Collecting Collocations

Dorthe Duncker
University of Copenhagen, Institute of Nordic Philology
Njalsgade 80
DK-2300 Copenhagen S
Denmark
duncker@hum.ku.dk

Abstract

This paper presents a methodology for retrieving collocations from a digital text corpus for practical lexicographical purposes. The methodology has been tested on the Danish PAROLE Corpus. It is argued that each lexical target item has a unique collocational profile, and that precision is enhanced markedly when the search for collocations is tailored to fit the individual profiles, even under sparse data conditions. The disclosure of the profiles is based on a notion of positional weight relative to the target item, i.e. some contextual positions carry a heavier collocational load than others. The actual weights are estimated by comparing the original corpus with a twin corpus with random word order. As word order is decisive for the bulk of collocations, this comparison reveals salient collocational positions for each target item of the original corpus. The retrieved collocations are marked up according to the SGML standard, thus facilitating easy overview and dynamic presentation.

Introduction

Retrieving collocations from a digital text corpus is an information retrieval task balancing between precision and recall. The ideal claim for the lexicographer is that both precision and recall are high. If recall is high, all valid collocations are retrieved, and if precision is high only valid collocations are retrieved. Recall is not really at stake here, though. For the past thirty years it has been feasible to prepare KWIC concordances, and a concordance will always list all instances of the keyword in context which implies a recall of 100%. The tricky problem is how to obtain high precision values. What the lexicographer needs is some augmentation of the concordance, facilitating easy and direct access to groups of valid collocations.

This task can be viewed as one of classification and presentation. First the concordance must be severed into two sections: one containing valid collocations and one containing combinations that are either accidental or hapax legomena. Secondly, the section containing valid collocations must be grouped according to sub-type. The dubious combinations should not be discarded, but arranged in a separate section for manual scrutiny.

In the following a set of heuristics for identifying and retrieving collocations for practical lexicographical purposes will be presented. The methodology has been tested on the Danish PAROLE Corpus, a corpus of 250,000 text words from ten different written genres [Keson 1999].

Identifying Collocations

The field of collocations is in a terminological muddle. Wray & Perkins [2000] list well over 40 terms, and argue that this inconsistency obscures genuinely deep-seated ontological differences. In corpus linguistics, a collocation consists of a target item and its collocates. Significant collocates are lexical items occurring 'close to' the target item more often than chance would predict, and the distinction between casual and significant collocation is made according to the frequency of the collocates in several occurrences of an item, e.g. [Sinclair 1966, 1991; Church et al. 1991; Smadja 1991a, 1991b, 1993; Clear 1993; Biber et al. 1998]. The notion of proximity is treated as a matter of context size. Collocates are located within a fixed window of typically four items to the left and to the right of the target item. This definition is considered reductionistic and too narrow by e.g. Howarth [1996] and Wray & Perkins [2000], partly because it fails to identify those collocational relationships that exist over distances greater than the window size, and partly because of the recurrence requirement.

A language is not a set of words without internal groupings or organization [Toolan 1996]. The bulk of produced language is made up largely of established groupings of words, collocational 'sets' or 'frameworks' [Renouf & Sinclair 1991; Wray & Perkins 2000]. According to Altenberg [1991, 2001] as much as 80% of speech is repetitive, estimated from the linear distribution of recurrent contiguous word combinations (n-grams). Most of these combinations are sequences of frequently co-occurring text words (e.g. "and the", "of the", "in the"), but they do not co-occur with a probability greater than chance.

The definition found in Perkins [1999] combines the two approaches and sets up a collocational scale ranging from semantic opacity and absolute fixedness at one end, to semantic transparency and free combination at the other end. He defines formulaicity as: "manifested in strings of linguistic items where the relation of each item to the rest is relatively fixed, and where the substitutability of one item by another of the same category is relatively constrained". Semantic opacity is not always concomitant of fixedness, cf. [Clear 1993] on idioms vs. stereotypes. Semantically opaque collocations are fixed (e.g. *købe katten i sækken* 'buy a pig in a poke'), but some fixed collocations are perfectly transparent (e.g. *gøre en dyd af nødvendigheden* 'make a virtue of necessity'). The continuum between absolute fixedness and free combination is filled with collocations containing slots for, more or less, open class items (e.g. *for ... skyld* 'for ... sake', *for afvekslingens skyld* 'for a change', *for en anden gangs skyld* 'next time', *for hans skyld* 'for his sake'etc.).

Perkins' definition covers the collocational scale, but it misses the point of recurrence. Rather collocation should be defined as *conventional* combination of two or more, contiguous or non-contiguous, linguistic items within a unit relevant to the linguistic practice form (e.g., a sentence, a speech turn, a stanza). A word combination can be recurrent without being conventional, whereas the opposite is not possible [Lewis 1969]. That it is conventional implies that native speakers and readers will expect the co-occurrence of one or more additional items when a specific item is encountered under specific circumstances. It also implies that they will prefer the co-occurrence of exactly these items rather than other items with the 'same' meaning or syntactic function [Duncker 2001], e.g. *tage ansvar for ...* 'take/shoulder the responsibility for ...' is preferred by speakers of Danish to *gribe ansvar for ...*, even though *tage* 'take' and *gribe* 'seize' are perfectly substitutable in other circumstances (e.g., *tage/gribe fat i nakken/kraven på ...* 'take ... by the scruff of the neck').

Heuristics for Collecting Collocations

A number of statistical measures have been developed to compute the association significance between two events. Among these, three measures in particular have become mainstream during the last decade: multiple information, t-score, and z-score, e.g. [Church et al. 1991; Clear 1993; Biber et al. 1998; Howarth 1996]. Multiple information measures the strength of association between two words, while t-score and z-score measures the statistical confidence with which it can be claimed that there is some association. All three measures are useful, but the list of possible collocates are quite noisy and inconclusive. The most serious limitations of these measures are, that they evaluate the correlation between pairs of words, but a collocation may very well involve more than just two words. Secondly they retrieve combinations irrespectively of the position of the collocate relative to the target item. In the work of Smadja [1991a, 1991b, 1993] position is taken into account as well, and this approach is much more viable. The method proposed by Smadja is composed of three stages. In the first stage pairwise relations are identified, and the output of this stage is passed on as input to the next two stages. In the second stage, collocations involving more than two words are retrieved, and in the third stage syntactic information is added in order to filter out further uninteresting combinations. This approach is very effective on high frequency collocates, but it does not seem to be effective on low frequency words [Smadja 1993].

Sparse data conditions pose a serious problem to the retrieval of collocations. The difficulties turn up when collocations involving low frequency words are retrieved from a large corpus or when the corpus is small. Either way the challenge is to identify dependencies among rare cases. (See also e.g. [Dunning 1993; Pedersen 1996; Lin 1998; Johnson to appear] for heuristics on this matter.) The problem is not solved by compiling still larger corpora, it is only staved off. With a large corpus, the chances of finding rare instances increase and frequent instances can be more thoroughly examined. Thus, in principle large corpora should be preferred, but if the corpus markup is missing or defective, the findings could be misleading and reliable findings cannot be kept apart from unreliable ones. When the choice is between a small consistently and fully marked up corpus and a large corpus with less consistent or incomplete markup, the small corpus should always be preferred. Before any searches are performed, the entire marked up corpus must be proofread by humans in order to secure the accuracy of the markup. Proofreading random samples does not suffice, especially not in relation to rare cases [Duncker & Ruus to appear].

The ambition to solve the collocational riddle algorithmically without any human intervention is mistaken, apart from being a case of wishful thinking. Humans are never made redundant, but humans and computers have qualitatively different strengths when it comes to large quantities of text. Computers are good at performing heuristic tasks, while humans are experts in the hermeneutic realm [Duncker & Ruus 2000; Ruus 2002]. It should never be left to any algorithm to decide in cases that require understanding or interpretation, and therefore even hapax legomena should not be discarded when collocational candidates are retrieved. In stead the retrieved instances should be presented in an order easy for the lexicographer to survey.

However, the outlook is not as black as Howarth [1998] seems to think by claiming that "phraseological significance means something more complex and possibly less tangible than what any computer algorithm can reveal". A computer algorithm cannot evaluate the conventionality of a given word combination, but it can compute its recurrence. Recurrence

is not only a matter of which exact items co-occur with the target item, but also where the collocates are located. Different words keep different company, and they keep it in different places. Each target item has a unique collocational profile where some contextual positions carry a heavier collocational load than others. By tailoring the search for collocations according to the individual profiles, precision is enhanced markedly, even in rare cases.
The chance for a recurrent word combination to be in fact conventional increases with the number of occurrences, but still a high frequency could be ascribed to the composition of the corpus. This uncertainty can be solved across texts and text type (i.e. genre or register): If a word combination occurs only in a single text or one text type, it is unlikely to be generally expected in any text, irrespective of text type.

Data Preparation

The Danish PAROLE Corpus is a rather small corpus. It contains only about a quarter of a million text words, and only 228 lemmas occur more than 100 times; thus, it presents an ideal frame for exploring the problems of retrieving collocations under sparse data conditions.
The corpus has been lemmatized and marked up according to PAROLE's Corpus Encoding Standard [Norling-Christensen 1996; Ridings 1996]. The information contained in the marked up corpus is converted into a multi level representation with five levels [Duncker & Ruus 2000; Ruus 2002]: A source level containing the original textual word forms, an orthographical level containing the orthographically neutral word form corresponding to the source word, a lemma level, an inflectional level containing information about inflected forms, and finally a word class level. Information for the inflectional level and the word class level is derived from an element attribute in the markup containing morpho-syntactic description.
The retrieval of collocations exploits the multi level representation. Thanks to the orthographical neutral level, text words can be retrieved irrespectively of their orthographical representations, and collocates of the target item can belong to the target level or any level below that level.

Disclosing Collocational Profiles

A candidate collocate is heuristically defined as an item which co-occur with a target item with a frequency greater than chance in a window of r positions before and after the target item. The size of the window, $r \cong 2$, is tailored to fit the collocational profile of each individual target item. For the PAROLE Corpus, the default size is experimentally set to $r=6$. Highly formulaic genres will require a larger window, e.g. ballads and other oral praxis forms, including speech [Duncker & Ruus 2000].
For configurational languages word order is decisive for the bulk of collocations. If word order is destroyed, collocational relations cease to exist. Beside the original corpus, a twin corpus is prepared containing exactly the same SGML elements, but in a random order; like a pack of cards being shuffled. Two concordances are made, one from each corpus version, and the number of occurrences of each collocate is counted for each position of the window. All frequencies of the original corpus are standardized to a weight value with a mean of 0.0 and a standard deviation of 1.0. This procedure prevents those items with a high overall frequency from having an inordinate influence on the total weight load at each position. The

mean value, :, and standard deviation, Φ, is computed for the joint distribution ($N=r\cong4$). These two values are used to compute the weight value, w, for each occurrence value, f, of the original corpus ($w=(f!:)/\Phi$). Hereby the differences between collocates with high or low overall corpus frequencies are neutralized, while the shape of each distribution is kept intact. In practice, the consequence is that open and closed word class items can be considered together.

	tage V											
	Positions before						Positions after					
	-6	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5	+6
fejl ADJ												
Original corpus	0	0	0	0	0	0	5	1	0	0	0	0
Shuffled corpus	0	0	0	0	0	0	0	0	0	0	0	0
Weight	-0,242	-0,242	-0,242	-0,242	-0,242	-0,242	4,602	0,727	-0,242	-0,242	-0,242	-0,242
af PREP												
Original corpus	1	2	2	1	1	0	2	26	9	7	4	4
Shuffled corpus	5	2	5	7	2	2	5	7	2	4	3	4
Weight	-0,674	-0,479	-0,479	-0,674	-0,674	-0,869	-0,479	4,197	0,885	0,495	-0,089	-0,089

Table 1: Positional weights on the collocation *tage fejl af* ... 'be mistaken about ...'.

Table 1 shows the positional weights on the collocation *tage fejl af* ... 'be mistaken about ...'. The verb *tage* occurs 404 times in the corpus, the adjective *fejl* occurs 8 times, and the preposition *af* occurs 3494 times. It appears that *fejl* occurs only to the right of *tage* at positions +1 and +2, and that *af* occurs with weights above zero shifted one position further to the right. The heaviest loads fall on positions +1 and +2 with weights exceeding 4 standard deviations.

Within the window, *tage* occurs with 1340 other items from the lemma level and below (i.e. lemma, inflectional form, and word class). Of these 958 co-occurs only once with *tage*, and cannot be considered collocational candidates. There are 382 items with co-occurrences ∃ 2, and of these 236 occur more frequently than chance, i.e. compared with the shuffled corpus with random word order. A co-occurrence value of 2 is the lowest possible value for evaluating recurrence, and most other approaches refrain from including co-occurrences below 3, e.g. [Clear 1993], but in order to deal with the sparse data problem, it is necessary for the retrieval method to handle low frequency collocates.

Not all collocates at all positions in the window are equally attached to *tage*. The salience of each position appears from the percentage of collocates with weight values exceeding a given threshold, e.g. above 3 or 4 standard deviations. Each target item has its own distributional profile, and this profile shows where to look for collocations involving the target item. In the case of *tage* the heaviest weight falls on position +1 immediately to the right of *tage*. Different target items have different profiles, and suggest a different tailoring of the retrieval process.

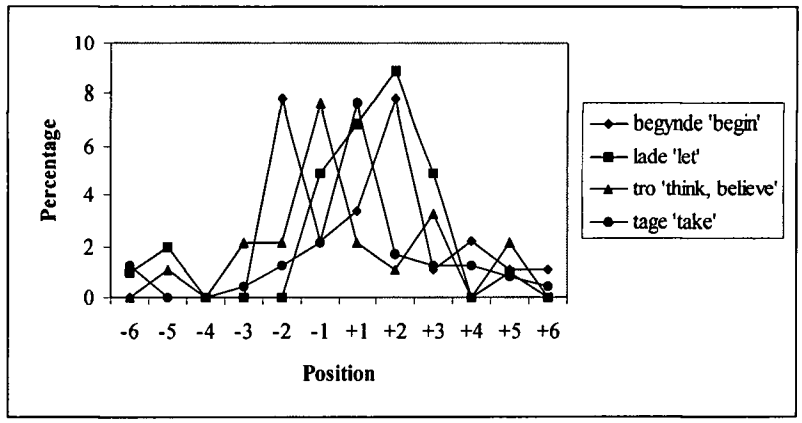

Figure 1: Collocational profiles with weight values exceeding 4 s.d. for four verbs *begynde* 'begin', *lade* 'let', *tro* 'think, believe', and *tage* 'take'.

Figure 1 shows the collocational profiles of weights above 4 s.d. for *tage* and three other verbs, *begynde* 'begin', *lade* 'let', and *tro* 'think, believe'. The profile for *begynde* is bimodal with peaks at positions !2 and +2, *lade* is mostly skewed to the right but includes positions !1 to +3, and *tro* is skewed to the left with a peak at position !1. The profile of a target item is not necessarily constant across registers or through time, and a diachronic study along these lines will be able to disclose changes in collocational practice [Duncker forthcoming].

Tailoring Retrieval for Editorial Purposes

Two parameters from the profile are used to retrieve collocations: the positions designated by the profile where collocates are supposed to be located, and a weight level, e.g. positions from !1 to +3 with weights above 4 s.d. On retrieval, all collocates within the profile on or above the weight level are marked up in a copy of the original context element, and a hyper link to the source location in the corpus is inserted. Further, relevant classificatory information is supplied so that the collocations can be arranged according to type or sub-type (i.e. according to the most significant collocate(s) of each instance). If one or more collocates are found within the profile, the combination is marked as valid, and if the primary collocate belongs to an open word class, the combination is marked as lexical, otherwise it is marked as grammatical (cf. [DS 2394-1:1998]). This information, and information about the syntagmatic pattern (cf. below), is implemented in SGML attributes; thus the concordance, rearranged as a collection of collocations, can be manipulated for editorial purposes by any standard SGML/XML application.

On closer examination is appears that different target items engage in different types of combination with their collocates. This type can be either lexical, syntagmatic, or a mix of both.

The collocational profiles cover a different proportion of the occurrences of the individual target items. In the case of *begynde* 'begin', the profile covers 69.63% of the occurrences, 98.03% of the occurrences of *lade* 'let' is covered by the profile, 72.12% of the occurrences

of *tro* 'think, believe', and half of the occurrences of *tage* 'take'. More than a third (39.60%) of the occurrences of *begynde* fit into the syntagmatic pattern (optional items in brackets):

(subordinate conjunction) ... **begynde** ... infinitive, e.g.

Det er slemt nok, at de indfødte taler engelsk med swahili accent, men her på hotellet er de sorte tjenere **begyndt** at **tale**[*inf*] engelsk med norsk swahili accent, og det er barske løjer.
Hvad lavede du **før**[*sub.conj*] du **begyndte** at **arbejde**[*inf*] for en kontrakt?
Jeg **begynder** at **forstå**[*inf*] palæstinensernes synspunkter," siger Darko Richter, der er børnelæge i Zagreb.

and almost half (49.66%) of the occurrences of *lade* 'let' have the form

(coordinate conjunction) **lade** (pronoun) infinitive (preposition), e.g.

Hun havde snakket med barnet og forsøgt at få hende til at acceptere tilstedeværelsen af lillebroren så vidt det nu kan **lade sig**[*reflex.pron*] **gøre**[*inf*] at få et knap seksårs barn til at acceptere noget, hvis konsekvenser var uforudsigelige for alle andre.
"Men de **lod** tyskerne **gøre**[*inf*] det beskidte arbejde," kommenterede en kroatisk embedsmand, da champagnepropperne røg til vejrs og Bonns særlige udsending, Klaus_Peter Klaiber, udvekslede håndtryk med Kroatiens præsident Franjo Tudjman.
Men en gang imellem skal vi stole på vores kolleger i ledelsen **og**[*co.conj*] **lade være**[*inf*] **med**[*prep*] at forhale beslutninger ved ustandseligt at give vores uforgribelige mening til kende om en kollegas ansvarsområde.

The single most frequent combination with *tro* 'think, believe' within the profile is the pronoun *jeg* 'I'; this combination covers almost a third (30.91%) of all the occurrences, e.g.

Jeg tror nok, jeg råbte meget ad min mor, mens jeg boede hjemme.
Jeg tror tv_annonce_kampagner skader dansk rock meget.
Jeg er nu engang optimist, og **jeg tror**, at intolerancen kun i perioder er i stand til at dominere vores liv _ i sidste ende vil tolerancen sejre.

With *tage* 'take' the profile of weights above 4 s.d. covers only half of all the occurrences, and no clear pattern emerges at this point. To get a more detailed picture of how *tage* combines with other items, weights below 4 s.d. are considered. Figure 2 shows the collocational profile of *tage* as the proportional weight distribution of the 236 collocates on different positions.

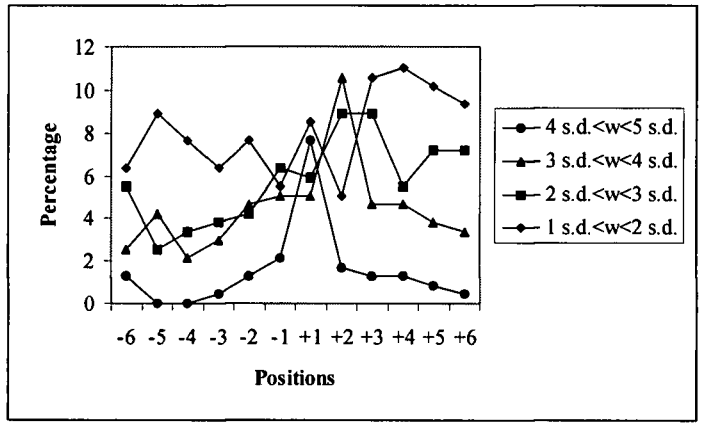

Figure 2: Collocational profiles for *tage* 'take' with weights (w) exceeding 1 s.d.

Beside the peak at position +1 with weights above 4 s.d., a peak with weights above 3 s.d. at the adjacent position, +2, is revealed. When this position is included, the profile covers 339

(83.91%) of the occurrences of *tage*. Weights above 2 s.d. have an equal load on positions +2 and +3, and when position +3 is included as well, the profile covers 394 (97.52%) of the occurrences. At this point a number of reiterated patterns have emerged, among which the three most frequent are:

tage noun preposition (noun), e.g.

"Ifølge advokaten vil det være bedst, om Louise får lov til at blive hos os, indtil landsretten har **taget stilling**[*noun*] **til**[*prep*], om hun i det hele taget skal anbringes på Josephine Schneiders Børnehjem i København for senere at komme hjem til sin far," siger Poul Wejlebjerg.

Regeringen vil efter et undersøgelsesarbejde til efteråret **tage initiativ**[*noun*] **til**[*prep*] at stramme administrationen af dapengelovgivningen og ændre forudsætningerne for at modtage dagpenge.

[D]e seneste overenskomster fra 1991 har **taget hul**[*noun*] **på**[*prep*] **problemet**[*noun*] med den manglende lønspredning.

tage pronoun noun (preposition), e.g.

Ved at analysere sine egne karriereankre, kan den enkelte medarbejder **tage et**[*indef.pron*] **medansvar**[*noun*] **for**[*prep*] at styre karrieren så den ikke ender i et job, han eller hun er mindre motiveret for.

Jeg **tog en**[*indef.pron*] **beslutning**[*noun*] **om**[*prep*] at prøve at se, om der var plads til mig i musikbranchen.

Arbejdsgruppen skal **tage sit**[*poss.pron*] **udgangspunkt**[*noun*] **i**[*prep*] de bestående lagre af lægemidler og forsyningsvejene i fredstid, herunder forsvarets lægemiddelberedskab.

tage pronoun ... noun, e.g.

Det **tager hende**[*pers.pron*] ingen **tid**[*noun*], hun glemmer aldrig noget, og hendes tøj er i topform ikke bare ved ankomsten, men under hele turen.

Det **tog mig**[*pers.pron*] syv **år**[*noun*] at få tingene på ret køl, efter jeg forlod Deep Purple første gang i 1973.

Voldtægtsforbryderen opholdt sig i sit offers lejlighed i ikke mindre end tre kvarter, men understregede selv, at det kun **tog ham**[*pers.pron*] seks **sekunder**[*noun*] at slå hende bevidstløs.

Within each pattern some lemmas are more prominent than others, and some lemmas figure in more than one pattern. The five most frequent noun collocates of *tage* are *stilling*, *chance*, *hensyn*, *initiativ*, and *betragtning*, all involved in several patterns including optional open slots:

tage (...) **stilling** (**til**) 'make up one's mind about'
tage (...) **chance** 'try one's fortune, take chances'
tage (...) **hensyn** (**til**) 'consider, take into account'
tage (...) **initiativ** (**til**) 'take the initiative in doing ...'
tage (...) (**med**) **i betragtning** 'consider, remember'

Each pattern attracts more collocations than free combinations, and thanks to the multi level representation even rare collocations are identified within the patterns designated by the profile because the inflected form or the word class itself has a high weight value. The pattern '**tage** noun preposition (noun)', for instance, occurs in 45 instances out of which 9 are hapax legomena but never the less valid collocations, e.g. **tage springet fra** 'take the plunge', **tage** ... **på sengen** 'take ... by surprise', **tage** ... **af bordet** 'withdraw ...':

Den 2.02 meter høje målmand Thomas Risum har **taget springet fra** Svendborg, Christian Lønstrup fra KB, Kenny Larsen fra Greve og ikke mindst Brian Rasmussen fra Vejle.

Og det så faktisk ud, som om vi **tog** ministeren **på sengen**.

I kan lige så godt **tage** Maastricht_traktaten **af bordet** med det samme.

The remainder of the instances include collocations with significant lemmas as *stilling, hensyn, initiativ* (cf. above), and open slot patterns as **tage** ... **fra** ... 'take ... (away) from ...' or **tage** ... **med** 'take away' e.g.

> Men da hun alligevel forsøgte at **tage** kniven **fra** ham, slog det klik for ham.
> M. **tog** bilnøglerne **fra** Svend Jensen før røveriet, og chaufføren måtte sidde magtesløs og vente på, at M. røvede banken og kom tilbage.
> Efter lidt snak frem og tilbage fik John Juhler Hansen lov til at **tage** skiltet **med** sig.

Conclusion

When the profile methodology is applied to the retrieval of word combinations from a digital text corpus, valid collocations are identified even under sparse data conditions. In the first place, the methodology reveals whether or not the target item chosen is involved in any significant relations with other items: If one or more peaks are found with high weight values, the answer is yes. The individual weight values indicate to what extent a particular item or a class of items is salient at a particular position relative to the target item. The comparison of the original corpus with a twin corpus with random word order secures, that the collocational candidates co-occur with the target item with a probability greater than chance. Each target item with collocational relations, has a unique collocational profile where some contextual positions carry a heavier load than others, and the profile is shaped by the number of items with weight values above a certain threshold. The profile delimits the collocational scope of the target item and therefore even low frequency collocates can be identified as long as they are located within the profile. Pursuant to the multi level representation, collocates that co-occur only once with the target item can be identified, provided that they belong to a class that fits into a significant syntagmatic pattern. Also salient collocates with weight values well above the threshold level within the profile can identify collocational relationships even outside the profile. These qualities of the profile methodology meet the objections against an approach to collocational significance based on restricted contextual position and recurrence. The collocational profile is like a syntagmatic finger-print for each target item. It shows which positions are especially salient with which significance and makes it possible to tailor the retrieval process with great precision.

References

[Altenberg 1990] Altenberg, B., 1990. Speech as linear composition, in: C. Caie et al. (eds.) *Proceedings from the Fourth Nordic Conference for English Studies, volume I*, pp. 133-143, Department of English, University of Copenhagen, Denmark.

[Altenberg 2001] Altenberg, B., 2001. On the Phraseology of Spoken English: The Evidence of Recurrent Word-Combinations, in: A.P. Cowie (ed.) *Phraseology. Theory, Analysis, and Applications*, pp. 101-124, Oxford University Press, Oxford.

[Biber et al. 1998] Biber, D., S. Conrad & Reppen, R. 1998. *Corpus linguistics. Investigating language structure and use*, Cambridge University Press, Cambridge.

[Church et al. 1991] Church, K., W. Gale, P. Hanks & D. Hindle, 1991. Using Statistics in Lexical Analysis, in: U. Zernik (ed.) *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pp. 115-164, Lawrence Erlbaum, Hillsdale, New Jersey.

[Clear 1993] Clear, J., 1993. From Firth Principles – Computational Tools for the Study of Collocation, in: M. Baker, G. Francis & E. Tognini-Bonelli (eds.) *Text and technology: In Honour of John Sinclair*, pp. 271-292, John Benjamins, Amsterdam.

[DS 2394-1:1998] DS 2394-1:1998. *Lexical data collections – Description of data categories and data structure – Part 1: Taxonomy for the classification of information types*, Dansk Standard, Copenhagen.

[Duncker 2001] Duncker, D., 2001. Modelling linguistic variation, in: A. Holmer et al. (eds.) *Proceedings of the 18th Scandinavian Conference of Linguistics*, Vol. 1., pp. 207-220, Lund University, Lund.

[Duncker forthcoming] Duncker, D., forthcoming. Phylogenetic variation in collocational practice.

[Duncker & Ruus 2000] Duncker, D. & H. Ruus, 2000. Multi Level Text Representation in an LCB, in: J.E. Mogensen et al. (eds.) *Symposium on Lexicography IX, Proceedings of the Ninth International Symposium on Lexicography April 23-25, 1998 at the University of Copenhagen*, Lexicographica, Series Mayor, Band 103, pp. 77-97, Max Niemeyer Verlag, Tübingen.

[Duncker & Ruus to appear] Duncker, D. & H. Ruus, to appear. Multi Level Text Markup, The Accumulative Approach.

[Dunning 1993] Dunning, T., 1993. Accurate Methods for the Statistics of Surprise and Coincidence, in: *Computational Linguistics* 19(1), pp. 61-74.

[Howarth 1996] Howarth, P.A., 1996. *Phraseology in English Academic Writing*, Max Niemeyer, Tübingen.

[Howarth 1998] Howarth, P.A., 1998. Phraseology and second language proficiency, in: *Applied Linguistics* 19(1), pp. 24-44.

[Johnson to appear] Johnson, M., to appear. Trading Recall for Precision with Confidence Sets.

[Keson 1999] Keson, B., 1999. *Vejledning til det danske morfosyntaktisk taggede PAROLE-korpus/The Danish Morphosyntactically Tagged PAROLE Corpus.* www.dsl.dk.

[Lewis 1969] Lewis, D.K., 1969. *Convention. A Philosophical Study*, Harvard University Press, Cambridge, Massachusetts.

[Lin 1998] Lin, D., 1998. Extracting Collocations from Text Corpora, in: *Workshop on Computational Terminology*, Montreal, Canada.

[Norling-Christensen 1996] Norling-Christensen, O., 1996. *Design and Composition of Reusable Harmonized Written Language Reference Corpora for European Languages*, MLAP PAROLE 63-384 WP 4.1.1.

[Pedersen 1996] Pedersen, T., 1996. Fishing for Exactness, in: *Proceedings of the South-Central SAS Users Group Conference (SCSUG-96), Austin, TX, Oct 27-29, 1996.*

[Perkins 1999] Perkins, M.R., 1999. Productivity and formulaicity in language development, in: C. Schelletter et al. (eds.) *Issues in Normal and Disordered Child Language: From Phonology to Narrative*, pp. 51-67, Special issue of The Bulmershe Papers, University of Reading.

[Renouf & Sinclair 1991] Renouf, A. & J. Sinclair, 1991. Collocational frameworks in English, in: K. Aijmer & B. Altenberg (eds.) *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, pp. 128-144, Longman, London.

[Ridings 1996] Ridings, D., 1996. *Text Representation in PAROLE.* MLAP PAROLE 63-384 WP 4.1.3.

[Ruus 2002] Ruus, H., 2002, A Corpus-based Electronic Dictionary for (Re)search, in: *Proceedings of the 10th EURALEX International Congress, Copenhagen (Denmark), 13-17 August 2002.*

[Sinclair 1966] Sinclair, J., 1966. Beginning the study of lexis, in: C. Bazell et al. (eds.) *In Memory of JR Firth*, Longman, London.

[Sinclair 1991] Sinclair, J., 1991. *Corpus, Concordance, Collocation*, Oxford University Press, Oxford.

[Smadja 1991a] Smadja, F.A., 1991a. Macrocoding the Lexicon with Co-occurrence Knowledge, in: U. Zernik (ed.) *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pp. 165-190, Lawrence Erlbaum, Hillsdale, New Jersey.

[Smadja 1991b] Smadja, F.A., 1991b. From n-grams to collocations. An evaluation of Xtract, in: *Association for Computational Linguistics. Proceedings of the conference 18-21 June 1991, University of California*, pp. 279-284, Berkeley, California.

[Smadja 1993] Smadja, F.A., 1993. Retrieving Collocations from Text: Xtract, in: *Computational Linguistics* 19(1), pp. 143-177.

[Toolan 1996] Toolan, M., 1996. *Total Speech. An Integrational Linguistic Approach to Language*, Duke University Press, Durham.

[Wray & Perkins 2000] Wray, A. & M.R. Perkins, 2000. The functions of formulaic language: an integrated model, in: *Language & Communication* 20, pp. 1-28.