2.63
5	*en* ‘and’	141,063	2.41
6	*ik* ‘ik’	125,162	2.13
7	*een* ‘a, an, one’	111,628	1.90
8	*’t* ‘the, it’	97,895	1.67
9	*je* ‘you, your’	90,155	1.54
10	*die* ‘that’	88,479	1.51
		1,316,640	**22.4**

Table 1: The most frequent words in the Dutch part of CGN
(N = 5,863,159, all counts courtesy WordSmith Tools)

Note that ten word types, all function words (as could be expected), comprise almost 20% of all word tokens in the corpus. Of course, these numbers should be handled with care: table 2 shows that there are considerable differences between the various subcorpora, i.e. between the various text types.

subcorpus	**# *ja***	**%**	**first in #**	**%**
face to face (N = 1,815,735)	1	4.52	*ja*	4.52
telephone (N = 1,286,962)	1	5.96	*ja*	5.96
radio (N = 1,001,366)	13	1.16	*de* 'the'	4.57
read aloud text (N = 558,543)	135	0.09	*de*	4.78
lesson (N = 307,876)	8	2.00	*dat* 'that'	2.91
interview (N = 264,621)	4	2.86	*uh*	3.98
(political) debate (N = 220,094)	35	0.47	*de*	5.02
television (N = 196,865)	17	0.92	*de*	4.02
simulated business (N = 140,349)	3	3.56	*uh*	4.33
presentation (N = 63,492)	97	0.13	*uh*	3.85

Table 2: The most frequent words in subcorpora of CGN

This table shows that *ja* is the most frequent word in the face-to-face and telephone parts of the corpus, but not in the other subcorpora. In the radio subcorpus, for example, *de* 'the' is the most frequent word, and the same holds for read-aloud texts, whereas *ja* in these two corpora is only in 13th or even 135th position respectively. In other words, *ja* may be the most frequent word in CGN, but that is partially due to it being the most frequent word in the largest subcorpora.

Note that the subcorpora in which *ja* is the most frequent word both involve dialogues. This suggests that *ja* might primarily be a dialogue word. We do, however, find *ja* in the middle of monologues. Consider the following extract from a television reporter:

(2) *maar eigenlijk valt in 't hele scenario de verdediging van 't binnenland verdediging tegen aanvallen binnen Amerika uh ja valt eigenlijk buiten de taakstelling van bijvoorbeeld uh 't Pentagon of van van 't Amerikaanse leger.*
but actually falls in the whole scenario the defense of the homeland defense against attacks within America uh yes falls actually outside the task of for example uh the Pentagon or of of the American army
'Actually, internal defense is outside the tasks of the Pentagon or the US Army'

In this passage, *ja* (or perhaps *uh ja*, cf. below) is used to fill a hesitation pause and to make a restart, in order to repair a syntactic construction that has grown too complex (cf. Levelt 1989).

More often, however, single *ja* is indeed a dialogue word. The following conversation fragment (from the same program on September 11, 2001) may serve as an illustration:

(3) A: *bedenk ook wel dat Amerika nooit een oorlog op eigen grondgebied wat dat betreft gevochten heeft uh...*
remember also that America never a war on own soil as far as that is concerned fought has uh
'Remember also that America has never fought a war on home ground'
B: *ja*
yes

A: *dus dat is een enorm verschil met bijvoorbeeld hoe we d'r in Europa over denken en hoe onze eigen defensie is opgebouwd*
so that is an enormous difference with for example how we in Europe think about it and how our own defense is constructed
'That's an enormous difference with European views for example and with the way our defense works'
B: *ja*
yes

At the end of the first part of his contribution, the reporter seems to be stuck. He hesitates, but the *ja* of the anchor man invites him to continue his contribution. After the second part, the reporter pauses, signaling that the anchor could take a conversation turn. By uttering *ja*, he 'accepts' the reporter's contribution.

After having discussed some of the usage possibilities of single *ja*, I will now turn to combinations with *ja*. We will see that there are indeed various phraseolexemes involving *ja* that have important functions in organizing dialogue structure. For the rest of this paper, I will restrict my attention to *ja*'s phraseological behavior in the dialogue and telephone subcorpora.

4. ja*'s collocational behavior*

In this section, I will investigate the collocational behavior of *ja*. Can we, using the Spoken Dutch Corpus and current lexicostatistical techniques, find interesting phraseolexemes involving *ja*? As a first result, table 3 below shows the most frequent clusters with *ja*.

N	word	frequency
1	*ja ja*	155,785
2	*ja ja ja*	73,180
3	*ja ja ja ja*	30,086
4	*oh ja*	17,683
5	*ggg ja*	16,323
6	*ja dat*	14,706
7	*ja maar*	13,836
8	*ja oh*	12,306
9	*uh ja*	11,721
10	*ja ggg*	11,281
11	*ja ja ja ja ja*	10,322
12	*ja nee*	9,684
13	*ja ik*	9,367
14	*ja nou*	9,077

Table 3: Most frequent clusters with *ja* in Dutch
(face-to-face and telephone conversations in CGN)

As reduplication is not a productive process in Dutch, it is rather surprising that *ja ja* is (by far!) the most frequent combination in which we find *ja*. Upon closer inspection, it turns out that various usage possibilities of *ja ja* should be distinguished. I will return to the most frequent ones of these below.

Even more surprisingly, double *ja* is immediately followed by triple and quadruple *ja*; I will return to these as well. Fourth in rank comes a combination with *oh* 'oh',[2] followed by one with *ggg*, which is how CGN represents all kinds of non-linguistic sounds, such as coughing, laughing, etc. Next come combinations with other function words, often polyfunctional, such as *dat* 'that' (demonstrative or relative pronoun, or complementizer), *maar* (conjunction or particle), etc. In the sections to come, I will take a closer look at some of the most frequent and interesting combinations with *ja*.

4.1. ja ja

The most frequent combination, *ja ja*, often appears to function as a unit. An external argument in favor of this position is the observation that the (unofficial) spelling *jaja*, as one word, is far from rare on the Internet:

(4) *jaja, t heeft even geduurd.* (sandraatje.punt.nl)
yeahyeah, it has briefly lasted
'yeah yeah, it has taken some time'

As far as I can see from the examples in the corpus, *ja ja* has various functions. I would propose to distinguish at least the following three:[3]

- confirmation: "pray continue, I won't interrupt you, you can go on"
- lexicalized irony: "no, I don't believe you"
- turn taker: "I agree (or: I pretend I agree), but now I want to make a contribution"

Although there is no one-to-one mapping between usage and prosody, the various usages each tend to have their own intonation pattern. For example, in the confirmation case, *ja ja* is usually pronounced fast and without stress. In the lexicalized irony case, on the other hand, the first *ja* is often stretched and pronounced with a rising intonation, as in the following examples.

(5) A: *en uh m'n bruine laarzen. dus dat was wel leuk*
and uh my brown boots. so that was quite nice
'And my brown boots, so that looked quite good'
B: *ja ja.*
yeah yeah
'I see'
A: *vond 'k wel. Wouter vond 't ook leuk dus.*
found I really. Wouter found it also nice thus
'No, really. Wouter liked it too, you know'

(6) *ja ja. nou d'r zit misschien wel iets van waarheid in maar ik geloof er niet in*
yeah yeah. now there sits maybe well something of truth in but I believe there not in
'yeah yeah. well, there may be some truth in it but I don't believe in it'

(7) A: *wat gek zomaar in the middle of nowhere*
what strange just in the middle of nowhere

[2] The variant *o* is more common in Dutch orthography to represent this interjection, but CGN has chosen *oh*.

[3] A fourth one might be impatience: "please go on and make your point", but I didn't find any clear examples.

‘How strange, just in the middle of nowhere’
B: *ja ja. nou ja konden ze zich goed concentreren misschien*
yeah yeah now yeah could they themselves well concentrate perhaps
‘Well, perhaps they could concentrate better’

Probably the strongest case for a separate lexiacal entry *ja ja* can be made for the lexicalized irony case: this meaning (essentially equivalent in meaning to *no*, cf. Horn 1989:55, note 19) is impossible with single *ja* (unless one employs extreme paralinguistic means such as heavy head shaking).

4.2. ja ja ja (and longer variants)

The next most frequent combination is *ja ja ja.* Even for native speakers, it is often not so clear what the difference is with double *ja ja* or longer combinations.

(8) A: *zijn zij ook verstoken van radio en TV?*
are they also deprived of radio and TV?
‘So they don't have radio or TV either?’
B: *ja ja ja*
yeah yeah yeah
‘Sure’
A: *oh dat wist ik niet.*
oh that knew I not
‘Oh, I didn’t know that’

It is therefore unclear whether a separate phraseolexeme (or lexical item) *ja ja ja* should be distinguished. In some cases, there can be no doubt, as these are clearly compositional. Consider the following example, in which there is a considerable pause after the first *ja*, and the double *ja* is used emphatically:

(9) A: en neemt ze ook nog wol mee naar Spanje om te breien?
and takes she also yet wool along to Spain for to knit?
‘And will she also take knitting wool to Spain?’
B: ja. ja ja
yeah. yeah yeah
‘Yeah. No really, I kid you not’

4.3. oh ja

Whereas *ja* is the most frequent word in the subcorpora, the combination of *oh* with *ja* is found in an unexpectedly high position. As far as I can see, *oh ja* has at least three meanings or usage possibilities. In one usage, it is a sort of counterpart to English *oh I see*:

(10) *oh ja dat zei je ja*
oh yeah that said you yeah
‘Oh yes I see that's what you said’

In another use, *oh ja* has a question intonation; in this case, it is comparable to English *oh really?*:

(11) A: jij hebt hier een keer hele tijden geleden je toilettas laten staan.
you have here a time whole times ago your toilet bag let stand
'Once, a long time ago, you left your toilet bag here'
B: oh ja? welke dan
oh yeah? which then?
'Oh really? which one is it?'

A final way of using *oh ja* to be discussed here is to start a new topic.

(12) *uhm oh ja ik heb wat vergeten die gebakken geitenkaas*
ehhm oh yeah I have something forgotten that baked goat's cheese
'Oh dear, I forgot something: the baked goat cheese'

The question now is, whether we should postulate lexical unit (or perhaps even several lexical units) *oh ja*. The mere frequency of the combination suggests that it is part of native speaker's lexis. On the other hand, it appears that in the examples above, *oh ja* can be interchanged by a shorter form without much change in meaning: both *oh dat zei je ja* and *ja dat zei je* could replace the utterance in (10) (on the other hand, *ja dat zei je ja* sounds very weird to my ears). In (11), the function of *oh ja* might have been fulfilled both by *oh* or *ja* (with question intonation, of course). And in (12), simple *oh* would have been (nearly?) as good as *oh ja* (simple *ja* is impossible in this case). Further research is needed to decide whether and why we would want to call *oh ja* a lexical item (or more than one). It is, however, clear that there is an interesting phraseological relationship between the two 'elements' of this combination.

4.4. ggg ja

As we have seen before, *ggg* is the CGN representation of all non-linguistic man-made sounds, such as laughing, sneezing, coughing, etc. We would hardly expect any phraseological effects between that type of sounds and a word like *ja*. Still, WordSmith's collocation tools tell us that *ggg* is found more often adjacent to *ja* than at a larger distance, and more often immediately to the left of *ja* than immediately to the right of it (approximately 8,000 times vs. approximately 5,000 times). This observation asks for an explanation. I am unable to offer such an explanation here, but I would like to suggest that such an thing might be found in discourse structure: non-linguistic sounds such as coughing and giggling break the normal (turn-taking) structure of the conversation, and *ja* is an appropriate means to take up a topic again. More research is needed to see whether I am on the right track here.

4.5. ja dat

Just like its English cognate *that*, Dutch *dat* functions both as a demonstrative pronoun (*dat boek – that book*) and as a complementizer (*ik denk dat ik kom – I think that I come*) (the third theoretical possibility, *dat* as a relative pronoun, was not found). Thanks to the fact that CGN has been annotated for Parts of Speech, we can quicly check that the complementizer *that* is very rare here: less than 5% of all cases of *dat* after *ja*:

(13) *dus dat iedereen dacht van ja dat je d'r zelf nog niet bent dus dat is eigenlijk niet zo leuk.*
so that everyone thought of yeah that you there self not yet are so that is actually not so nice
'So everyone was like, it's not good that you aren't there yourself yet'

Much more common is *that* as a demonstrative, and then much more often used independently, as a pronoun (exemplified in (14)), than determiner-like (15):

(14) *die zag er heel lelijk uit ja dat heb jij ook gezien*
that saw there very ugly out yes that have you also seen
'That looked very ugly indeed, you have seen that too'

(15) *oh ja dat verhaal heb je gehoord ja ja.*
oh yes that story have you heard yes yes
'Oh you know that story of course'

Still, I hesitate to classify *ja dat* as a phraseolexeme. In my opinion, the fact that the combination occurs very often in the corpus can be explained from independent principles of Dutch syntax (such as V2), of conversation (start your contribution with a discourse particle showing that you are taking the floor), and of information structure (start the first sentence of your turn with an old topic, which is typically done by means of a demonstrative pronoun).

4.6. ja maar

The next potential phraseolexeme to be discussed is *ja maar*. As far as I can see, *ja maar* 'yes but' usually functions as a turn taker, expressing something like 'I admit your last point (or at least I say so), but I want to add/correct/...'. The following example is a clear case:

(16) A: *ja is redelijk rustig.*
yeah is reasonably quiet
yes is rather quiet
B: *ja maar je zit wel aan de weg. meteen.*
yeah but you sit well on the road. immediately.
yes but you are next to the road. immediately next to it.

The reverse combination, *maar ja*, occurring somewhat less frequent, can be used to express resignation or concession:[4]

(17) A: *ja da's ook niet ideaal.*
yeah that's also not ideal
'Yes that isn't ideal either'
B: *maar ja goed 't is wel gratis dan.*
but yeah well it is well free then
'But if anything, it'll be free'

We find *maar ja* in the middle of an utterance as well: in (18), *maar ja* clearly expresses a combination of concession ("I admit, we had planned it") and resignation ("we can't help it").

[4] An editor suggests that *ja maar goed* might be a construction in its own right.

(18) A: *we hadden eigenlijk gezegd van uh net na de kerst uh kerstballen te kopen maar ja die zullen nu al wel weg zijn.*
we had actually said of uh just after the Christmas uh Christmas balls to buy but yes those will no already well away be
'We had planned to buy Christmas decoration just after Christmas but it'll probably be sold out by now'

Should we classify *ja maar* and *maar ja* as phraseolexemes? As regards *ja maar*, there is a tradition to do so: in the lemma *maar*, the Van Dale dictionary (Den Boon and Geeraerts 2005) explicitly mentions *ja maar* "ter inleiding van een tegenwerping" ('to introduce an objection', which may be slightly beside the point). Its mirror image *maar ja*, on the other hand, is not found in the dictionaries I consulted. It might be argued that its meaning is compositional: *maar* 'but' contributing concession, and *ja* resignation. The high frequency of the combination, however, may be taken as an argument that the combination is somehow functioning as a unit in the native speaker's language competence. Although I haven't investigated this, I wouldn't be surprised if native speakers, when asked about the meaning of *maar ja*, would come up with a description rather close to the one I gave above. Further research should decide here.

4.7. ja oh

In contrast to *oh ja* discussed above, there is no evidence that *ja oh* is a phraseolexeme. The WordSmith Tools found a considerable number of instances, but most (approximately 95%!) cases cross the border of two utterances. An example is given below:

(19) A: *dadelijk even noteren hoor want dat onthoud ik niet.*
immediately briefly write-down hear because that remember I not
'Write that down immediately, since I won't remember'
B: *oh.*
C: *ja.*

4.8. uh ja

The hesitation marker *uh* (or *eh*, as it is rendered in the Van Dale dictionary) is the second most frequent "word" in the dialogue and telephone subcorpora. Therefore, it is not too surprising, statistically speaking, that a combination with *ja*, the number one in the ranking, scores very high as well. Chance alone, however, cannot explain the very high frequency of *uh ja* all by itself, for the simple reason that chance predicts that *ja uh* should occur just as often, which it doesn't (*ja uh* 2,446 (rank 55), *uh ja* 11,721 (rank 8)):

(20) A: *ja ik heb zoiets van uh ja ik ga niet uh*
yes I have something of uh yes I go not uh
'Yes, I am like uh, yes I'm not going to uh'

My hunch would be that *uh ja* is a lexicalized discourse marker, signaling that the speaker wants to keep the floor.

5. Concluding remarks

In this paper, I have argued that, at least in spoken language, function words may have interesting collocational properties, and/or may be part of larger lexical units. All examples were taken from Dutch, but is not too difficult to come up with corroborating evidence from other languages. For example, a quick Google search revealed the following high frequency combinations with English *yes:*

combination	**frequency**
oh yes	8.7 million
yes yes	5.2 million
but yes	2.1 million
yes yes yes	1.9 million

Table 4: Combinations with *yes* (Google, February 16, 2006)

In this paper, I hope to have shown once again that non content words may show collocational effects that are at least as interesting as those of content words. So far, the collocational properties of high frequency function words (‘stop words’) have not received the attention they deserve: in dictionaries and grammars, the kind of combinations discussed above are covered only cursorily, at best. Their frequency alone, however, should be a strong argument for language teachers to systematically include them in their teaching materials. The same holds for computational linguistics: NLP systems that are designed to deal with spontaneous spoken language should be prepared for these function words and the larger units they are part of. Of course, I have only scratched the surface of what may turn out to be an enormous haystack of spoken language phenomena, filled with loads of phraseolexemic needles.

Acknowledgments

An earlier version of this paper was presented at Phraseology 2005: The many faces of Phraseology. Université Catholique de Louvain (Belgium), October 13-15, 2005. Thanks are due to the audience, reviewers and editors. Remaining errors are mine.

Ton van der Wouden
Leiden University
Leiden University Centre for Linguistics / Dutch Department
t.van.der.wouden@let.leidenuniv.nl

References

den Boon, T. & Geeraerts, D. (2005), *Van Dale Groot woordenboek der Nederlandse taal*, 14th ed., Utrecht/Antwerpen, Van Dale Lexicografie.

Croft, W. (2001), Radical Construction Grammar: Syntactic Theory in Typological Perspective, Oxford [etc.], Oxford University Press.

Goldberg, A. (1995), *Constructions. A Construction Grammar Approach to Argument Structure*, Chicago, University of Chicago Press.

Haussmann, F. (1989-91), Collocations, in Haussmann, F. (ed.), *Wörterbücher: ein internationales Handbuch zur Lexikographie* [...], Berlin [etc.], De Gruyter, I: pp. 1010-19.

Hoeksema, J. & Rullmann, H. (2001), Scalarity and Polarity: a Study of Scalar Adverbs as Polarity Items, in J. Hoeksema et al. (ed.), *Perspectives on Negation*, Amsterdam, John Benjamins, pp. 129-171.

Horn, L. (1989), *A natural History of Negation*, Chicago, University of Chicago Press.

Huesken, N. (2001), *Mirrorsentences. Repetition of inflected verb and subject in Spoken Dutch*. Master's thesis General Linguistics, Utrecht University.

Jackendoff, R. (1997), *The Architecture of the Language Faculty*, Cambridge, Mass., MIT Press.

Jansen, F. (1981), *Syntaktische konstrukties in gesproken taal*, PhD thesis Leiden University.

Kuiper K. (1996), *Smooth Talkers: The Linguistic Performance of Auctioneers and Sportscasters*, Mahwah, NJ, Lawrence Erlbaum Associates.

Levelt, W. (1989), *Speaking: From Intention to Articulation*, Cambridge, Mass., MIT Press.

Manning, C. & Schütze, H. (1999), *Foundations of Statistical Natural Language Processing*, Cambridge, Mass., MIT Press.

Miller, J. & Weinert, R. (1998), *Spontaneous Spoken Speech. Syntax and Discourse*, Oxford, Clarendon.

Oostdijk, N. et al. (2002), Experiences from the Spoken Dutch Corpus Project, *Proc. LREC 2002*, pp. 340-347.

Pawley, A. & Syder, F.H. (1983), Two puzzles for linguistic theory: nativelike selection and nativelike fluency, in J.C. Richards and R.W. Schmidt (eds.), *Language and Communication*, London, Longman, pp. 191-226.

Strik, H., Binnenpoorte, D. & Cucchiarini, C. (2005), Multiword Expressions in Spontaneous Speech: Do we really speak like that? in *Proc. of InterSpeech* 2005 (IS2005), Lisbon 4-8 Sept. 2005, pp. 1161-1164.

Wotjak, W. (1992), *Verbale Phraseolexeme in System und Text*, Tübingen, Niemeyer.

van der Wouden, T. (1997), *Negative Contexts. Collocation, Polarity, and Multiple Negation*. London, Routledge.

van der Wouden, T. (2001), Collocational Behaviour in Non Content Words, in B. Daille & G. Williams (eds.), *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*, pp. 16-23.

van der Wouden, T. (2002), Particle Research Meets Corpus Linguistics: On the Collocational Behavior of Particles, in *Belgian Journal of Linguistics* 16, pp. 151-174.

van der Wouden, T. et al. (2003), Harvesting Dutch Trees: Syntactic properties of spoken Dutch, in T. Gaustad (ed.), *Computational Linguistics in the Netherlands 2002. Selected Papers from the Thirteenth CLIN Meeting*, Amsterdam, Rodopi, pp. 129-141.

Wray, A. (2002), *Formulaic Language and the Lexicon*, Cambridge, Cambridge University Press.