A statistical and structural approach to extracting collocations likely to be of relevance in relation to an LSP sub-domain text

Bjarne Blom
Department of Lexicography and Computational Linguistics
The Aarhus Business School
bb@lng.hha.dk

1. Background.

This article[*] sketches a method by means of which likely text-relevant collocational word strings may be extracted from sub-domain LSP-texts. Because of the repetitive nature of collocations, a statistical and structural approach is suggested. The goal of this project is two-fold: **(a)** to explore the degree to which computational methods are suitable for extracting collocations; **(b)** to explore the degree to which it is possible to extract information particular to the overall topic or domain of a LSP text, i.e. terminology, by means of knowledge-poor techniques. As this study has run on a time-limited basis, most importance has been placed on (a).

2. What is understood by "collocation".

A "collocation" is often defined as either "an arbitrary and recurrent word combination" (Benson 1990) or "the cooccurrence of two or more words within a short space of each other" (Sinclair 1991). However, such definitions tend to leave two important questions out of consideration, cf. the two following examples:

(a) ... shouldn't have done *that. This* prompted........
(b) .... seems unlikely *that this* should be the case

The two definitions above, which are in effect fairly representative of the major part of definitions offered, do not take two points into considerations, namely the distinction between lexical collocations like *prime minister* and syntactic collocations like *that this* - or the issue whether words should be allowed to cross sentence boundaries. Relying on either of the above definitions would result in massive overgeneration, so consequently we will define the concept of "collocation" as to better suit a quantitative approach:

> A collocation is a word string consisting of a minimum of two words[1] with the following characteristics:

[*] This is a brief description of the method designed and the findings made in connection with the research project UDOG (Udforskelse af Dansk Ordforråd og Grammatik - *Exploration of Danish Vocabulary and Grammar*) under the auspices of The Danish Research Council for the Humanities. A broader outline is given in Danish in Blom 1997 and (1998 - forthcoming).

[1] For the purposes of this study, we will impose the restriction on the minimum word string length that it must contain at least one content word, cf. section six.

(1) the entire word string is occuring within a given textual segment *a*.

(2) the least frequent word(s) in the word string occur(s) at least *b* times in the corpus.

(3) all words in the word string occur with a span of *c* words to its neighbour collocate

(4) all words in the word string occur in *d* particular sequences

This definition is rather a set of parameters to be adjusted according to the particular scientific context in which they are to be utilised, thus **(1)** refers to whether a word string is considered to be a phrase, a syntagma, an entire sentence, or even permitted to cross sentences boundaries; **(2)** refers to a lower threshold value decided on for a given purpose in order to filter off infrequent occurrences ; **(3)** refers to the issue whether interrupted structures should be taken into account; **(4)** refers to the question whether inverted structures should be taken into account. In this article we let *a* be sentence level, i.e. we accept a clausual structure as a collocation rather than imposing some arbitrary value as a maximum word string length; we let *b* be four, i.e. a word occurring three times or less is not considered to be significant; we let *c* be zero, i.e. we do not include interrupted structures, as such structures require quite an extensive statistical setup (cf. Ikehara, Shirai & Uchino 1996), and such structures are outside the scope of this study; we let *d* be left-to-right, in the sense that we do not lemmatize bigram structures, i.e. *afgift betales* = *betales afgift* in order to secure the inclusion of underrepresented items. However, in case inverted structures are significantly represented in the corpus, they will be included as valid collocations.

3. Finding likely relevant words (unigram level).

The approach is founded on the assumption that a word has both statistical and grammatical properties which may serve as a clue to its terminologicalness, at least this is believed to be the case for very hard-core LSP texts. As an example of such a text, the Danish VAT act has been chosen. The legislative genre has been carefully selected, as legislative terminology tends to be unambiguous in the sense that one particular concept is usually backed by one particular orthography. As opposed to LGP, the use of synonymy, paraphrases and other stylistic ways of representing a concept orthographically is minimised in order to avoid interpretative scope on the legal circumstances relating to a given concept. As a bonus, we may expect polysemy not to constitute a serious problem for this kind of study.

Zipf (1949) suggested that there seems to exist an inversely proportional interdependence between frequency and rank in a text, in that a very limited number of words occur with very high frequencies, whereas a very large number of words occur with very low frequencies. Damerau (1965), Dennis (1967) and Stone & Rubinoff (1968) discovered a link between frequency, typicality - and word class, so that so-called function words (which tend to account for quite a considerable part of words found at the high-frequency end of the Zipfian distribution) and content words have different statistical behaviour in texts, as the former category follows a poisson distribution closely, whereas the latter does not. These works provide theoretical foundation for the intuition that a word like "that" has a greater intertextual dispersion than a word like "thermodynamics". Bookstein & Swanson (1974) (1975) and Harter (1975) sharpened this point by offering an explanation why certain content words (like e.g. *red, say, man*) have greater intertextual dispersion than other content words (like e.g. *enterprise, sub-clause, taxable*). Their claim was that words which show a random distribution in a poisson process are likely not to contain information about the text in which they occur, whereas words which do not follow a poisson distribution tend to contain information about the text in which they occur.

For this study, we will rely on the very simple assumption that given the fact that we are dealing with a very specialised sub-domain LSP text, i.e. a legislative genre text within the domain of taxation, text-relevant words are likely to be found among high-frequency content words. For this end, three steps have been necessary: (1) the text has been tagged for word class, (2) words sharing an identical stem have been lemmatised to avoid insignificance arising from the underrepresentation of a given inflected form, and (3) a very simple and limited stop list has been applied to filter off spatio-temporal items (eg. names of various EU countries or words referring to a deadline by which a given task is to be carried out), which tend to be represented with some frequency owing to the nature of the text, however is likely to be irrelevant for the particular legal domain in question. In order to determine a suitable threshold value for filtering off words with no or very little relevance, a number of sub-sections of the VAT act have been randomly picked in order to compare the frequency of occurrence of words in the sub-section with their frequency of occurrence in the entire text. This was done in order to find a suitable cut-off point below which words are likely not to be of relevance to the text (for an example of this, see Blom 1997), and there seemed to be a case for omitting items below the frequency of four.

4. Finding words with collocational potential (unigram level).

The output consists of a list of unigrams rendered text-relevant[2] by the method. What we need to find out now is which of these unigrams tend to occur as either single words or as part of a multi-word unit.

There are existing statistical methods for testing the "bondness" of word pairs, the most prominent one being mutual information, cf. Church & Hanks (1990). Mutual information is a widely used frequency-based formalism that calculates the probability whether a word pair occurs together or separately in a text. However, this approach is not adequate for this study, as we place our focus on content words. In mutual information statistics, both words of a bigram receive equal weight irrespective of grammatical status. This would mean that function words might exercise undue influence over the overall MI-value and thus bias the measure. This is no unreasonable assumption, if we consider the fact that the most frequent words in a Zipfian distribution tend to be function words. In this study, we restrict our scope to keeping content words as "nodes" and then examine the way they combine with their either left or right adjacent "collocates".

We will use the heuristics that a content word's ability to enter into a "bond" with another word depends on its contextual distribution, i.e. its number of adjacent either left or right words in a text. The fewer such adjacent words, the more a word's unigram frequency will be distributed upon *particular* collocates. In case a content word has a large number of collocates, its unigram

[2] Future versions of the method will try to incorporate more sophisticated heuristics such as the way a given content word is distributed over the entire number of sentences, or the number of content words used to describe a given content word in the entire text.

frequency will be watered out. Consider a content word which occurs fourteen times in a text and has an equal number of collocates, then each bigram gets a frequency of one. On the basis of these statistics, we can safely assume that this content word definitely constitutes a single word unit. If, on the other hand, a content word occurring fourteen times is represented two times with one particular adjacent word and twelve times with another particular adjacent word, we can deduct that this content word is predominantly part of a multi-word unit. For the purposes of this heuristics, we apply the following formula in which *u* is the unigram frequency of the content word and *c* is the number of collocates of the content word:

5. Finding word-pairs with collocational strength (bigram level).

Our present output consists of the unigram list from the previous step exclusive of words which the method does not render potential multi-word items. We will now explore the above idea a bit further, as we will focus on the strength of any given bigram in which a particular content word occurs in order to find the strength of the bond of the word-pair. The bond strength depends on the number of times a word occurs in a given bigram compared to how often the word occurs as a unigram. Our assumption is that *content word n* exercises good "attractive power" on *word x* if the bigram frequency (fq_{nx}) accounts for a major part of the unigram frequency of content word *n* (fq_n). We will try to make this statistical parameter differentiate between words with the same unigram/bigram-ratio, but with different frequencies, eg. the instance where two words with ten collocates each occur with a unigram frequency of ten and a hundred, respectively. We will use this formula, in which *u* is the unigram frequency and *b* is the bigram frequency, to discriminate between superior and less superior collocational strength:

6. Expanding the bigrams (sentence level).

The output from the last two steps consists of a list of bigrams which are, on the basis of the statistics applied, considered to be of a multi-word nature rather than a single-word nature. Now we will expand each bigram to find each possible word string in which the bigram occurs. In stead of merely extracting the longest possible bigram like for instance Smadja (1993), we will permit all syntactically well-formed word strings in which a candidate bigram occurs. To this end a contextual array is applied, in which a given bigram is shown in all its contexts, so that the horizontal axis accounts for syntagmatic values, while the vertical axis accounts for paradigmatic values. The following rules apply to the expansion of orthographic word strings: **(1)** Only word strings (non-expanded as well as expanded bigrams) occurring at least three times are accepted; **(2)** A bigram may be expanded by ±one position only if the same word occurs at the same syntagmatic position in the contextual array; **(3)** A bigram is left out of consideration in case a longer word string occurs with at least the bigram frequency minus one occurrence. This is to avoid syntactically incomplete structures, the frequent occurrence of which can be ascribed to syntactic reasons only. The longer word string which the same or nearly the same frequency as the bigram is likely to be the valid one. An easier solution would be to favour certain syntactic structures, typically noun syntagmas or predicative-like structures like Smadja (1993), or to limit the study to bigrams containing only content words, however this would idiosyncratically favour *betale afgift* {C-C}to *betaling af afgift* {C-F-C}.

In a Zipfian distribution of words, the high-frequency end is dominated by grammatical words like prepositions, conjunctions and the like, which is why a great amount of syntactic collocations is to be expected in any given bigram output. In order to reduce the likely amount of overgeneration, we will take into account only bigrams that contain at least one content word and which does not include a punctuation mark. Having taken this preventive step, we will assume that syntactic noise is likely to occur at the left- and rightmost positions in a word string where a function word may be found. In stead of merely leaving an expert with the tedious task of filtering off items irrelevant owing to syntactic noise, we will apply a set of syntactic rules to serve as a filter for such syntactic noise.

There are four syntactic main rule classes (1-3), one syntactic exception rule class (4) and one statistical exception rule class (5) as follows: **(1)** a particular tag (or tag sequence) cannot take up the leftmost slot (or slots) of a word string; **(2)** a particular tag (or tag sequence) cannot take up the rightmost slot (or slots) of a word string; **(3)** a particular tag sequence cannot form either an entire word string or a part of a given word string counting from the second left- or rightmost slot; **(4)** a particular tag sequence discharged by a syntactic rule may qualify, if it matches a specific syntactic pattern; **(5)** a particular tag sequence left out by a syntactic rule may qualify if a content word is immediately succeeded and/or preceded by a preposition in at least 80% of all cases. The latter rule is used to statistically identify phrasal verbs (*berettige til*), as well as complex prepositions (*i henhold til*). All occurrences of {vb-prep}and {prep-sb-prep}have been examined in order to evaluate the performance of this 80%-rule. All instances of phrasal verbs and complex prepositional phrases are correctly identified. Consequently, this technique might even prove to be a theoretical spinoff of this project.

Example of a Class (1) rule:
input: "den i stk. #num#" {det-prep-noun.abbr-num}
main rule: a word string cannot be initiated by tags {det-prep}- reduce word string by these words.
output: "stk. #num" {noun.abbr-num}

Example of a Class (2) rule:
input: "registrere som landbrug og" {vb.inf-konj-noun-konj}
main rule: a word string cannot be ended by tag {konj}- reduce word string by this word.
output: "registrere som landbrug"{vb.inf-konj-noun}

Example of a Class (3) rule:
input: "opgøres på grund af" {vb.pres.pass-prep-noun-prep}
main rule: a word string cannot be equivalent to tags {vb.opt[3].opt-prep-noun-prep}where *prep-noun-prep* is a
complex prepositional phrase - omit entire word string.
output: Ø

Example of a Class (4) rule:
input: "den i stk #num# omhandlede afgift" {det-prep-noun.abbr-num-vb.part.adj-noun}
main rule: a word string cannot be initiated by tags {det-prep}- reduce word string by these words.

[3] This notation refers to optional filling of slot. The slots may be filled by a tag referring to a verbal tense, voice and mood.

exception: a word string initiated by tags {det-prep}cannot be reduced, if the word string is ended by tags
{num-vb.pret.adj-noun}[4].
output: "den i stk #num# omhandlede afgift" {det-prep-noun.abbr-num-vb.part.adj-noun}

Example of a Class (5) rule:
input: "fritage for" {vb.inf-prep}
main rule: a word string cannot be ended by tag {prep}- reduce word string by this words.
exception: a word string ended by tag {prep}cannot be reduced, if {prep}has in its immediate left position the
tag {vb.opt.opt}, and {vb.opt.opt}has in all its lemmatized forms the orthographic representation of
{prep}in 80% or more of all cases.
output: "fritage for" {vb.opt.opt-prep}

.... mstændigheder	**fritage**	en virksomhed for.......
.... kan indrømme	**fritagelse**	*for*..........................
.........er kan meddele	**fritagelse**	*for* forhøjelse af..........
...........omhandlende	**fritagelser.**	d . Spørgsmål 1........
........f virksomheder	**fritage**	*for* at svare afgift........
.........arer ville være	**fritaget**	*for* afgift efter §.........
........dre EF-lande er	**fritaget**	*for* afgift : 1)............
.......heder ville være	**fritaget**	*for* afgift . Stk. 5........
........ndet ville være	**fritaget**	*for* afgift . 2)............
...... stk. 1 og 3 , er	**fritaget**	*for* at svare afgift........

Fig. 1. Concordance for the lemma "fritage[-]" (frequency of occurrence: 10) and its immediate right collocates. In eight out of ten cases, the lemma "fritage" has "for" as its adjacent right collocate. In this case, the {prep}tag is deemed to be a particle rather than a preposition, and thus a valid part of a phrasal verb.

7. Evaluating the syntactic filter.

The syntactic filter was created on the basis of the orthographic output strings. The output was examined in order to distinguish between word strings which should either be permitted as syntactically well-formed or omitted as syntactically ill-formed. This was done in order to deduct syntactic rules to be generalised with a view to creating the best possible recall-precision rate.

48% of all orthographic word strings were syntactically ill-formed, whereas 52% were well-formed. This 50-50 ratio does certainly imply a certain higher principle of randomness, which might make syntactic formalization impossible. However, if we examine syntactic tag sequences among the group of omitted word strings, we discover that some 96% are actually unique,

[4] This rule aims at including a syntactic pattern particular to the legal domain in Danish in which the head of a nominal phrase is pre-modified by a determiner followed by a preposition.

whereas some 4% share a syntactic sequence with a member of the group of syntactically well-formed word strings. Within the group of omitted word strings with a unique tag sequence, 99% may be subject to formalization, whereas only 1% could not be formalised into any operational rule. This one per cent account for a minor loss of precision (-0,6%). The formalization of the 99% word strings did not conflict with rules applicable to the syntactically well-formed word strings. All of the remaining 4% ambiguous word strings could be formalised, however with a moderate loss of recall (-3,4%). The syntactic filter works with 96,6% recall and 99,4% precision, and provides some indication that syntax usage in sub-domain LSP texts like the present one is of a nature which accommodates formalisation rules.

8. How to evaluate the terminological relevance of the word strings extracted.

The syntactic filter ensures a well-formed output, but this is of course no guarantee that the output meets a certain qualitative standard as to text typicality. How can we know that the word strings extracted are not merely commonplace collocations found in virtually any text. One obvious way of judging the output would be to have an expert evaluate it, however one obvious drawback to this approach is the circumstance that experts tend to have different and sometimes even conflicting opinions. However another method might be to test how the word strings are dispersed on a cross-section of texts from various domains in order to find out whether the same collocations tend appear in one or more other texts, or whether they tend to be restricted to the Danish VAT Act. Obviously, we cannot look for the collocations in all texts ever written, but we can ascertain with statistical certainty whether a given word string is restricted to the VAT Act or not. Evaluation of LSP relevance has been left out of consideration for this present study, but is an interesting issue for further research.

References

Benson, M. (1990): Collocations and general-purpose dictionaries. *International Journal of Lexicography 2,* pp. 1-14.

Blom, B. (1997): Om statistisk og strukturel afgrænsning af sandsynlige teksttypiske kollokationer i Momsloven. In: *UDOG-rapport 6,* pp.3-23.

Blom, B. (1998 - forthcoming): A method for identifying collocations likely to be relevant in relation to a sub-domain LSP-text. *UDOG-rapport 7.*

Bookstein, A. & Swansson, D. R. (1974): Probabilistic models for automatic indexing. *Journal of the American Society for Information Science, 25,* pp. 312-318.

Bookstein, A. & Swansson, D. R. (1975): A decision-theoretic foundation for indexing. *Journal of the American Society for Information Science 26,* pp. 45-50.

Church, K. W. & Hanks, P. (1990): Word association norms, mutual information and lexicography. *Computational Linguistics 16-1* pp. 22-29.

Damerau, F. J. (1965): An experiment in automatic indexing. *American Documentation 16,* pp. 283-289.

Dennis, S. F. (1967): The design and testing of a fully automatic indexing-search system for documents consisting of expository text. In: G. Schechter (ed.) *Information Retrieval: A critical review*, pp. 67-94. Thompson Books Co., Washington.

Frantzi, K. T. & Ananiadou, S. (1996): Extracting nested collocations. *Proceedings from COLING'96*, pp. 41-46.

Harter, S. P. (1975): A probabilistic approach to automatic keyword indexing. Part 1: On the distribution of specialty words in a technical literature, Part 2: An algorithm for probabilistic indexing. *Journal of the American Society for Information Science 26*, pp. 197-206; pp. 280-289.

Ikehara,S., Shirai, S & Uchino, H. (1996): A statistical method for extracting uninterrupted and interrupted collocations from very large corpora. *Proceedings from* COLING'96, pp. 574-579.

Sinclair, J. (1991): Corpus, concordance and collocation. Oxford University Press.

Smadja, F. (1993): Retrieving collocations from text: Xtract. *Computational Linguistics 19-1*, pp. 143-178.

Stone, D. C. & Rubinoff, M. (1968): Statistical generation of technical vocabulary. *American Documentation*, pp. 411-412.

Zipf, G. K. (1949): Human behaviour and the principle of least effort. Addison-Wesley, Cambridge, Massachusetts.