Collocations and statistical analysis of n-grams: Multiword expressions in newspaper text

Gunn Inger Lyse and Gisle Andersen

Abstract

Multiword expressions (MWEs) are words that co-occur so often that they are perceived as a linguistic unit. Since MWEs pervade natural language, their identification is pertinent for a range of tasks within lexicography, terminology and language technology. We apply various statistical association measures (AMs) to word sequences from the Norwegian Newspaper Corpus (NNC) in order to rank two- and three-word sequences (bigrams and trigrams) in terms of their tendency to co-occur. The results show that some statistical measures favour relatively frequent MWEs (e.g. *i motsetning til* 'as opposed to'), whereas other measures favour relatively low-frequent units, which typically comprise loan words (*de facto*), technical terms (*notarius publicus*) and phrasal anglicisms (*practical jokes*; cf. Andersen this volume). On this basis we evaluate the relevance of each of these measures for lexicography, terminology and language technology purposes.

1 Introduction

Multiword expressions (MWEs) may be defined as words that co-occur so often that they are perceived as a linguistic unit (for instance *pit stop* and *by and large*). Linguistically, MWEs comprise several phenomena, ranging from idioms, semi-fixed expressions, foreign expressions (such as anglicisms in Norwegian) and technical terminology. MWEs are surprisingly ubiquitous in natural language, being estimated to be as frequent as one-word expressions Jackendoff (1997). The identification of MWEs is therefore pertinent for a range of tasks within lexicography, terminology and language technology, including for instance the correct segmentation of phraseological units and the extraction of terminology (*ulcerøs kolitt, notarius publicus*).

In the field of Natural Language Processing, MWEs are sometimes referred to as "a pain in the neck" (Sag et al. 2002), because their meanings usually cannot be determined compositionally from the meanings of the individual words. In the context of machine translation, for instance, this means that the system needs to know if a sequence of words can be translated word by word or if it has a special meaning, requiring a particular translation, in virtue of being an MWE. Moreover, since many MWEs have a marked syntax, they may seriously impede syntactic parsers (cf. the expression *by and large*, which is a juxtaposition of a preposition and an adjective). Norwegian, like German but unlike English, follows the convention of representing compounds as one word; hence MWEs are not as relevant for the identification of (domestic) compounds as is the case for English. Nevertheless, for the Norwegian newspaper project it is desirable to explore the vast amount of data by identifying MWEs that are lexicographically or terminologically relevant.

The identification of MWEs is valuable for several purposes. First, multiword expressions are needed in lexical databases used for general lexicographical purposes as well as for NLP purposes. Recurring MWEs should be systematically identified and correctly segmented in a corpus-driven approach, and added to the national lexical database, the Norwegian Word Bank (cf. Fjeld this volume). Second, we know that the syntactic tagger used by the Norwegian Newspaper Corpus (NNC), the Oslo-Bergen tagger (Fjeld and Nygaard this volume) makes errors related to MWEs, especially pertaining to phrasal prepositions such as *på grunn av* 'because of, due to' and adverbs such as *i tide* 'on time'. These should be segmented as phrasal units and not processed further by the tagger. Therefore the overall performance of the tagger may be improved through added knowledge about multiword expressions in Norwegian text. Third, technical terminology is very often realised as MWEs, and the identification of recurrent collocational patterns is relevant for term extraction, even in non-technical texts such as newspaper language.

In line with the "re-emergence of empirical linguistics" Abney (2000), statistical methods have been introduced as a way to quantify an intuition about words that "belong together" (e.g. Church et al. 1991, Baldwin and Bond 2002, Banerjee and Pedersen 2003, McInnes 2004, Evert 2004). So-called association measures (AMs) analyse the relation between how often words in a sequence occur together and the frequency of each of the words individually.

However, the interplay between statistical measures, corpus material and the identification of MWEs is still not very well explored, among other things because the choice of statistical measure depends on which type of MWEs one is attempting to extract. The norm seems to be to provide a variety of statistical measures with minimal guidelines as to which will probably suit the needs of the user best (Banerjee and Pedersen 2003, Evert 2004, Baldwin 2004, Baldwin and Kim 2010). In other words, there seems to be a knowledge gap in terms of how to use (and how to choose) association measures to extract MWEs. To explore this relation further, we have applied nine common statistical measures to two-word sequences (bigrams) in the NNC and four statistical measures to three-word units (trigram).

Our main objective is to evaluate the usefulness of the alternative association measures, when applied to a large set of Norwegian data, in terms of their ability to pick out relevant MWEs representing the different lexical and terminological categories sketched above. This is based on the *a priori* assumption that certain AMs will pick out items with a relatively low frequency and thereby be better at finding rarely used technical terms (*trojansk hest* 'trojan horse' in computing), anglicisms (*corned beef, practical jokes*), other foreign expressions (*per capita* 'per head', *gefundenes fressen* 'sensational news'), and possibly also different domestic MWEs (e.g. the dessert *tilslørte bondepiker*). Other AMs may be better suited for picking out high-frequency multiword units such as phrasal prepositions and adverbs (*stort sett* 'mostly', *blant annet* 'among other things'). Yet other AMs may well be better suited for the automatic detection of multi-word proper nouns (*Gro Harlem Brundtland*), which are of less lexicographical value but important for NLP purposes such as named entity recognition.

The paper is structured as follows. Section 2 provides an overview of general concepts relevant for the study of phraseology and MWEs. Section 3 describes the material used and the methods applied in order to extract collocational statistics and test the association measures. In section 4 we present the results and discuss the relevance of the different association measures from the point of view of lexicography and terminology, while section 5 contains some concluding remarks and proposals for future work.

2 Background

2.1 Multiword Expressions (MWEs)

The concept of MWE is an attempt to capture the intuition that meaningful units in our language are often larger than individual words. Intersecting with a wide range of linguistic phenomena, Baldwin and Bond conclude that "there is much descriptive and analytic work on MWEs that has yet to be done" (Baldwin and Bond 2002: 3). In so-called compositional semantics, open-class words (nouns, verbs, adjectives and adverbs) are commonly assumed to have a lexical meaning which contributes to the meaning of an utterance, i.e. the meaning of the utterance is composed of the meanings of the parts. Linguistic analyses of meaning often rest solely on words as a basic lexical unit, which is also reflected in the way we tend to organise our vocabulary in terms of lexicons or dictionaries that are usually based on looking up individual words.

This approach poses problems when we encounter sequences of words where the meaning is *not* unambiguously composed of the meaning of the parts. There is an emerging awareness that MWEs are not just sporadic exceptions in our vocabulary (Sinclair 1991, Stubbs 1996, Sinclair 1996, Tognini-Bonelli 2001, Biber 2009). Jackendoff (1997) estimates that MWEs are as common as simplex words in our vocabulary; similarly Sag et al. (2002) assert that 41 per cent of the entries in WordNet (Fellbaum 1998) are multiword units.

We adopt the definition by Baldwin and Kim (2010: 3), who define MWEs as units that (i) can be decomposed into more than one (space-separated) lexical unit; and (ii) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity. This is a useful definition because it states that some form of idiomaticity is a necessary feature of MWEs, as opposed to applying loose criteria (as an example, the 'single-word phrasability' criterion states that if an MWE can be paraphrased by a single word, then it usually counts as an MWE).

'Idiomaticity' is taken to mean that the expression is somehow 'marked' or deviates from the linguistic properties of the component words, and may apply at the lexical, syntactic, semantic, pragmatic, and/or statistical levels (Baldwin and Kim 2010: 4), as described below. From this definition it also follows that MWEs are somehow lexicalised, in virtue of having some additional property which cannot be deduced directly from its component words.

Following Baldwin and Kim (2010), *Lexical idiomaticity* is defined as an expression being "lexically marked" in that one or more lexical components of the expression are not part of the conventional lexicon of the language in question (consider the Latin phrase *de facto*, or a phrasal anglicism such as *bottom line* in Norwegian). Since such phrases are not part of the domestic vocabulary, lexical idiomaticity usually entails semantic idiomaticity, since the meaning of the expression is then usually not deduced from the meanings of the parts (unless the listener has adequate knowledge of the language from which an expression is taken and the meaning of the phrase can be deduced from its parts in that language). For the same reason, lexical idiomaticity usually also entails syntactic idiomaticity.

Syntactic idiomaticity, as defined by Baldwin and Kim (2010), occurs when the syntax of the MWE deviates from the constructions that one would expect in the given language—although in that case, it is perhaps more accurate to call it *morphosyntactic* idiomaticity. Consider the Norwegian adverb *i tide* 'on time', whose archaic word form *tide* does not belong to the ordinary inflectional paradigm of the lexeme *tid* 'time', making *i tide* syntactically idiomatic. The noun cannot be derived by the rules of the

grammar, but its use in this idiomatic context has to be learnt.

Semantic idiomacity is the property that the meaning of an MWE is not fully predictable from its component parts. The Norwegian expression *på kant med* 'in disagreement with' is semantically marked in that its meaning is not derivable from the components *på* 'in/on', *kant* 'edge', *med* 'with'. The expression is, incidentally, also syntactically marked in that one cannot inflect the noun or insert a modifier before the noun (**på kanten med* lit. 'on the edge with'), illustrating why such idiomatic expressions often cause problems for foreign learners.

Baldwin and Kim (2010: 5) observe that there are a range of borderline examples of semantic idiomacity, where the meaning is partially predictable, for instance due to metonymic extensions (Halverson this volume) or the use of metaphor (*around the clock*, meaning '24 hours', refers to counting the hours on a clock as a metaphor.)

Pragmatic idiomaticity pertains to MWEs that are associated with particular communicative situations or discourse contexts, for instance expressions like *good morning, how do you do* or *welcome back*, performing specific pragmatic functions, e.g. at the speech act level (e.g. Sag et al. 2002).

Finally, *statistical idiomacity* is the phenomenon of particular combinations of words occurring with markedly higher frequency in comparison to alternative phrasings of the same concept. Statistical idiomaticity thus appears to correspond to what Sag et al. (2002: 7) refer to as 'institutional phrases'. For instance, there is no principled linguistic reason for not saying *computer translation* when meaning *machine translation*, or *pepper and salt* instead of *salt and pepper*, but statistically we find that particular lexicalisations are simply more frequent. Statistical idiomacity encompasses notions such as 'naturalness' and 'conventionalisation' of word sequences; for instance one may say *strong tea* and *powerful car*, but their so-called anti-collocations **powerful tea* and **strong car* are markedly less common.

A problem in delimiting MWEs is that there are many sequences of words which intuitively have a strong association but which may or may not be characterised as MWEs– consider for instance so-called "formulaic sequences" such as *I don't want to, the fact that*. Biber et al. (1999: 999) introduce the term 'lexical bundles' to denote "sequences of word forms that commonly go together in natural discourse" and that are characterized by a high frequency. Lexical bundles are compositional expressions that are not lexicalized, although they may become so in the course of time, and which may allow for a certain variability (e.g. Biber 2009).

There are several phenomena that intersect with the MWE category, although not all examples of these phenomena are multiword units. Therefore, when we attempt to identify MWEs automatically, we expect to discover that the identified MWE candidates belong to different linguistic categories. One category is technical terminology, defined by Baldwin (2004) as "a lexical unit consisting of one or more words which represents a concept inside a domain". Although terms are not necessarily multiword units, Sag et al. (2002: 2) observe that "specialized domain vocabulary, such as terminology, overwhelmingly consists of MWEs". According to Baldwin and Kim (2010: 11), the field of terminology is "broader in scope than MWEs in the sense that simple lexemes can equally be technical terms, and narrower in the sense than non-technical MWEs are not of interest to the field". Anglicisms are often lexicalised phrases (Andersen this volume) such as *due diligence, easy listening, straight edge*, etc.. Further, even though compounds are written without whitespace in Norwegian, it is not inconceivable that the systematic retrieval of MWEs may identify certain compounds that are commonly spelt as separate words in disagreement with the spelling norms (the phenomenon known as *særskriving*). Finally there are named entities (names of persons,

places, events, organisations, titles, expressions of time and quantity, numerical expressions, etc. (*New York, Melodi Grand Prix, cand. scient., tomorrow morning, one million*).

Among these phenomena, there are certain kinds of multiword sequences that we are particularly interested in identifying in the context of the Norwegian Newspaper Project, due to their potential relevance for general lexicography and terminology. This applies to multi-word terms from various professional domains, such as *amyotrofisk lateralsklerose* (medicine), *pater familias* (law) or *vennligsinnet oppkjøp* 'friendly acquisition' (business), as well as recurrent MWEs in general language, i.e. idioms such as *gemene hop* 'common folk', conventionalised metaphors such as *hellige kuer* 'holy cows' and multiword anglicisms such as *make or break* or imported interjections and discouse markers like the irony marker *yeah right*.

As we will see, different association measures vary in terms of their ability to retrieve elements in these groups.

2.2 Collocations

Sag et al. (2002) define a 'collocation' as an arbitrary statistically significant association between co-occurring items; that is, collocations subsume MWEs.

Association measures (AMs) are statistical measures that calculate the association strength between tokens in an n-gram. An n-gram is a sequence of n units, in our case a unit is a string of characters separated by white space. A bigram contains two units, a trigram contains three units, etc. There are many statistical AMs available; some are relatively simple to calculate whereas other are more complex. Two freely available statistics software packages illustrate the range of AMs that have been suggested in the literature: The UCS Toolkit[1] provides a repository for almost 30 AMs for bigrams (Evert 2004). The n-gram Statistics Package (NSP)[2] provides 13 AMs for bigrams and four measures for trigrams (Banerjee and Pedersen 2003).

The relation between statistical measures of collocation and linguistic concepts such as MWEs has not been fully explored in the literature. Baldwin and Kim (2010: 23) observe that although AMs have been applied to a wide range of extraction tasks over a number of languages, the general finding is that it is often unpredictable which association measure will work best for a given task in a given language. Thus, it seems to be the norm to present a group of "common statistical measures" with minimal statements relating to the applicability of each statistical test.

The approach taken in this paper is that, although the numerical values for ranking n-grams according to collocational strength are not comparable across statistical measures, we may evaluate the different measures applied on the same data by comparing whether one n-gram is ranked higher or lower than another. Our intension is therefore to scrutinise the kinds of MWE candidates that are captured by each statistical measure, in an attempt to improve our understanding of the relation between statistical AMs and linguistic phenomena of MWEs, as observable in a large Norwegian dataset. To this end, we adopt a subset of the "common statistical measures" that are given in the UCS toolkit and in the NSP package. Although we follow the definitions given from these two sources, all calculations have been performed in Common LISP by the authors.

[1] URL: http://www.collocations.de. Last verified July 8, 2010.

[2] URL: http://www.d.umn.edu/ tpederse/nsp.html. Last verified July 8, 2010.

3 Methodology

3.1 Data and n-gram extraction

In order to apply statistical AMs, all word sequences in the NNC first had to be organised into lists of n-grams. In the description below, n-grams are represented in square brackets in which the first element is the frequency of the n-gram, and the rest is the sequence of words. For instance, the bigram [38759 *på forhånd*] tells us that the bigram *på forhånd* 'in advance, beforehand' occurred $38,759$ times in the corpus.

In order to maximise the amount of text from which n-grams were extracted, we merged the texts from the NNC (Andersen and Hofland this volume) with Norwegian newspaper texts collected by the company Nordisk Språkteknologi. This additional dataset contains electronic text from the newspapers *Bergens Tidende*, *Aftenposten* and *VG* from the 1980s and 1990s. The two sources were combined in order to broaden the time span of the newspaper articles, and the resulting corpus material thus covers three decades and contains both Bokmål and Nynorsk, the two written standards of Norwegian.

The extraction of n-gram lists from the corpus of newspaper texts was done by Knut Hofland at Uni Computing. He first filtered out boilerplates and other elements that do not belong to the core text (Andersen and Hofland this volume). For the subsequent multiword extraction, only bigrams and trigrams have been considered so far, motivated by considerations of re-usability, allowing a user to modify the original list according to the specific needs of a given task. The n-grams were compiled according to the principle of keeping the data as unmodified as possible, in order to allow subsequent users of the data to modify the original n-gram lists according to the specific needs of a given task. Specifically, case-sensitivity was retained, i.e. n-grams that occur both with lower and upper case are counted separately, allowing for subsequent users to count them separately or collapse them into joint n-grams.

For the same reason, the n-grams also record all punctuation marks as separate tokens in a sequence. Finally, in this strictly corpus-driven approach, the extraction of n-grams or the subsequent analysis does not rely on any linguistic or other annnotation of the data, such as parts of speech, lemma information or the like.

3.2 Post-processing of n-gram lists

The initial analysis of the raw n-gram lists and preliminary experiments for MWE extraction revealed that it would be beneficial to filter the data by removing certain n-grams prior to the application of statistical AMs. First, more than half of the n-grams only occurred once in this large dataset, and we have no reason for claiming that they represent "recurrent" phenomena. Moreover the hapax legomena occupy much computer memory, and therefore only bigrams that occur more than once were extracted. Based on initial MWE experiments we decided to weed out all n-grams that occur less than 5 times. Only the results of applying AMs on this subset of n-grams are reported here.

Second, statistical AMs generally capture sequences of words that co-occur more often than would be expected by chance. Among the categories that are ranked high are multiword proper nouns and titles, which, for our current purposes are not seen as relevant. For instance, the upper-case version of the adjective form *Røde* 'Red' predominantly occurs in the sequence *Røde Kors* 'Red Cross'. We therefore needed to perform processing operations to remove or collapse n-grams according to their relevance.

Third, the preliminary experiments revealed another category of unwanted high-ranked items, namely string sequences that are not words but the result of formatting errors and boilerplates that should have been removed in the initial text processing. For example we found [*reC sultat*] as an erroneous division of the noun *resultat* 'result'. Although this erroneous collocation was not frequent overall, the AMs that favour low-frequent associations ranked such examples high. Thus, the n-gram lists to be used as input for the AMs were filtered according a set of rules, as described in points 1-5, in which each string separated by a space in an n-gram is referred to as a *token*):

1. *Only alphanumeric characters*: remove n-grams with non-alphanumeric characters ([*i 1998*] 'in 1998', [*, og*] ', and').

2. *Remove proper nouns*: Proper nouns are identified simply on the basis of case sensitivity and are discarded from the main file. An n-gram is assumed to be a proper noun if each token is capitalised (*Røde Kors*) or all letters are written in upper-case (*RØDE KORS, SAS Norge*), although we define some exceptions from this general rule (see main text below).

3. *Merging the frequencies of pairs of n-grams*: We merge the frequencies of n-grams pairs that are only distinguished by whether the first token is capitalised or not (*[Det er]* 'It is'–*[det er]* 'it is'), and we also merge the frequencies of n-gram pairs where an assumed non-proper noun has a capitalised/upper-case variant (see main text below).

4. *Remove erroneous word segmentations, when possible*: erroneous word segmentations such as *reC sultat* 'result' are removed by discarding all n-grams where the initial character is lower-case and one or more succeeding characters are upper-case.

5. *Frequency*: Of the remaining n-grams after filtering, remove n-grams that occur less than 5 times (a heuristic threshold).

It may be added that some n-grams were recorded both in upper-case (all-caps or capitalized) and lower-case. For instance, the proper noun *Røde Kors* had an appellative lower-case variant *røde kors*, whereas the adverbial *stort sett* 'generally' and the anglicism *easy listening* had upper-case variants (*STORT SETT* and *Easy Listening*, respectively).

In such cases, the most frequently used variant was taken to be the most probable one. If the lower-case variant is the most probable, we tentatively merged the frequencies of the case variants and then discarded the upper-case. In the opposite case (cf. *røde kors–Røde Kors*), both variants were filtered from our data file, since they were assumed to represent a proper noun. We also discarded n-grams where the latter token(s), but not the first one, are capitalised (*i Europa* 'in Europe'). Possible abbreviations–characterised by a period after each token (*dr. med.*)–were moved from the data file and moved to a separate list of proper noun candidates, for the benefit of future users of such a resource.

The size of the n-gram lists before and after filtering is shown in Table 1. The figures show the number of unique bigrams and trigrams, respectively, and does not show how many times each n-gram occurred in the corpus. The top row shows the total number of unique n-grams in the unfiltered list of n-grams occurring more than once. Rows two and three show the number of proper nouns and other items that were filtered out. The fourth row shows the number of remaining words after the filtering.

It is this data set that will be used for the statistical ranking of MWE candidates in this study.

As can be seen, the number of token combinations is by far higher for trigrams than for bigrams. The high number of discarded proper nouns illustrates their preponderance in newspaper language.

Table 1: *The number of n-grams (type level) before and after filtering proper nouns and other unwanted items.*

	Bigrams	**Trigrams**
n-grams (where $n > 1$) before filtering	$24,512,294$	$62,040,589$
discarded proper nouns	$1,102,927$	$347,260$
other discarded n-grams	$8,476,622$	$29,063,290$
n-grams after filtering (where $n \geq 5$)	$4,945930$	$8,542,891$

It is worth pointing out that when weeding out elements *a priori*, as we have done here, this has consequences for the statistical basis for our study. We nevertheless believe that such a filtering of names and other unwanted material is necessary in order to enhance the extraction of relevant candidates for MWEs. The NSP package offers two alternatives: either include unwanted items while counting frequencies, but ignore them in the final ranked lists of MWE candidates; alternatively–as we do–discard unwanted n-grams prior to frequency counts.

3.3 Contingency tables

Having filtered the list of n-grams, each remaining n-gram was associated to a set of observed frequencies and estimated frequencies. These frequency counts were conveniently plotted into a *contingency table* (henceforth CT). We will first consider the bigram CT and then the trigram CT.

3.3.1 Bigram Contingency Tables

Each observed frequency in a bigram CT is represented as a numeric value, o_{ij}, where i and j represent the presence (value=1) or absence (value=0) of each token in the n-gram. For instance, the value o_{11} tells how many times the bigram sequence [a b] occurred in the corpus, and o_{12} indicates how many times a occurred in n-grams without being followed by b. The marginal frequencies are the sums of each line: o_{1p} is the sum of $o_{11} + o_{12}$, and so on. Marginal frequencies are sometimes referred to as R1, R2, C1 and C2, respectively (cf. Table 2). o_{pp} is the sum of the marginal frequencies (sometimes the notation N is used), and thus sums up the total number of n-grams.

Table 2: *Contingency table (CT) for a bigram [a b]: observed frequencies*

	b	**not b**	
a	o_{11}	o_{12}	o_{1p} **(R1)**
not a	o_{21}	o_{22}	o_{2p} **(R2)**
	o_{p1} **(C1)**	o_{2p} **(C2)**	o_{pp} **(N)**

Each CT of observed frequencies has a parallel table of *estimated frequencies*, which provides the *expected* frequencies, given the null hypothesis that there is no

association between the words in the given n-gram. If there is no association between a and b we would expect that the chance of seeing them together is proportional to the frequency of each of the tokens individually. Thus, if one or both tokens are highly frequent, then we may expect a high frequency for their estimated e_{11}, too.

Table 3: *Contingency table for a bigram [a b]: estimated frequencies*

	b	**not b**
a	$e_{11} = \frac{R1C1}{N}$	$e_{12} = \frac{R1C2}{N}$
not a	$e_{21} = \frac{R2C1}{N}$	$e_{22} = \frac{R2C2}{N}$

Table 4 contains a number of illustrative examples of bigrams and trigrams that we consider relevant for the purposes of lexicography or technical terminology, because they represent genuine MWEs, as well as some n-grams that are not relevant but nevertheless included in the table for comparison (bottom section of the table). Each n-gram is listed with the observed value of the full n-gram, and the number of times each word of the n-gram occurred in contexts other than the n-gram (bigrams: o_{11}, o_{12}, o_{21}; trigrams: o_{111}, o_{122}, o_{212} and o_{221}). Each expression in the table is given an English gloss in parenthesis. Due to their foreign origin, we consider multiword anglicisms as a separate category, which may subsume English-based technical terms (*due dilligence*) or other lexicalised items.

Intuitively bigrams such as the phrasal anglicism *banana split* (a dessert), the linguistic term *perfektum partisipp* 'perfect participle' and the phrasal idiom *på fersken* 'red-handed, in the act' are more strongly associated than highly frequent formulaic multiword sequences such as *det er* 'it is' or less frequent syntactic phrases such as *på bananen* 'on the banana'. In other words, for our purposes a 'good' AM measure is one that ranks highly the members of the first-mentioned categories (seen in the upper five sections of Table 4), since they represent true MWEs.

For instance, the phrase [*på fersken*] may happen to instantiate a non-lexicalised regular prepositional phrase in a situated context (*tenke på fersken* 'think of a peach'), but the bigram is usually instantiations of an idiomatic expression that means 'red-handed, in the act'. In this meaning, one cannot insert anything between the two elements, nor can the noun be inflected and still convey the meaning of the idiom. By contrast, the bigram [*på bananen*] is intuitively just a regular prepositional phrase (cf. *tenke på bananen* 'think of the banana'). This is clear because the noun can be inflected and still retain the same lexical meaning, and a modifier may be inserted between the two bigram words (*på en uhyre liten banan* 'on an incredibly small banana').

As Table 4 shows, the bigram frequencies appear to reflect this intuition. Considering the anglicism and the linguistic term, the bigrams are not particularly frequent, but both tokens in both bigrams nonetheless occur more often in this particular collocation than in any other collocations. Comparing the observed frequencies of [*på fersken*] (idiom) and [*på bananen*] (prepositional phrase), we note that the idiomatic *på fersken* is by far more frequent than the non-lexicalised (1133 against 5, respectively).

3.3.2 Trigram Contingency Tables

Trigram contingency tables are more complicated since they are three-dimensional. To generate the trigram contingency tables, we re-implemented in Common LISP the NSP package. Given a trigram [a b c], where 1 denotes the presence of a token and 0 the

Table 4: *Examples of bigram and trigram multiword expressions (MWEs). Observed frequencies (o11) and the number of times each token occurs in this position.*

n-gram	o_{11}/o_{111}	o_{12}/o_{122}	o_{21}/o_{212}	o_{221}
Anglicisms:				
banana split (a dessert)	15	0	16	
due dilligence (appraisal of a business)	35	805	0	
practical jokes (a prank)	67	33	0	
corned beef (brine-cured beef)	78	0	16	
easy listening (music style)	85	129	13	
get a life (multiword discourse marker)	29	258	6298	403
Other foreign expressions:				
gefundenes fressen (Ger. 'sensational news')	47	0	0	
in vitro (Lat. 'in a test tube')	66	14859	0	
quod erat demonstrandum (Lat. 'that which was to be demonstrated')	12	0	0	0
Terms:				
anaerob terskel 'anaerobic threshold' (medicine)	6	9	1429	
perfektum partisipp 'perfect participle' (linguistics)	8	0	4	
ulcerøs colitt 'imflammation' (medicine)	24	18	0	
notarius publicus 'public secretary' (law)	69	4	0	
per capita 'per head' (statistics)	96	43651	3	
trojansk hest 'virus, Trojan horse'(computing)	140	4	9546	
acute respiratory infection	5	0	5	0
Idioms:				
tilslørte bondepiker (a dessert)	50	68	4	
tenners gnissel 'despair'	106	23	5	
navns nevnelse 'by name'	156	16	2	
hellig ku 'holy cow'	277	5362	1593	
på fersken 'in the act'	1133	9051906	266	
i tide 'on time'	9527	16400611	18348	
på tide 'about time'	18199	9034840	9676	
rusk og rask 'bric-a-brac'	190	329	6650950	15421
katta i sekken 'buy a pig in a poke'	255	123	11940228	916
på kant med 'in opposition to'	1741	5985152	1299	5694890
rett og slett 'simply'	48827	161950	6598278	18904
Complex grammatical expressions (syntactically idiomatic):				
på forhånd 'beforehand'	38769	9014270	22	
stort sett 'generally'	52403	159845	216228	
blant annet 'among other things'	257627	338086	208987	
på en måte 'in a way'	16600	5612405	7619609	94091
i motsetning til 'as opposed to'	26154	11463279	1304	7848663
i forhold til 'compared to, in relation to'	134768	11462650	55016	7811917
på grunn av 'because of, due to'	178525	5655710	146367	7170521
Multiword formulaic sequences that are *not* multiword expressions:				
på bananen 'on the banana'	5	9053034	74	
millioner kroner 'million NOK'	411623	245778	250644	
det er 'it is'	1170464	9424660	7343062	
jeg vil ikke 'I don't want to'	4738	1974296	1428425	3640209
grunn til å 'reason to'	85136	203451	5087841	7028596

absence of a token, the observed frequencies are shown in Table 5. The marginal and estimated frequencies are listed in Tables 6 and 7, respectively. As with bigrams, p means that the count is not conditioned by what appears in the position p.

Table 5: *Contingency table for a trigram [a b c]: observed frequencies*

		c	**not c**
a	**b**	o_{111}	o_{112}
a	**not b**	o_{121}	o_{122}
not a	**b**	o_{211}	o_{212}
not a	**not b**	o_{221}	o_{222}

Table 6: *Marginal frequencies: trigrams (three-dimensional)*

o_{1pp}, o_{p1p}, o_{pp1}	= the number of trigrams where the first token is a, b and c, respectively.
o_{2pp}, o_{p2p}, o_{pp2}	= the number of trigrams where the first token is **not** a, b and c, respectively.
o_{11p}, o_{1p1}, o_{p11}	= the number of trigrams where the first and second token; first and third token and second and third token are (respectively) a—b; a—c and b—c.
(the marginal frequencies o_{22p}, o_{2p2}, o_{p22} are not needed for any of the AMs)	
o_{ppp}	= the total number of occurrences of all trigrams.

Table 7: *Estimated frequencies: trigrams (three-dimensional). The o_{ppp} corresponds to the N value in the bigram contingency table, i.e. the total number of trigram occurrences.*

e_{111}	$= \frac{o_{1pp} * o_{p1p} * o_{pp1}}{(o_{ppp})^2}$	e_{222}	$= \frac{o_{2pp} * o_{p2p} * o_{pp2}}{(o_{ppp})^2}$		
e_{112}	$= \frac{o_{1pp} * o_{p1p} * o_{pp2}}{(o_{ppp})^2}$	e_{121}	$= \frac{o_{1pp} * o_{p2p} * o_{pp1}}{(o_{ppp})^2}$	e_{211}	$= \frac{o_{2pp} * o_{p1p} * o_{pp1}}{(o_{ppp})^2}$
e_{122}	$= \frac{o_{1pp} * o_{p2p} * o_{pp2}}{(o_{ppp})^2}$	e_{212}	$= \frac{o_{2pp} * o_{p1p} * o_{pp2}}{(o_{ppp})^2}$	e_{221}	$= \frac{o_{2pp} * o_{p2p} * o_{pp1}}{(o_{ppp})^2}$

3.4 Bigram Association Measures

For bigrams, our experiments mainly follow the definitions found in the UCS Toolkit[3]. Some of the statistical tests have several variants (for instance with and without statistical correction measures). In the following we provide all formulae, as implemented

[3]URL: http://www.collocations.de. Last verified July 8, 2010.

in Common LISP by the authors, and motivate the various choices that have been made. For each formula, an abbreviation is given in parenthesis. For a more thorough, theoretical discussion of the different measures, we refer to Evert (2004) and Banerjee and Pedersen (2003).

Pearson's chi-squared homogeneity corrected ($X^2_{h,c}$) Pearson's chi squared test (X^2) measures the difference between the observed values and the estimated values, i.e. those values one would expect if the tokens in the bigram were independent. The higher the score, the more strongly they are associated. It was chosen to implement a special version of the X^2 formula, namely the *chi-squared homogeneity corrected*, as this version, according to Evert (2004), is often used in applications.

$$X^2_{h,c} = \frac{N(|o_{11}o_{22} - o_{12}o_{21}| - \frac{N}{2})^2}{R1R2C1C2} \tag{1}$$

Log-likelihood ratio (LL) The log likelihood ratio measures the difference between the observed values and the expected values. It is the sum of the ratio of the observed and expected values. According to Evert (2004), the log-likelihood ratio is expected to perform better than the Pearson's chi-squared for lexical word collocations, since lexical (as opposed to grammatical) words tend to have a low o_{11} in comparison to a generally high N value. The standard formula is:

$$\text{Log-likelihood ratio} = 2 * \sum_{ij} o_{ij} log(\frac{o_{ij}}{e_{ij}}) \tag{2}$$

Logarithmic Odds Ratio$_{disc}$ (OR) The Logarithmic odds ratio returns the proportion between how many times the tokens in an n-gram co-occur and how many times each of the tokens occur individually. Since the logarithm is undefined if any of the numbers in the denominator (o_{21} or o_{22}) are zero, Evert (2004) proposes a 'discounting' technique by which 0.5 is added to every observed value (written in the formula below as $_{disc}$). As an alternative solution, Banerjee and Pedersen (2003) propose a "smooth-by-one" technique by which only zero values are replaced by 1. We chose to apply this discounting technique because Evert (2004) claims that since it produces slightly higher figures, it might be beneficial for low-frequency bigrams.

$$\text{Odds ratio}_{disc} = log\frac{(o_{11} + 0.5)(o_{22} + 0.5)}{(o_{12} + 0.5)(o_{21} + 0.5)} \tag{3}$$

Z-score (regular and corrected) (Z-s, Z-s$_{corr}$) The z-score is a relatively simple measure which computes a probability score for the observed frequency in comparison to the expected value. According to Evert (2004) it can be used to find "significant word pairs".

$$\text{z-score} = \frac{o_{11} - e_{11}}{\sqrt{e_{11}}} \tag{4}$$

A problem of the z-score measure is its use of the continuous normal distribution to approximate a discrete (binomial) distribution. According to Evert

(2004), *Yates' continuity correction* improves this approximation by adjusting the observed frequencies according to the following rules, which we also implemented:

$$\text{z-score}_{corrected} = \frac{o_{11} - e_{11}}{\sqrt{e_{11}}} \begin{cases} o_{ij} - 0.5 & \text{if } o_{ij} > e_{ij}, \\ o_{ij} + 0.5 & \text{if } o_{ij} < e_{ij}. \end{cases} \tag{5}$$

Yates' continuity correction can be applied to all cells of the contingency table, although only o_{11} is relevant for the z-score measure.

T-score (T-s) Church et al. (1991) use the co-called Student's t-test as an alternative to the z-test. The t-score determines whether the association between two words is non-random, by computing the quotient of the observed and estimated value divided by the square root of the observed frequency value. As opposed to the z-test, the variance (the denominator) is estimated directly from data, and not through the estimated frequency. According to Evert (2004), the t-test is theoretically dubious for collocations and produces extremely conservative values.

$$\text{z-score} = \frac{o_{11} - e_{11}}{\sqrt{o_{11}}} \tag{6}$$

Pointwise Mutual Information (PMI) In general, Mutual Information ranks n-grams according to the principle of comparing the frequency of the MWE candidate to the frequency of the components of the MWE. This is expressed in the formula below in that the o_{11} value tells how often the sequence occurs, whereas the estimated value of the sequence is based on how often each of the two words in the sequence occur independently. This measure is biased towards low-frequency n-grams, i,e, n-grams where o_{11} is low (Evert 2004, Manning and Schütze 1999). The pointwise MI is calculated as follows:

$$\text{Pointwise MI} = \log(\frac{o_{11}}{e_{11}}) \tag{7}$$

Dice coefficient, Jaccard coefficient (D, J) The two measures Dice coefficient and Jaccard coefficient are often used in information retrieval technology and are easily calculated. The dice coefficient considers the frequency of a and b occurring together and their individual frequencies.

$$\text{Dice} = \frac{2o_{11}}{R1 + C1} \tag{8}$$

$$\text{Jaccard} = \frac{o_{11}}{o_{11} + o_{12} + o_{21}} \tag{9}$$

3.5 Trigram Association Measures

The calculations are not straightforward when treating sequences that are longer than bigrams. Banerjee and Pedersen (2003) list the following tests as suitable for trigrams: Log-likelihood ratio, Mutual Information, Pointwise Mutual Information and Poisson Stirling. The trigram measures below are based on the perl code of the NSP package.

Log-likelihood (LL) As with bigrams, the log-likelihood method measures the tendency for words to co-occur by considering the deviation between observed and expected values for each observed value in (Table 5):

$$\text{Log-likelihood} = 2\sum_{ijk} o_{ijk} log \frac{o_{ijk}}{e_{ijk}} \tag{10}$$

Poisson-Stirling (PS) The Poisson-Stirling measure is computed as:

$$\text{Poisson-Stirling} = o_{111}(log(o_{111}) - log(e_{111} - 1) \tag{11}$$

Pointwise Mutual Information (PMI) As with bigrams, Pointwise Mutual Information measures the association strength by considering the frequency of the MWE candidate in comparison to the frequency of the components of the expression.

$$\text{Pointwise Mutual Information} = log_2(\frac{o_{111}}{e_{111}}) \tag{12}$$

True Mutual Information (TMI) True Mutual Information measures the extent to which the observed frequencies differ from the expected frequencies, by computing the weighted average of the pointwise mutual informations for all the observed and expected value pairs.

$$\text{True Mutual Information} = \sum_{ijk}(\frac{o_{ijk}}{N})(log_2 \frac{o_{ijk}}{e_{ijk}}) \tag{13}$$

4 Results

4.1 Bigrams

In what follows, we will propose a grouping of the nine tested AMs for bigrams, according to their ability to rank bigrams that we perceive to be of high lexicographical or terminological relevance. First we compare how the different AMs rank the example bigrams in Table 4 using *Spearman's rank correlation coefficient*. Then we present an evaluation based on a manual inspection of the 500 most highly ranked items for each AM.

Given a set of n-grams which has been ranked by two different AMs, Banerjee and Pedersen (2003) suggest to compare the rankings using *Spearman's rank correlation coefficient*. The formula is given in Equation (14), in which r is the rank, n is the total number of n-grams considered and D_i is the difference between the rank assigned to an n-gram i by two different AMs (if an n-gram i was ranked second by the first AM and fourth by the second AM, then D_i = 4-2 = 2). An r value close to 1 indicates that the two measures rank n-grams in the same order, -1 that the two rankings are exactly opposite to each other, and the value 0 indicates that they are not related.

$$\text{r} = 1 - \frac{6\sum_{i}^{n} D_i^2}{n(n^2-1)} \tag{14}$$

It was beyond the scope of the present experiments to run comparisons on the entire material (almost 5 million bigrams); instead a pairwise comparison of the AMs was run

based on the list of example bigrams in Table 4. As will be shown below, the results from this sample-based comparison seem to concord with the general findings when considering the top-ranked bigrams for each AM. With nine bigram measures and 36 combinations, we only present the main conclusions from the Spearman's comparison.

Ranking all 36 r scores from 1 to -1, a clear "upper ten" of AM comparisons is singled out with values in the range of 0.99–0.77. For instance, the highest r value was found when comparing the Z-score and Z-score$_{corrected}$ (r=0.9993162), which indicates that they are very similar in how they rank the set of example bigrams. The next (11^{th}) r score is as low as 0.36. Among the top ten most similar AMs, some of them have direct or indirect links (cf. the way the r score close to 1 linked Z-score and Z-score$_{corrected}$) whereas others are never linked to each other. Thus, three clusters of AMs with a similar behaviour are suggested through Spearman's, as itemized below. Based on the small example set of bigrams, the Chi-squared$_{corr}$ measure is "the odd one out" and is not found to be similar to any other measure. Its highest score is 0.36 (which was in comparison to Dice), but being closer to 0 this value rather indicates unrelatedness.

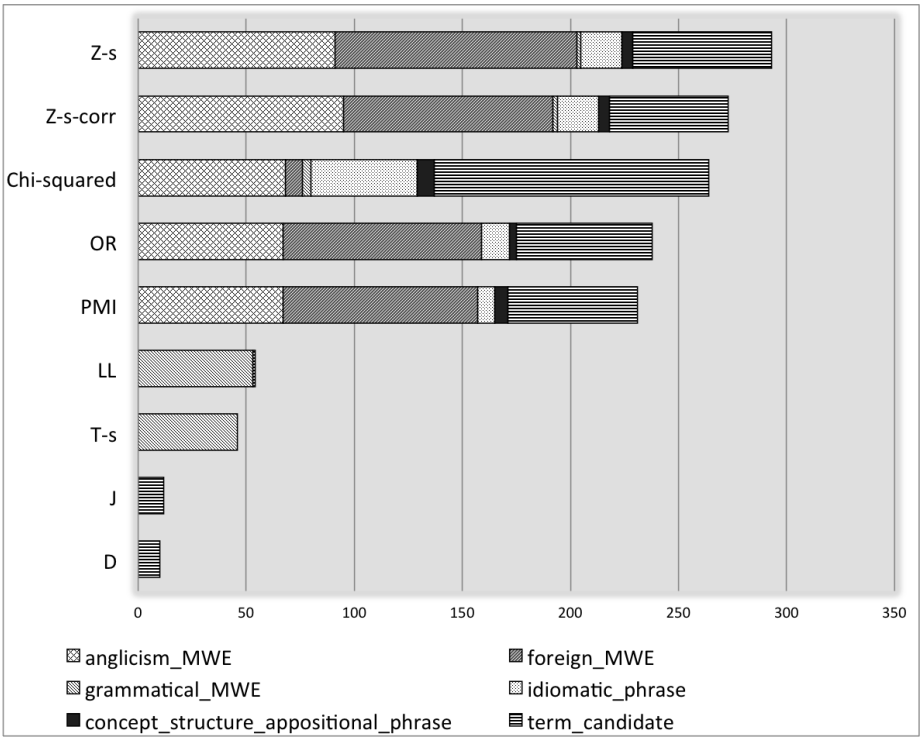
- Z-score, Z-score$_{corr}$, PMI and Odds Ratio
- Log-Likelihood and T-score
- Jaccard and Dice
- Chi-squared$_{corr}$

In order to analyse the behaviour of each association measure in more detail, all the 500 top-ranked MWEs for each AM were classified manually according to the following set of categories: *anglicism MWE*, *foreign MWE*, *grammatical MWE*, *idiomatic phrase*, *concept structure appositional phrase*, *term candidate*. Thus we may analyse the kinds of MWEs singled out by each AM. The findings for the bigram AMs are summed up in Figure 2 (which does not include n-grams that were not classified as belonging to any of the mentioned catgories).

The manual analysis indicates that there are major differences between the different measures in their ability to retrieve bigrams that are considered terminologically of lexicographically relevant. Two AMs, Jaccard and Dice, are only able to retrieve a very limited number of lexicographically or terminologically relevant items, amounting to a mere 2 per cent of the manually inspected ranked n-grams, including *langvarig konjunkturoppgang* 'sustained cyclical expansion' and *maritime industri* 'maritime manufacturing'. Two measures, T-score and Log Likelihood, are particularly suited for detecting grammatical MWEs and not any other MWE types. The retrieved items include multiword adverbials and prepositions such as *for eksempel* 'for example', *i tillegg* 'in addition', *etter hvert* 'gradually' and *blant annet* 'among others' as well as one phrasal verb, *regne(r) med* 'take into account'. Their respective 10.8 and 9.2 per cent must be considered a high proportion of grammatical MWEs, given that this is a closed category, which generally can be expected to have fewer members than open categories such as nouns, which is where most terms would be included.

The remaining five AMs are all relatively successful in retrieving lexically and terminologically relevant items, ranging from 46.2 (PMI) to 58.6 per cent (Z-score regular). One of these measures, Pearson's chi square, is particularly able to pick out term candidates, including *alternative energikilder* 'alternative energy sources' and *blokkerende mindretall* 'blocking minority' as well as appositional noun phrases of the type

Figure 1: *A manual classification of the* 500 *top ranked* MWEs *for each bigram Association Measure (AM)*



tungmetallet kadmium 'the heavy metal cadmium', which we also consider to be relevant for term extraction purposes since it gives not only a term but also its superordinate concept. The other four measures are to a lesser degree able to identify domestically based term candidates but are better than Pearson's at extracting MWEs (including terms) of foreign or English origin, such as *consumer confidence, joint ventures, annus horribilis* and *garam masala*.

The similarities and differences between the AMs might be better understood by considering the frequency patterns that tend to be ranked high with the different AMs. Table 8 shows the top ten ranked bigrams of four AMs that seem well-suited for the identification of extracting MWEs (including terms) of foreign or English origin, namely Pointwise Mutual Information, Odds Ratio$_{discr}$, Z-score and Z-score$_{corr}$. Recall that we set a lower threshold of 5 tokens; hence no bigrams occurring fewer times are included in the tables. These four measures favour bigrams where the o_{11} is low, and where the individual words of the bigrams only occur in the context of this particular bigram, as can be seen from the contingency figures of almost all n-grams in Table 8. (It may be noted that if several bigrams have the same statistical score in the leftmost column, they are simply ranked alphabetically, for instance this applies to the PMI and the Z-score measure).

We note that the n-grams typically consist of words that only occur in this particular expression, which we would expect to be the case for xenomorphic MWEs. Several of

Table 8: *Top ten ranked bigrams: 4 AMs favouring low-frequent associations*

Pointwise Mutual Information (PMI)				
18.409534	vilkårsett skattefritaking	5	0	0
18.409534	varannan damernas	5	0	0
18.409534	unio mystica	5	0	0
18.409534	twam asi	5	0	0
18.409534	tussilago farfara	5	0	0
18.409534	suvas bohciidit	5	0	0
18.409534	skrimmi nimmi	5	0	0
18.409534	rollon rolloff	5	0	0
18.409534	rødøret terrapin	5	0	0
18.409534	radiær keratotomi	5	0	0
Odds Ratio$_{discr}$ (OR)				
25.34685	chop suey	51	0	0
25.265997	gefundenes fressen	47	0	0
24.916813	nobis pacem	33	0	0
24.855255	jaska beana	31	0	0
24.789658	lorem lipsum	29	0	0
24.75517	lipsum lorem	28	0	0
24.75517	hæ hæ	28	0	0
24.603941	retinitis pigmentosa	24	0	0
24.518784	eines fahrenden	22	0	0
24.425692	haemophilus influenzae	20	0	0
Z-score (Z-s)				
22236.418	byssan lull	15	0	0
22236.416	yada yada	6	0	0
22236.416	whistle blowers	7	0	0
22236.416	visibility corp	6	0	0
22236.416	vilkårsett skattefritaking	5	0	0
22236.416	varannan damernas	5	0	0
22236.416	utsletti respateksbord	8	0	0
22236.416	unio mystica	5	0	0
22236.416	uisge beatha	11	0	0
22236.416	twam asi	5	0	0
Z-score$_{corr}$ (Z-s$_{corr}$)				
22018.41	chop suey	51	0	0
21999.857	gefundenes fressen	47	0	0
21899.498	nobis pacem	33	0	0
21877.762	jaska beana	31	0	0
21862.246	hokus pokus	263	6	2
21853.03	lorem lipsum	29	0	0
21839.336	lipsum lorem	28	0	0
21839.336	hæ hæ	28	0	0
21773.154	retinitis pigmentosa	24	0	0
21744.455	unit linked	111	2	2

Table 9: *Top ten ranked bigrams for Pearson's chi-square*

Chi-squared$_{h,corr}$ ($X^2_{h,c}$)				
9996430.0	ss tomatpure	220	5361	207
9996173.0	knus hvitløken	22	63	247
9983254.0	våpentekniske korps	88	62	2440
9983179.0	obstruktive lungesykdommer	10	0	437
9980703.0	all sannsynlighet	5164	143091	3739
9980339.0	red anm	87	1864	103
9980115.0	buddhistiske munker	112	400	1091
9967788.0	ferjefritt veisamband	9	55	47
9966970.0	nordatlantiske fiskeriorganisasjonen	22	446	27
9962959.0	tissue engineering	5	0	196

the English-based n-grams that are highly ranked with these measures indeed represent technical terminology from different professional domains. This includes anglicisms such as *rollon rolloff* 'ships designed to carry wheeled cargo', *unit linked* 'a type of (insurance) fund' and *whistle blower* 'person who alerts about a wrongdoing', which are linked to shipping and the economic-administrative domains.

By contrast, multiword anglicisms that are part of the general vocabulary does not achieve a particularly high rank according to these four measures, due to the higher overall frequency of the n-grams and their components, of which *body lotion* ($9 - 78 - 12$), *dark horse* ($153 - 28 - 22$), *sudden death* ($1158 - 5 - 261$) and *all right* ($738 - 147517 - 370$)) are notable examples.

Furthermore, Latin expressions tend to dominate in these lists, and the most highly ranked items are drawn from technical terminology rather than general language. In the top ten lists we find for instance the biological term *tussilago farfara*, which is the Latin name of the flower coltsfoot, and the medical term *Retinitis pigmentosa* 'an eye disease'. There are also terms of Norwegian origin among these highly ranked n-grams, such as the biological *rødøret terrapin* 'red-eared terrapin' (a turtle) and the medical *radiær keratotomi* 'eye surgery'. Incidentally, these lists also highlight the importance of detecting and investigating longer n-grams before shorter ones. For instance, *twam asi* is part of the Sanskrit sentence *tat tvam asi* 'that thou art', while the bigram *jaska beana* is part of the Sami sentence (which is also the title of a song) *oro jaska beana* 'be quiet, dog'.

Pearson's chi-square could have been grouped with those in Table 8, but we list it alone (Table 9) since it is particularly well-suited to single out domestically based term candidates. The vast majority of highly ranked n-grams with this are straightforward grammatical phrases with rather low overall frequencies, such as *knus hvitløken* 'crush the garlic' and *buddhistiske munker* 'Buddhist monks'. However, we do find the odd n-gram of terminological relevance, namely the medical term *obstruktive lungesykdommer* 'obstructive lung diseases' and the biology term *tissue engineering*.

The next category of AM measures, in Table 10, include AMs that seem to favour grammatical MWEs, namely the Log-likelihood ratio and the T-score. These lists predominantly consist of highly frequent formulaic sequences of lexical and function words with a low relevance to lexicograpical or terminological purposes. Virtually all of them consist of at least one function word, e.g. prepositions (*på* 'on, at', *til* 'to, for'), the infinitive marker (*å* 'to'), subjunctions (*at* 'that', *som* 'which'), etc.

According to Evert (2004), the Log-Likelihood ratio works better when bigrams containing grammatical words are ignored. However, given our strict corpus-driven method, we did not exclude any items on the basis of syntactic category. If we isolate only the Log-likelihood bigrams where both components are lexical words, i.e. verbs, nouns, adjectives or adverbs, or combinations thereof, the top-ranked items seem to have a varying degree of lexicalised status. Some, like *millioner kroner* 'Million NOK', are clearly not lexicalised MWEs, while others, like *administrerende direktør* 'managing director', can be seen to have terminological value in the economic-administrative domains. In other words, we cannot rule out the relevance of these three AMs for the identification of multiword lexemes and terminology.

However, and importantly, these AMs do appear to be capable of identifying highly frequent MWEs that are grammatically significant, namely multiword prepositions. An example of this is *blant annet* 'among other things', which is ranked among the top ten in two of the AMs in Table 10. In passing, we ought to mention that it may be fruitful in a future study to test all AMs on lists of n-grams that solely contain lexical words, and lists of n-grams where at least one of the words is a lexical word.

Table 10: *Top ten ranked bigrams: 2 AMs favouring grammatical MWEs*

Log-likelihood				
6512751.0	til å	1965359	6805145	8333315
6221586.0	for å	1904212	6759186	8394462
4913251.0	millioner kroner	411623	245778	250644
3253937.2	å få	642768	9800294	634097
3172030.5	at det	1233788	6813598	8599093
2953856.0	blant annet	257627	338086	208987
2943812.0	har vært	544440	7822106	575711
2596295.2	det er	1170464	9424660	7343062
2009919.6	at han	619027	7428359	2766184
1954526.4	løpet av	264615	21286	9123515
T-score				
1271.6094	til å	1965359	6805145	8333315
1249.1697	for å	1904212	6759186	8394462
966.686	at det	1233788	6813598	8599093
913.2606	det er	1170464	9424660	7343062
850.10693	er det	1205511	12478853	8627370
768.0912	å få	642768	9800294	634097
716.75714	at han	619027	7428359	2766184
712.1745	har vært	544440	7822106	575711
709.274	at de	671512	7375874	4876299
686.3624	som er	803092	10116108	7710434

Table 11 shows the top ten bigrams for the final two AMs, Dice and Jaccard, which only seem to identify very few terminologically relevant items. By and large they seem to extract a bit of a 'word salad' not particularly fitted for any of our needs. Characteristic for them is that the high-ranked bigrams are mainly composed of lexical, as opposed to grammatical words.

Table 12 shows how the example bigrams of Table 4 are ranked in relation to each other with each of the nine AMs. The rankings seem to confirm what we saw when considering the top ten lists for each individual AM, namely that the first category, consisting of Pointwise Mutual Information (PMI), Odds Ration(OR), Z-score and Z-s_{corr}, is best at retrieving low-frequent items such as technical terms, domain-specific multiword anglicisms and foreign expressions in general.

As for common Norwegian idioms and fixed expressions, such as *tenners gnissel* 'gnashing of teeth' and *hellig ku* 'holy cow', it does not seem that any of the measures are particularly good at picking them out. Dice and Jaccard are the only ones to rank an idiom highest among the selected bigrams (*i tide, på tide*, respectively), but in return they have their low-ranked items in the same category (*på fersken, i tide*, respectively).

4.2 Trigrams

As with bigrams, we compared how the different AMs rank the example trigrams in Table 4 using *Spearman's rank correlation coefficient*; then the 500 most highly ranked items for each AM were manually inspected. The Spearman's comparison did not shed much light on the relation between the four AMs being compared, maybe due to the low number of trigram examples in the comparison. Based on the manual evaluation, on the other hand, there seem to be striking differences between the four different trigam AMs in their ability to pick out word sequences that are of terminological or lexicographical relevance (Table 2).

Based on the analysis of the 500 top-ranked items, two of the AMs, Log-Likelihood (LL) and True Mutual Information (TMI), were unable to rank highly any relevant

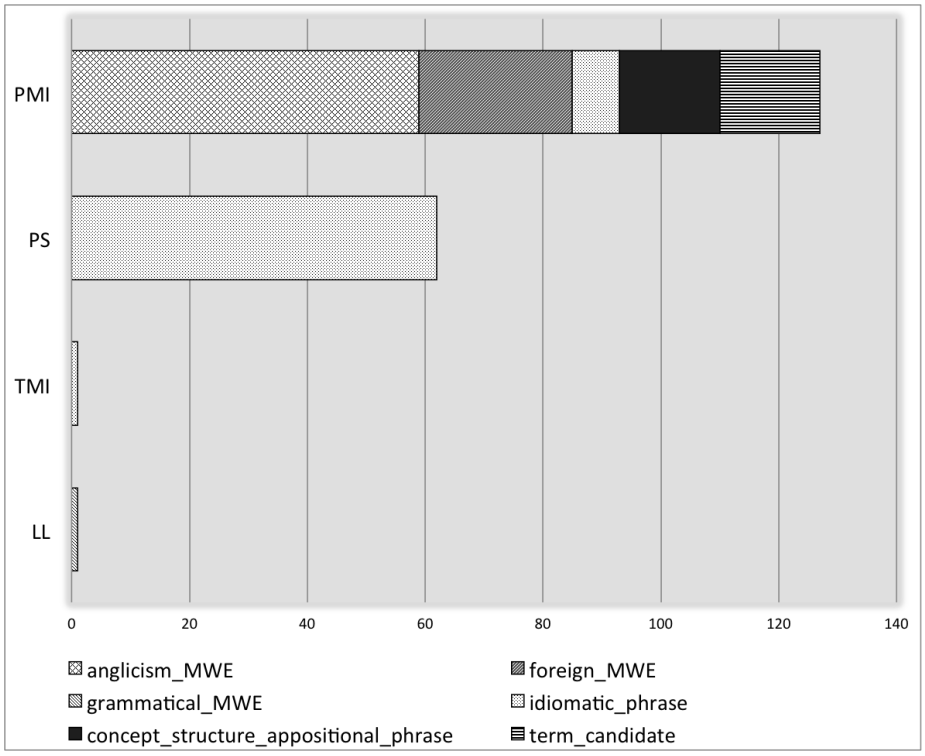
Table 11: *Top ten ranked bigrams: 2 final AMs, not well-suited for MWEs*

Dice (D)				
0.001	engasjerende valgkamp	7	1225	12761
0.001	folk sa	251	368926	132572
0.001	froskemenn lette	11	437	21541
0.001	innkalle statsråden	9	2764	15218
0.001	konsernet realiserte	6	11549	439
0.001	manglende gjennomslag	28	41751	14193
0.001	militæret opplyser	18	2744	33220
0.001	overveie konsekvensene	11	1667	20311
0.001	strengere sikkerhetsforskrifter	7	13857	129
0.001	undersøkt påstander	15	11606	18364
Jaccard (J)				
0.001	arrangerte seminar	14	5815	8171
0.001	begår voldelige	11	3945	7044
0.001	fekk raudt	8	7906	86
0.001	flott representant	49	26459	22492
0.001	hjemlige filmmiljø	6	5790	204
0.001	høyerestående planter	7	203	6790
0.001	human kapital	19	1115	17866
0.001	kapitalistisk ånd	5	545	4450
0.001	konfliktene oppstår	21	3928	17051
0.001	lidende dyret	5	868	4127

Table 12: *Rank of selected bigrams with ten different association measures (AMs)*

n-gram	PMI	OR	z-S	z-S_c	LL	PS	t-S	D	J	$X^2_{h,c}$
banana split	2	8	9	9	23	23	23	16	15	21
due dilligence	12	11	15	15	21	21	21	9	8	22
practical jokes	7	4	6	6	16	16	17	19	18	17
corned beef	6	3	4	4	13	14	15	21	20	15
easy listening	11	12	12	12	14	13	14	13	12	25
gefundenes fressen	4	1	1	1	18	18	20	24	23	12
in vitro	15	13	19	19	20	20	18	6	5	4
anaerob terskel	13	16	22	22	25	25	25	4	3	7
perfektum partisipp	1	6	7	7	24	24	24	18	17	18
ulcerøs colitt	3	7	8	8	22	22	22	17	16	19
notarius publicus	5	2	2	2	15	15	16	23	22	13
per capita	17	15	21	21	17	17	13	3	2	6
trojansk hest	14	14	16	16	12	12	11	7	6	1
tilslørte bondepiker	8	10	10	11	19	19	19	14	13	24
tenners gnissel	9	9	5	5	11	11	12	20	19	16
navns nevnelse	10	5	3	3	9	10	10	22	21	14
hellig ku	16	18	18	18	10	9	9	8	7	3
på fersken	22	22	25	25	8	8	8	26	24	10
i tide	24	24	24	24	7	7	7	1	26	9
på tide	23	23	23	23	6	6	6	2	1	8
på forhånd	21	17	20	20	5	5	5	5	4	5
stort sett	20	21	14	14	4	4	4	11	10	20
blant annet	19	20	13	13	2	2	3	12	11	26
på bananen	26	26	26	26	26	26	26	25	25	11
millioner kroner	18	19	11	10	1	1	2	15	14	23
det er	25	25	17	17	3	3	1	10	9	2

Figure 2: *A manual classification of the* 500 *top ranked* MWEs *for each trigram Association Measure*



items, with the exception of one token each, the idiomatic phrase *grøss og gru* 'shiver and horror' (TMI) and the phrasal verb *kommer til å* 'is going to' which counts as a grammatical MWE (LL). To give an indication of the kinds of items ranked highly with these measures, the top ten trigrams with these two measures are listed in Tables 13 (LL) and 14 (TMI).

As we also saw with bigrams, the Log-Likelihood tends to assign a high rank to high-frequency items which include grammatical words and which are formulaic sequences rather than lexicalised phrases (*for å få* 'in order to get', *til å ha* 'to have/having'). According to McInnes (2004), Log-likelihood generally encounters problems with large sample sizes. Since the marginal values will then be very large, the expected values are increased, yielding extremely low observed values compared to the expected values. Paradoxically, this means that very independent and strongly dependent n-grams may end up with the same values. She notices similar characteristics with true mutual information; if this is true this may mean that these AMs are not well-suited for our material. The top-ranked trigrams with TMI are rather reminiscent of the 'word salad' that we saw for Dice and Jaccard with bigrams. To illustrate, the three top trigrams are translated as 'us the necessary confidence', 'researchers that a', 'end the boycott of', respectively.

Based on the manual inspection of the 500 top-ranked items using Poisson-Stirling,

Table 13: *Top ten ranked trigrams: Log-likelihood*

Log-likelihood					
	AM-score	o_{111}	o_{1pp}	o_{p1p}	o_{pp1}
til å få	9680734.0	47806	6391902	9844607	1099442
for å få	9259922.0	204938	6485314	9844607	1099442
til å ha	8226067.0	26364	6391902	9844607	1058478
til å ta	7820512.0	71232	6391902	9844607	505870
for å ha	7761713.0	100968	6485314	9844607	1058478
til å bli	7748718.0	85448	6391902	9844607	922553
til å være	7624434.0	61043	6391902	9844607	1164046
til å gjøre	7609440.0	39591	6391902	9844607	393839
til å i	7546252.0	76	6391902	9844607	15570426
for å ta	7411975.5	35027	6485314	9844607	505870

Table 14: *Top ten ranked trigrams: True Mutual Information*

True Mutal Information					
	AM-score	o_{111}	o_{1pp}	o_{p1p}	o_{pp1}
oss nødvendig selvtillit	$9.999997e-7$	5	329746	83870	7994
forskere at et	$9.999991e-8$	7	14296	6738145	3119515
avslutte boikotten av	$9.999989e-6$	10	6383	1826	7602876
det store endringer	$9.9999845e-5$	69	10219653	478697	31431
politiske strid i	$9.9999794e-5$	11	99326	39182	15570426
om hvorvidt amerikanerne	$9.999977e-5$	11	3391818	12656	12779
er kjent ugyldig	$9.999971e-5$	52	11305572	152137	1956
an å bygge	$9.999971e-4$	156	51684	9844607	91208
vei et lavt	$9.99997e-6$	5	99877	3552803	15550
han kræsjet med	$9.999967e-6$	8	3596157	154	5800181

Table 15: *Top ten ranked trigrams: Poisson-Stirling*

Poisson-Stirling					
	AM-score	o_{111}	o_{1pp}	o_{p1p}	o_{pp1}
i løpet av	1607657.0	263659	11929615	278965	7602876
på grunn av	1110109.4	178525	6090499	327900	7602876
for å få	988856.9	204938	6485314	9844607	1099442
at det er	897022.25	259845	7064920	9035186	5503500
når det gjelder	848167.8	102058	783030	9035186	151384
i forbindelse med	845152.0	133840	11929615	149389	5800181
først og fremst	828941.2	81313	236761	6652225	82358
i forhold til	746500.1	134768	11929615	227855	8315633
etter å ha	620644.8	105565	1209267	9844607	1058478
ferd med å	589744.1	85702	88570	4773377	8883188

this measure is highly capable of picking out one specific type, namely grammatical MWEs, as 12.4 per cent of the inspected trigrams were of this category, and no other categories were represented. Its top ten trigrams are given in Table 15 to give an idea of the kinds of frequencies that yield a high rank with this measure. Grammatical MWEs are exemplified by complex prepositions like *i løpet av* 'in the course of' and *i forbindelse med* 'in connection with', and adverbs like *først og fremst* 'first and foremost', etc.

Table 16: *Top ten ranked trigrams: Pointwise Mutal Information*

Pointwise Mutal Information					
	AM-score	o_{111}	o_{1pp}	o_{p1p}	o_{pp1}
lars myhren holand	51.51715	6	6	6	6
nil nisi bene	51.195225	5	5	5	9
viral hemoragisk septikemi	51.102116	6	8	6	6
remote operated vehicle	51.102116	6	6	8	6
geht besser nun	51.102116	6	8	6	6
acute respiratory infection	51.043217	5	5	10	5
ludens nomadic machines	50.932186	6	6	6	9
zero visibility corp	50.905716	5	11	5	5
sub specie aeternitatis	50.709797	7	9	7	7
hypertext markup language	50.401676	6	6	6	13

Finally, Pointwise Mutual Information is a more versatile measure that is capable of picking out a variety of MWEs, totalling 25.4 per cent of its 500 most highly ranked items. Note that no types representing grammatical MWEs were picked out by this AM. This shows very clearly the need for selecting the right AM depending on the specific objectives of the term extraction or lexical acquisition. However, all the other types were identified. The multiword anglicisms include multiword terms from various domains such as *hypertext markup language*, *deficit hyperactivity disorder*, *joint stock companies*, *checks and balances*, *frequently asked questions*, *catch and release* and *stream of consciousness*, as well as other salient multiword anglicisms of a more general nature, such as *worst case scenario* and *trick or treat*. The foreign multiword trigrams are especially culinary terms, such as *gambas al ajillo*, *spaghetti alla carbonara*, *chili con carne*, *biff chop suey*, *cafe au lait* and *pain au chocolat*, but also include terms from other domains such as *homo sapiens sapiens* and *tae kwon doe*, and also more general foreign multiwords such as *quod erat demonstrandum*, *cage aux folles* and *persona non grata*. Further, this AM picks out idiomatic phrases like the formulaic *snipp snapp snute* (used at the conclusion of fairy tales) and *bitte litte granne* 'teeny weeny bit'. The term candidates are mostly from medicine and include *viral hemoragisk septikemi*, *amyotrofisk lateral sklerose* and *hemolytisk uremisk syndrom*.

The top ten ranked trigrams using Pointwise Mutual Information can be inspected in Table 16, showing that it seems to be capable of extracting relevant term candidates. The top-ranked item is a proper noun written in lower-case that failed to be weeded out during filtering. The second item illustrates the importance of looking for longer sequences before approaching trigrams and bigrams. *nil nisi bene* is part of the Latin proverb *de mortuis nil nisi bene* (approximately: '(speak) of the dead only good'), whereas *ludens nomadic machines* is part of the sequence *Homo Ludens Nomadic Machines*. The third item is a term referring to a fish disease, other terms are *acute respiratory infection* (medicine) and–below the top ten–*file transfer protocol* (computers).

5 Conclusion and Future Work

This paper has approached multiword expressions by applying statistical association measures to two- and three word sequences (bigrams and trigrams) from the Norwegian Newspaper corpus. To this end, lists of bigrams and trigrams were generated, and they were pre-filtered for proper nouns and other named entities. This was necessary since statistical measures of collocation strength capture a wide range of phenomena, among these named entities. Another way of looking at it is to say that proper nouns are easily detected, by virtue of getting high ranks, by the same measures that work well for n-grams where the observed frequency of the components outside of the n-gram is low.

Based on our observations when testing nine different AMs on bigrams and four different ones for trigrams, the following conclusions seem justifiable. First, there are great differences with respect to the different AMs' abilities to extract n-grams that are relevant for lexicographical and terminological purposes, and a grouping of AMs according to these abilities has been proposed. Second, a necessary methodological strategy for future work is to begin with the longest sequences using a methodology for weeding out longer-sequence n-grams before moving to shorter-sequence n-grams. Third, future work should also aim at testing the AMs on a morphologically analysed version of the corpus, to test the different AMs on lists of n-grams that contain at least one lexical word or exclusively lexical words, to see if this may improve the performance of some of them.

References

Abney, Steven. 2000. *Statistical methods*. Nature Publishing Group, Macmillian.

Baldwin, Timothy. 2004. Multiword Expressions. Advanced course at the Australasian Language Technology Summer School (ALTSS 2004), Sydney, Australia.

Baldwin, Timothy and Francis Bond. 2002. Multiword Expressions: Some Problems for Japanese NLP. In *The Eigth Annual Meeting of the Association of Natural Language Processing*, pages 379–382. Keihanna.

Baldwin, Timothy and Su Nam Kim. 2010. Multiword Expressions. In N. Indurkhya and F. J. Damerau, eds., *Handbook of Natural Language Processing, Second Edition*. Boca Raton, FL: CRC Press, Taylor and Francis Group. ISBN 978-1420085921.

Banerjee, Satanjeev and Ted Pedersen. 2003. The Design, Implementation, and Use of the n-gram Statistic Package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, pages 370–381. Mexico City.

Biber, Douglas. 2009. A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics* 14(3):275–311.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, Edward Finegan, and Graeme Hirst. 1999. Longman Grammar of Spoken and Written English. London: Longman.

Church, Kenneth, William Gale, Patrick Hanks, and Donald Hindle. 1991. Using statistics in lexical analysis. In *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pages 115–164. Erlbaum.

Evert, Stefan. 2004. *The Statistics of Word Co-occurrences: Word Pairs and Collocations*. Ph.D. thesis, IMS, University of Stuttgart.

Fellbaum, Christiane, ed. 1998. *WordNet. An Electronical Lexical Database*. MIT Press.

Jackendoff, Ray. 1997. *The Architecture of the Language Faculty*. Cambridge, MA: The MIT Press.

Manning, Christopher and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge: Massachusetts: MIT Press.

McInnes, Bridget T. 2004. *Extending the Log Likelihood Measure to Improve Collocation Identification*. Master's thesis, University of Minnesota. (For the degree of master of science).

Sag, Ivan, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics*. Mexico City, Mexico.

Sinclair, John M. 1991. *Corpus, Concordance, Collocation*. Oxford : Oxford University Press.

Sinclair, John M. 1996. The search for units of meaning. In *Textus*, vol. 9, pages 75–106.

Stubbs, Michael. 1996. *Text and corpus analysis*. London: Blackwell.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins.