On the phraseology of stop words

Ton van der Wouden
ULCL/Dutch, Leiden University
PO Box 9515, 2300 RA Leiden, The Netherlands
t.van.der.wouden@let.leidenuniv.nl

KEYWORDS: stop words, grammaticalisation, phraseology, discovery procedures

Spoken language usually precedes language represented in writing. Children know how to speak and listen years before they learn to read and write. The history of language is estimated to be in the order of magnitude of hundreds of thousands of years, the history of writing in thousands of years. There are many language communities without writing, but only in the case of dead languages such as Latin the number of readers may outnumber the number of speakers.

The fact that writing originates as a representation of spoken language does not imply that the only difference between written and spoken language is the modality (Bolinger and Sears 1981). For example, intonation in spoken language is replaced by punctuation in written language, but only to a certain extent, as there is no direct mapping between the two. Miller and Weinert (1998) even go so far as to claim that spoken and written languages are (partially) different systems in almost every aspect, that is to say, of morphology, syntax, vocabulary, and the organization of texts.

For various reasons, such as time pressure and habituation (Wray, 2002), spontaneous spoken language heavily deploys readymade linguistic building blocks (that are also referred to with other names, such as formulae (Wray 2002), extended lexical units (Stubbs 2002), idioms (van der Linden 1992), phrasal lexical items (Kuiper 2004) etc.), usually at a larger scale than consciously, scrupulously composed written language does. Kuiper (1996) observes that in certain high pressure situations, most of a speaker's utterances consist of stored (or, to use another metaphor, pre-compiled) linguistic material, with very little syntactic computation going on. We may assume that this is one extreme of a continuum, the other extreme being the Chomskyan ideal of a creative language user with an infallible memory and infinite processing power, with enough time to verbalize new ideas in an original way.

Estimates as to the number of such larger lexical items differ widely, depending, among other things, on the definition used. Altenberg (1998) reports that his corpus of nearly half a million running words of spoken English contains over 200000 recurrent word-combinations. "A rough estimation indicates that over 80 per cent of the words in the corpus form part of a recurrent word-combination in one way or another." Jackendoff (1997) claims that "[t]here are too many idioms and other fixed expressions for us to simply disregard them as phenomena 'on the margin of language'" whereas Kuiper (2004) quotes Mel'čuk suggesting "that the phrasal lexicon is an order of magnitude larger than the one-word lexicon".

Linguistics is only beginning to appreciate the importance of prefabricated language pieces in everyday language usage, as traditional grammars and dictionaries traditionally deal for the greater part with written variants of the language (Miller and Weinert 1998). However, as larger and better corpora of spoken variants of languages are becoming available, new perspectives are opening to investigate them in a systematic way.

There is a considerable tradition of trying to extract, in an automated way, collocations and other phraseological units from text corpora. Under the most general interpretation of the notion collocation, any two lexical elements occurring more often in each other's neighborhood than chance predicts should be considered having a collocational relationship. From this perspective, text book combinations such as *collect stamps* and *proud of* qualify as collocations, but the same holds for *have to* and *an apple* (Van der Wouden 1997).

Cases such as the latter two show the weakness of a purely quantitative approach to collocation. Still, for lack of a better one that can be operationalized, quantitative definitions of collocation are often used, especially in the automatic retrieval of such fixed combinations.

Two popular strategies employed widely (Manning and Schütze 1999) in order to reduce the number of uninteresting combinations such as *an apple* and *have to* (uninteresting in the sense that they are either transparent or can be explained better by grammar) are "part of speech filters" (e.g. Ross and Tukey (1975), Justeson and Katz (1995)) and "stop word lists" (e.g. Smadja and McKeown (1990)).

- part of speech filters: only let through those syntactic structures that are likely to be 'phrases'.
- stop word lists: neglect certain words (usually high frequency function words) as parts of higher than chance bigrams and N-grams.

However, it has been argued (van der Wouden 2001) that these strategies of excluding certain high frequency elements from the set of potential collocations are, although highly effective in some cases, not without its dangers in others. It was demonstrated there that, at least in Dutch, certain high frequency function words show all kinds of collocational effects. Part of speech filters and stop word lists in the usual sense effectively filter out these effects.

In line with this tradition, we will concentrate in our paper on the collocational properties of the most frequent words in the Dutch part of the Spoken Dutch Corpus (CGN) (Oostdijk *et al.* 2002). It will be shown that these high frequency words, which are all function words, occur in fixed combinations far more often than chance predicts. Many combinations of high frequency words, be they frequently occurring or not, have developed special properties, either syntactic (specialization or grammaticalisation), phonological (special stress patterns (van der Wouden (2002), fossilized reduced pronunciations (Binnenpoorte et al. 2004)), semantic (cf. e.g. Hoeksema and Rullmann (2001)) or pragmatic (e.g. development into discourse markers). We briefly point at a few prominent examples of the latter here, restricting ourself to combinations involving *ja* 'yes', which is the most frequent word, occurring more than 190000 times in the ca. 5.7 M word Dutch subcorpus, or over 3%:

- Although reduplication is not a productive process in Dutch, there exists a combination *ja ja* which is very frequent (over 70000 occurrences). Although there exist other usage possibilities, the combination is often used as a discourse marker or back channel, expressing something like 'I see, pray continue'. In this case, it has a lexicalized intonation: a rising contour on the first ja, and a falling one on the second. (Incidentally, *ja ja ja, ja ja ja ja, ja ja ja ja ja* and even longer chains of *ja*'s occur as well.)

- There is also a frequent combination oh ja, built out of exclamatory *oh* (in isolation used to express a wide area of emotions such as delight, amasement, surprise, pain, fear, anger and impatience) and *ja*. The combination, occurring some 16000 time in the corpus, gives the speaker some time to prepare the continuation of his speech turn and, by filling the pause, it makes it more difficult for the conversation partner to take over.
- The next combination worth mentioning is *uh ja*, consisting of the hesitation marker *uh* (fourth most frequent word, ca. 150000 occurrences) and *ja* ‘yes’. Uh ja appears to be a lexicalized discourse marker, signifying that the speaker wants to keep the floor.
- Finally, there is the combination *ja maar*, which consists of *ja* ‘yes’ and *maar* ‘but’ (the thirteenth most frequent word in the corpus, ca. 80000 occurrences). Occurring more than 13000 times, this combination too can function as a discourse marker, in at least three different situations, in wich it can be paraphrased as 'oh I see', 'really?', and 'to start a new topic', respectively.

Many of the combinations exemplified above and their special, unpredictable properties have remained unnoticed (or at least undocumented) so far in the descriptive literature (grammars, dictionaries). As they are also of the highest important for language learners, our research is of potential relevance for grammarians, lexicographers and language teachers alike.

References

Altenberg, B. (1998) On the phraseology of spoken English: The evidence of recurrent word-combinations. In *Phraseology. Theory, Analysis, and Applications*, ed. by A.P. Cowie, 101–22. Oxford: Clarendon Press.

Binnenpoorte, D., C. Cucchiarini, L. Boves & H. Strik (2004) Multiword expressions in spoken language: An exploratory study on pronunciation variation. Paper presented at Computational Linguistics in the Netherlands 15, Leiden, December 17th, 2004.

Bolinger, D. & D. Sears (1981) *Aspects of Language*. New York: Harcourt Brace Jovanovitch, 3rd edition. (1st edition 1968).

Ernestus, M. (2000) *Voice assimilation and segment reduction in casual Dutch: a corpus-based study of the phonology-phonetics interface*. Vrije Universiteit, Amsterdam dissertation.

Hoeksema, J. & H. Rullmann (2001) Scalarity and polarity: a study of scalar adverbs as polarity items. In *Perspectives on Negation*, ed. by Jack Hoeksema, Hotze Rullmann, Víctor Sánchez Valencia, Ton van der Wouden, 129–171. Amsterdam: John Benjamins.

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

Justeson, J. & S. Katz (1995) Technical terminology: some linguistic properties and an algorithm for indentification in text. *Natural Language Engineering* 1, 9–27.

Kuiper, K. (1996) *Smooth talkers: the linguistic performance of auctioneers and sportscasters.* Mahwah, NJ: Lawrence Erlbaum Associates.

Kuiper, K. (2004) (review of) A. Wray: Formulaic language and the lexicon. *Language* 80, 868–872.

Manning, C. & H. Schütze (1999) *Foundations of statistical natural language processing*. Cambridge, Mass.: The 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. *Proceedings LREC* 2002.

Ross, I. & J. Tukey (1975) Introduction to these volumes. In *Index to Statistics and Probability*, ed. by John W. Tukey, iv–x. Los Altos: R & D Press.

Smadja, F. & K. 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.

Stubbs, M. (2002) *Words and phrases: corpus studies of lexical semantics*. Oxford (etc.): Blackwell.

van der Linden, E.-J. (1992) Incremental processing and the hierarchical lexicon. *Computational Linguistics* 18, 2, 219–238.

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

van der Wouden, T. (2001) 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.

van der Wouden, T. (2002) Partikels: naar een partikelwoordenboek voor het Nederlands. *Nederlandse Taalkunde* 7, 20–43.

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