Corpora and collocations

Stefan Evert (stefan.evert@uos.de)
Institute of Cognitive Science, University of Osnabrück
49069 Osnabrück, Germany

Extended Manuscript, 13 October 2007

Contents

This is an extended manuscript of

Evert, Stefan (to appear). Corpora and collocations. To appear in A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, article 58. Mouton de Gruyter, Berlin.

The text has been shortened considerably for publication. In particular, all text in framed boxes (such as this one), all footnotes, the appendices, and some tables *do not appear* in the printed version. Please keep this in mind when citing the article.

1 Introduction

1.1 The controversy around collocations

The concept of *collocations* is certainly one of the most controversial notions in linguistics, even though it is based on a compelling, widely-shared intuition that certain words have a tendency to occur near each other in natural language. Examples of such collocations are *cow* and *milk*, *day* and *night*, *ring* and *bell*, or the infamous *kick* and *bucket*.[1] Other words, like *know* and *glass* or *door* and *year*, do not seem to be particularly attracted to each other.[2] J. R. Firth (1957) introduced the term "collocations" for characteristic and frequently recurrent word combinations, arguing that the meaning and usage of a word (the *node*) can to some extent be characterised by its most typical *collocates*: "You shall know a word by the company it keeps" (Firth 1957, 179). Firth was clearly aware of the limitations of this approach. He understood collocations as a convenient first approximation to meaning at a purely lexical level that can easily be operationalised (cf. Firth 1957, 181). Collocations in this Firthian sense can also be interpreted as empirical statements about the predictability of word combinations: they quantify the "mutual expectancy" (Firth 1957, 181) between words and the statistical influence a word exerts on its neighbourhood. Firth's definition of the term remained vague, though,[3] and it was only formalised and implemented after his death, by a group of British linguists often referred to as the Neo-Firthian school. Collocations have found widespread application in computational lexicography (Sinclair 1966, 1991), resulting in corpus-based dictionaries such as COBUILD (Sinclair 1995; see also Article 8).[4]

In parallel to the development of the Neo-Firthian school, the term "collocations" came to be used in the field of phraseology for semi-compositional and lexically determined word combinations such as *stiff drink* (with a special meaning of *stiff* restricted to a particular set of nouns), *heavy smoker* (where *heavy* is the only acceptable intensifier for *smoker*), *give a talk* (rather than *make* or *hold*) and *a school of fish* (rather than *group*, *swarm* or

[1] The first two examples are from Firth (1957), the third came up in a corpus of Dickens novels (as the second most strongly associated verb-noun combination after *shake* and *head*, for cooccurrence within sentences and the simple-ll measure). *Bell* is also the top collocate of the verb *ring* in the British National Corpus, according to the BNCweb collocation analysis (robustly for several association measures and span sizes).

[2] Both examples can be validated in the British National Corpus, using BNCweb. In the corpus of Dickens novels, *know* and *glass* show no significant association despite a cooccurrence frequency of $f = 27$ (two-sided Fisher's test $p = .776$, for verb-noun cooccurrences within sentences). In the Brown corpus, the nouns *door* and *year* show marginally significant evidence for a negative association (two-sided Fisher's test $p = .0221$, for noun-noun cooccurrences within sentences).

[3] "Moreover, these and other technical words are given their 'meaning' by the restricted language of the theory, and by applications of the theory in quoted works." (Firth 1957, 169)

[4] Firth himself obviously had lexicographic applications of collocations in mind: "It is clearly an essential procedure in descriptive lexicography" (Firth 1957, 180). He also anticipated the advent of corpus-based dictionaries and gave a "blueprint" of computational lexicography (Firth 1957, 195–196).

flock). This view has been advanced forcefully by Hausmann (1989) and has found increasingly widespread acceptance in recent years (e.g. Grossmann and Tutin 2003). It is notoriously difficult to give a rigorous definition of collocations in the phraseological sense and differentiate them from restricted word senses (most dictionaries have separate subentries for the special meanings of *stiff*, *heavy* and *school* in the examples above).[5] There is considerable overlap between the phraseological notion of collocations and the more general empirical notion put forward by Firth (cf. the examples given above), but they are also different in many respects (e.g., *good* and *time* are strongly collocated in the empirical sense, but *a good time* can hardly be understood as a non-compositional or lexically restricted expression). This poor alignment between two interpretations of the same term has resulted in frequent misunderstandings and has led to enormous confusion on both sides.[6] The situation is further complicated by a third meaning of "collocations" in the field of computational linguistics, where it is often used as a generic term for any lexicalised word combination that has idiosyncratic semantic or syntactic properties and may therefore require special treatment in a machine-readable dictionary or natural language processing system. This usage seems to originate with Choueka (1988) and can be found in standard textbooks, where collocations are often defined in terms of non-compositionality, non-modifiability and non-substitutability (Manning and Schütze 1999, 184). It has recently been superseded by the less ambiguous term *multiword expression* (cf. Sag *et al.* 2002).

An excellent overview of the competing definitions of collocations and their historical development is given by Bartsch (2004). Interestingly, she takes a middle road with her working definition of collocations as "lexically and/or pragmatically constrained recurrent co-occurrences of at least two lexical items which are in a direct syntactic relation with each other" (Bartsch 2004, 76).[7] For a compact summary, refer to Williams (2003).

1.2 Definitions and recommended terminology

In order to avoid further confusion, a consistent terminology should be adopted. Its most important goal is to draw a clear distinction between (i) the *empirical* concept of recurrent and predictable word combinations, which are a directly observable property of natural language, and (ii) the *theoretical* concept of lexicalised, idiosyncratic multiword expressions, defined by linguistic tests and speaker intuitions. In this article, the term "*collocations*" is used exclusively in its empirical Firthian sense (i), and we may occasionally speak of "*empirical collocations*" to draw attention to this fact. Lexicalised word combinations as a theoretical, phraseological notion (ii) are denoted by the generic term "*multiword expressions*", following its newly established usage in the field of computational linguistics. In phraseological theory, multiword expressions are divided into subcategories ranging from completely opaque idioms to semantically compositional word combinations, which are merely subject to arbitrary lexical restrictions (*brush teeth* rather than *scrub teeth*) or carry strong pragmatic connotations (*red rose*). A particularly interesting cate-

[5] The Oxford American Dictionary shipped with the Mac OS X operating system, for instance, has "*stiff* 2: (of an alcoholic drink) strong" and "*school*^{2}: a large group of fish or sea mammals".

[6] Some researchers even seem to have formed the impresssion of a war raging between the two camps, prompting them to offer peace talks "aus einer Position der Stärke" ["from a position of strength"] (Hausmann 2004).

[7] When Bartsch operationalises her notion of collocations, this working definition is amended with "an element of semantic opacity such that the meaning of the collocation cannot be said to be deducible as a function of the meanings of the constituents" (Bartsch 2004, 77). Thus, one might argue that collocations are a subclass of lexicalised multiword expressions, and perhaps even fall under the narrower phraseological concept of semi-compositional combinations.

gory in the middle of this cline are semi-compositional expressions, in which one of the words is lexically determined and has a modified or bleached meaning (classic examples are *heavy smoker* and *give a talk*). They correspond to the narrow phraseological meaning of the term "collocations" (cf. Grossmann and Tutin 2003) and can be referred to as "*lexical collocations*", following Krenn (2000). As has been pointed out above, it is difficult to give a precise definition of lexical collocations and to differentiate them e.g. from specialised word senses. Because of this fuzziness and the fact that many empirical collocations are neither completely opaque nor fully compositional, similar to lexical collocations, the two concepts are easily and frequently confused.

This article is concerned exclusively with empirical collocations, since they constitute one of the fundamental notions of corpus linguistics and, unlike lexicalisation phenomena, can directly be observed in corpora. It is beyond the scope of this text to delve into the voluminous theoretical literature on multiword expressions, but see e.g. Bartsch (2004) and Grossmann and Tutin (2003) for useful pointers. There is a close connection between empirical collocations and multiword expressions, though. A thorough analysis of the collocations found in a corpus study will invariably bring up non-compositionality and lexicalisation phenomena as an explanation for many of the observed collocations (cf. the case study in Section 2.2). Conversely, theoretical research in phraseology can build on authentic examples of multiword expressions obtained from corpora, avoiding the bias of relying on introspection or stock examples like *kick the bucket* (which is a rather uncommon phrase indeed: only three instances of the idiom can be found in the 100 million words of the British National Corpus).[8] *Multiword extraction* techniques exploit the often confusing overlap between the empirical and theoretical notions of collocation. Empirical collocations are identified as candidate multiword expressions, then the "false positives" are weeded out by manual inspection. A more detailed account of such multiword extraction procedures can be found in Section 6.2.

Following the Firthian tradition (e.g. Sinclair 1991), we define a collocation as a combination of two words that exhibit a tendency to occur near each other in natural language, i.e. to *cooccur* (but see the remarks on combinations of three or more words in Section 7.1). The term "*word pair*" is used to refer to such a combination of two words (or, more precisely, word *types*; see Article 36 for the distinction between types and tokens) in a neutral way without making a commitment regarding its collocational status. In order to emphasise this view of collocations as word pairs, we will use the notation (*kick*, *bucket*) instead of e.g. *kick (the) bucket*. In general, a word pair is denoted by (w_1, w_2), with $w_1 =$ *kick* and $w_2 =$ *bucket* in the previous example; w_1 and w_2 are also referred to as the *components* of the word pair. The term "word" is meant in the widest possible sense here and may refer to surface forms, case-folded surface forms, base forms, etc. (see Article 25). While collocations are most commonly understood as combinations of orthographic words, delimited by whitespace and punctuation, the concept and methodological apparatus can equally well be applied to combinations of linguistic units at other levels, ranging from morphemes to phrases and syntactic constructions (cf. Article 43).

In order to operationalise our definition of collocations, we need to specify the precise circumstances under which two words can be said to "cooccur". We also need a formal definition of the "attraction" between words reflected by their repeated cooccurrence, and a quantitative measure for the strength of this attraction. The *cooccurrence* of words can be defined in many different ways. The most common approaches are (i) *surface cooccurrence*,

[8] There are 20 instances of the collocation (*kick*, *bucket*) in the British National Corpus. Of these, 8 instances are literal uses (e.g., *It was as if God had kicked a bucket of water over.*), 9 instances cite the idiom *kick the bucket* in a linguistic meta-discussion, and only 3 are authentic uses of the idiom. See Appendix A.2 for a complete listing of the corpus examples.

where words are said to cooccur if they appear close to each other in running text, measured by the number of intervening word tokens; (ii) *textual cooccurrence* of words in the same sentence, clause, paragraph, document, etc.; and (iii) *syntactic cooccurrence* between words in a (direct or indirect) syntactic relation, such as a noun and its modifying adjective (which tend to be adjacent in most European languages) or a verb and its object noun (which may be far apart at the surface, cf. Goldman *et al.* (2001, 62) for French). These three definitions of cooccurrence are described in more detail in Section 3, together with appropriate methods for the calculation of cooccurrence frequency data.

The hallmark of an attraction between words is their frequent cooccurrence, and collocations are sometimes defined simply as "recurrent cooccurrences" (Smadja 1993, 147; Bartsch 2004, 11). Strictly speaking, any pair of words that cooccur at least twice in a corpus is a potential collocation according to this view. It is common to apply higher *frequency thresholds*, however, such as a minimum of 3, 5 or even 10 cooccurrences. Evert (2004, Ch. 4) gives a mathematical justification for this approach (see also Section 7.1), but a more practical reason is to reduce the enormous amounts of data that have to be processed. It is not uncommon to find more than a million recurrent word pairs ($f \geq 2$) in a corpus containing several hundred million running words, but only a small proportion of them will pass a frequency threshold of $f \geq 10$ or higher, as a consequence of Zipf's law (cf. Article 37).[9] In the following, we use the term "*recurrent word pair*" for a potential collocation that has passed the chosen frequency threshold in a given corpus.

Mere recurrence is no sufficient indicator for a strong attraction between words, though, as will be illustrated in Section 4.1. An additional measure of attraction strength is therefore needed in order to identify "true collocations" among the recurrent word pairs, or to distinguish between "strong" and "weak" collocations. The desire to generalise from recurrent word pairs in a particular corpus (as a sample of language) to collocations in the full language or sublanguage, excluding word pairs whose recurrence may be an accident of the sampling process, has led researchers to the concept of *statistical association* (Sinclair 1966, 418). Note that this mathematical meaning of "association" describes a statistical attraction between certain events and must not be confused with psychological association (as e.g. in word association norms, which have no direct connection to the statistical association between words that is of interest here). By interpreting occurrences of words as events, statistical *association measures* can be used to quantify the attraction between cooccurring words, completing the formal definition of empirical collocations.

The most important association measures will be introduced in Sections 4 and 5, but many other measures have been suggested in the mathematical literature and in collocation studies. Such measures assign an *association score* to each word pair, with high scores indicating strong attraction and low scores indicating weak attraction (or even repulsion) between the component words. Association scores can then be used to select "true collocations" by setting a threshold value, or to rank the set of recurrent word pairs according to the strength of their attraction (so that "strong" collocations are found at the top of the list). These uses of association scores are further explained in Section 2.1. It is important to keep in mind that different association measures may lead to entirely different rankings of the word pairs (or to different sets of "true collocations"). Section 6 gives some guidance on how to choose a suitable measure.

[9] In the British National Corpus, there are ca. 3.6 million bigram types (excluding punctuation etc.) with $f \geq 2$. Less than 700,000 pass a threshold of $f \geq 10$, and only 160,000 pass $f \geq 50$.

1.3 Overview of the article

Section 2 describes the different uses of association scores and illustrates the linguistic properties of empirical collocations with a case study of the English noun *bucket*. The three types of cooccurrence (surface, textual and syntactic) are defined and compared in Section 3, and the calculation of cooccurrence frequency data is explained with the help of toy examples. Section 4 introduces the concepts of statistical association and independence underlying all association measures. It also presents a selection of simple measures, which are based on a comparison of observed and expected cooccurrence frequency. Section 5 introduces more complex statistical measures based on full-fledged contingency tables. The difficulty of choosing between the large number of available measures is the topic of Section 6, which discusses various methods for the comparison of association measures. Finally, Section 7 addresses some open questions and extensions that are beyond the scope of this article, and lists references for further reading.

Readers in a hurry may want to start with the "executive summaries" in Section 4.3 and at the beginning of Section 7, which give a compact overview of the collocation identification process with simple association measures. You should also skim the examples in Section 3 to understand how appropriate cooccurrence frequency data are obtained from a corpus, find out in Section 4.1 how to calculate observed cooccurrence frequency O and expected frequency E, and refer to Figure 4 for the precise equations of various simple association measures.

2 What are collocations?

2.1 Using association scores

Association scores as a quantitative measure of the attraction between words play a crucial role in the operationalisation of empirical collocations, next to the formal definition of cooccurrence and the appropriate calculation of cooccurrence frequency data. While the interpretation of association scores seems straightforward (high scores indicate strong attraction), they can be used in different ways to identify collocations among the recurrent word pairs found in a corpus. The first contrast to be made is whether collocativity is treated as a categorical phenomenon or as a cline, leading either to *threshold* approaches (which attempt to identify "true collocations") or to *ranking* approaches (which place word pairs on a scale of collocational strength without strict separation into collocations and non-collocations). A second contrast concerns the grouping of collocations: the *unit* view is interested in the most strongly collocated word pairs, which are seen as independent units; the *node–collocate* view focuses on the collocates of a given node word, i.e. "the company it keeps". The two contrasts are independent of each other in principle, although the node–collocate view is typically combined with a ranking approach.

In a threshold approach, recurrent word pairs whose association score exceeds a (more or less arbitrary) threshold value specified by the researcher are accepted as "true collocations". We will sometimes refer to them as an *acceptance set* for a given association measure and threshold value. In the alternative approach, all word pairs are ranked according to their association scores. Pairs at the top of the ranked list are then considered "more collocational", while the ones at the bottom are seen as "less collocational". However, no categorical distinction between collocations and non-collocations is made in this approach. A third strategy combines the ranking and threshold approaches by accepting the first n word pairs from the ranked list as collocations, with n either determined interactively by the researcher or dictated by the practical requirements of an application.

Typical choices are $n = 100$, $n = 500$, $n = 1{,}000$ and $n = 2{,}000$. Such n-best lists can be interpreted as acceptance sets for a threshold value determined from the corpus data (such that exactly n word pairs are accepted) rather than chosen at will. Because of the arbitrariness of pre-specified threshold values and the lack of good theoretical motivations (cf. Section 4.2), n-best lists should always be preferred over threshold-based acceptance sets. It is worth pointing out that in either case the ranking, n-best list or acceptance set depends critically on the particular association measure that has been used. The n-best lists shown in Tables 4 and 5 are striking examples of this fact.

The unit view interprets collocations as pairs of words that show a strong mutual attraction, or "mutual expectancy" (Firth 1957, 181).[10] It is particularly suitable and popular for multiword extraction tasks, where n-best lists containing the most strongly associated word pairs in a corpus are taken as candidate multiword expressions. Such candidate lists serve e.g. as base material for dictionary updates, as terminological resources for translators and technical writers, and for the semi-automatic compilation of lexical resources for natural language processing systems (e.g. Heid *et al.* 2000). The node–collocate view, on the other hand, focuses on the predictability of word combinations, i.e. on how a word (the node) determines its "company" (the collocates). It is well suited for the linguistic description of word meaning and usage in the Firthian tradition, where a node word is characterised by ranked lists of its collocates (Firth 1957). Following Firth (1957, 195–196) and Sinclair (1966), this view has also found wide acceptance in modern corpus-based lexicography (e.g. Sinclair 1991; Kilgarriff *et al.* 2004), in particular for learner dictionaries such as COBUILD (Sinclair 1995) and the Oxford Collocations Dictionary (Lea 2002).[11]

In addition to their "classic" applications in language description, corpus-based lexicography and multiword extraction, collocations and association scores have many practical uses in computational linguistics and related fields. Well-known examples include the construction of machine-readable dictionaries for machine translation and natural language generation systems, the improvement of statistical language models, and the use of association scores as features in vector space models of distributional semantics. See Evert (2004, 23–27) for an overview and comprehensive references.

2.2 Collocations as a linguistic epiphenomenon

The goal of this section is to help readers reach an intuitive understanding of the empirical phenomenon of collocations and their linguistic properties. First and foremost, collocations are observable facts about language, i.e. primary data. From a strictly data-driven perspective, they can be interpreted as empirical predictions about the neighbourhood of a word. For instance, a verb accompanying the noun *kiss* is likely to be either *give*, *drop*, *plant*, *press*, *steal*, *return*, *deepen*, *blow* or *want*.[12] From the explanatory perspective of theoretical linguistics, on the other hand, collocations are best characterised as an epiphenomenon: idioms, lexical collocations, clichés, cultural stereotypes, semantic compatibility and many other factors are hidden causes that result in the observed associations between words.[13]

[10] "The collocation of a word or 'piece' is not to be regarded as mere juxtaposition, it is an order of *mutual expectancy*. The words are mutually expectant and mutually prehended." (Firth 1957, 181)

[11] Another example are recent approaches to language teaching, in particular the *profiles combinatoires* of Blumenthal *et al.* (2005).

[12] This prediction is correct in about a third of all cases. In the British National Corpus, there are 1,003 instances of the noun *kiss* cooccurring with a lexical verb within a span of 3 words. For 343 of them (= 34%), the verb is one of the collocates listed above.

[13] Firth's description of collocations as "an order of mutual expectancy" (Firth 1957, 181) may seem to suggest that collocations are pre-fabricated units in which the node "primes" the collocate, and vice versa.

collocate	$f \geq 3$	MI
fourteen-record	4	13.31
ten-record	3	13.31
full-track	3	12.89
single-record	5	12.63
randomize	10	10.80
galvanized	4	10.67
groundbait	3	10.04
slop	17	10.03
spade	31	9.41
Nessie	4	9.34
leaky	3	8.59
mop	20	8.57
bottomless	3	8.33
douse	4	8.28
galvanised	3	8.04
oats	7	7.96
shovel	8	7.84
Rhino	7	7.77
synonym	7	7.62
iced	3	7.41

collocate	$f \geq 3$	simple-ll
water	184	1083.18
a	590	449.30
spade	31	342.31
plastic	36	247.65
size	42	203.36
slop	17	202.30
mop	20	197.68
throw	38	194.66
fill	37	191.44
with	196	171.78
into	87	157.30
empty	27	152.72
and	479	152.19
record	43	151.98
bucket	18	140.88
ice	22	132.78
randomize	10	129.76
of	497	109.33
kick	20	108.08
large	37	88.53

Table 1: Collocates of *bucket* in the BNC (all words). [*extended manuscript only*]

In order to gain a better understanding of collocations both as an empirical phenomenon and as an epiphenomenon, we will now take a look at a concrete example, viz. how the noun *bucket* is characterised by its collocates in the British National Corpus (BNC, Aston and Burnard 1998). The data presented here are based on surface cooccurrence with a span size of 5 words, delimited by sentence boundaries (see Section 3). Observed and expected frequencies were calculated as described in Section 4.1. Collocates were lemmatised, and punctuation, symbols and numbers were excluded. Association scores were calculated for the measures MI and simple-ll (see Section 4.2).

The BUCKET data set was extracted from the British National Corpus (XML Edition), using the open-source corpus search engine CQP, which is a part of the IMS Corpus Workbench.[14] Instances of the node were identified by searching for the base form *bucket* (according to the BNC annotation), tagged unambiguously as a noun (`NN1` or `NN2`). Base forms of all orthographic words (delimited by whitespace and punctuation) within a symmetric span of 5 words (excluding punctuation) around the node instances were collected. Spans were further limited by sentence boundaries, and punctuation, other symbols and numbers were excluded as collocates. Finally, a frequency threshold of $f \geq 3$ was applied.

A first observation is that different association measures will produce entirely different rankings of the collocates. For the MI measure, the top collocates are *fourteen-record*, *ten-record*, *full-track*, *single-record*, *randomize*, *galvanized*, *groundbait*, *slop*, *spade*, *Nessie*. Most of them are infrequent words with low cooccurrence frequency (e.g., *groundbait* occurs only 29 times in the BNC). Interestingly, the first five collocates belong to a technical sense of *bucket* as a data structure in computer science; others such as *groundbait* and *Nessie* (the name of a character in the novel *Worlds Apart*, BNC file ATE) are purely accidental combinations. By contrast, the top collocates according to the simple-ll measure are dominated by high-frequency cooccurrences with very common words, including several function words: *water*, *a*, *spade*, *plastic*, *size*, *slop*, *mop*, *throw*, *fill*, *with*.

However, as any statistics textbook will point out, association does not imply causality: the occurrences of both words might be triggered by a hidden third factor, resulting in an indirect association of the word pair.

[14] See `http://cwb.sourceforge.net/`.

noun	*f*	simple-ll	verb	*f*	simple-ll	adjective	*f*	simple-ll
water	183	1063.90	*throw*	36	165.32	*large*	37	92.72
spade	31	338.21	*fill*	29	129.69	*single-record*	5	79.56
plastic	36	242.63	*randomize*	9	115.33	*cold*	13	52.63
slop	14	197.65	*empty*	14	106.51	*galvanized*	4	52.35
size	41	193.22	*tip*	10	62.65	*ten-record*	3	49.75
mop	16	183.97	*kick*	12	59.12	*full*	20	46.34
record	38	155.64	*hold*	31	58.52	*empty*	9	36.41
bucket	18	138.70	*carry*	26	55.68	*steaming*	4	36.37
ice	22	131.68	*put*	36	48.69	*full-track*	2	33.17
seat	20	78.35	*chuck*	7	48.40	*multi-record*	2	33.17
coal	16	76.44	*weep*	7	44.14	*small*	21	30.90
density	11	66.78	*pour*	9	39.35	*leaky*	3	30.14
brigade	10	66.78	*douse*	4	37.85	*bottomless*	3	29.04
algorithm	9	66.54	*fetch*	7	35.22	*galvanised*	3	28.34
shovel	7	64.53	*store*	7	30.77	*iced*	3	25.46
container	10	62.40	*drop*	9	21.76	*clean*	7	25.17
oats	7	62.32	*pick*	11	21.74	*wooden*	6	24.14
sand	12	61.91	*use*	31	20.93	*old*	19	18.83
Rhino	7	60.50	*tire*	3	20.58	*ice-cold*	2	17.66
champagne	10	59.28	*rinse*	3	20.19	*anti-sweat*	1	16.58

Table 2: Collocates of *bucket* in the BNC (nouns, verbs and adjectives).

Table 1 shows the 20 highest-ranked collocates according to each association measure, together with the calculated association scores and raw cooccurrence frequencies (in the column labelled f). An interesting case is the noun *synonym* at the bottom of the MI table, which is both coincidental and related to the technical sense of *bucket*: 6 of the 7 cooccurrences come from a single computer science text (BNC file FPG), whose authors use *synonym* as an ad-hoc term for data records stored in the same bucket. The simple-ll table contains a striking number of function words: *a*, *with*, *into*, *and*, *of*.

Although human readers will easily spot some more "interesting" collocates such as *kick*, it would be sensible to apply a stop word list in order to remove function words and other very general collocates (although it can be interesting in some cases to look for collocations with function words or syntactic constructions, see e.g. Article 43). The highest-ranking collocates according to simple-ll and restricted to words from lexical categories are *water*, *spade*, *plastic*, *slop*, *size*, *mop*, *throw*, *fill*, *empty*, *record*, *bucket*, *ice*, *randomize*, *kick*, *large*, *seat* and *single-record*. Clearly, this list gives a much better impression of the usage of the noun *bucket* than the unfiltered list above.

A clearer picture emerges when different parts of speech among the collocates (e.g. nouns, verbs and adjectives) are listed separately, as shown in Table 2 for the simple-ll measure. Ideally, a further distinction should be made according to the syntactic relation between node and collocate (node as subject/object of verb, prenominal adjective modifying the node, head of postnominal *of*-NP, etc.), similar to the lexicographic *word sketches* of Kilgarriff *et al.* (2004). Parts of speech provide a convenient approximation that does not require sophisticated automatic language processing tools. A closer inspection of the lists in Table 2 underlines the status of collocations as an epiphenomenon, revealing many different causes that contribute to the observed associations:

- the well-known idiom *kick the bucket*, although many of the cooccurrences represent a literal reading of the phrase (e.g. *It was as if God had kicked a bucket of water over.*, G0P: 2750);[15]

[15] A complete listing of the cooccurrences of *kick* and *bucket* in the BNC can be found in Appendix A.2. Note the lower cooccurrence frequency in Table 2 because only collocates with unambiguous part-of-speech tags were included there.

- proper names such as *Rhino Bucket*, a hard rock band founded in 1987;[16]
- both lexicalised and productively formed compound nouns: *slop bucket*, *bucket seat*, *coal bucket*, *champagne bucket* and *bucket shop* (the 23rd noun collocate);
- lexical collocations like *weep buckets*, where *buckets* has lost its regular meaning and acts as an intensifier for the verb;
- cultural stereotypes and institutionalised phrases such as *bucket and spade* (which people prototypically take along when they go to a beach, even though the phrase has fully compositional meaning);
- reflections of semantic compatibility: *throw*, *carry*, *kick*, *tip*, *take*, *fetch* are typical things one can do with a bucket, and *full*, *empty*, *leaky* are some of its typical properties (or states);[17]
- semantically similar terms (*shovel*, *mop*) and hypernyms (*container*);
- facts of life, which do not have special linguistic properties but are frequent simply because they describe a situation that often arises in the real world; a prototypical example is *bucket of water*, the most frequent noun collocate in Table 2;[18]
- linguistic relevance: it is more important to talk about *full*, *empty* and *leaky* buckets than e.g. about a rusty or yellow bucket; interestingly, *old bucket* ($f = 19$) is much more frequent than *new bucket* ($f = 3$, not shown);[19] and
- "indirect" collocates (e.g. *a bucket of cold, warm, hot, iced, steaming water*), describing typical properties of the liquid contained in a bucket.[20]

Obviously, there are entirely different sets of collocates for each sense of the node word, which are overlaid in Table 2. As Firth put it: "there are the specific contrastive collocations for *light/dark* and *light/heavy*" (Firth 1957, 181). In the case of *bucket*, a technical meaning, referring to a specific data structure in computer science, is conspicuous and accounts for a considerable proportion of the collocations (*bucket brigade algorithm*, *bucket size*, *randomize to a bucket*, *store records in bucket*, *single-record bucket*, *ten-record bucket*). In order to separate collocations for different word senses automatically, a sense-tagged corpus would be necessary (cf. Article 26).

[16] Note that the name was misclassified as a sequence of two common nouns by the automatic part-of-speech tagging of the BNC.

[17] Sometimes the connection only becomes clear when a missing particle or other element is added to the collocate, e.g. *pick (up) – bucket*.

[18] Many other examples of facts-of-life collocates can be found in the table, e.g. *galvanised, wooden bucket* and *bucket of sand, ice, coal*.

[19] It is reasonable to suppose that an *empty* bucket is a more important topic than a *full* bucket, but the data in Table 2 seem to contradict this intuition. This impression is misleading: of the 20 cooccurrences of *bucket* and *full* in the BNC, only 4 refer to full buckets, whereas the remaining 16 are instances of the construction *bucket full of sth.* Thus, *empty* is indeed much more strongly collocated than *full* as an intersective adjectival modifier.

[20] On a side note, the self-collocation (*bucket*, *bucket*) in the list of noun collocates is partly a consequence of term clustering (cf. Article 36), but it also reflects recurrent constructions such as *from bucket to bucket* and *bucket after bucket*.

adjective-noun	$f \geq 5$	simple-ll	noun-noun	$f \geq 5$	simple-ll	verb-preposition	$f \geq 5$	simple-ll
last year	87	739.76	*no one*	40	505.32	*looked at*	79	452.88
same time	95	658.47	*carbon tetrachloride*	18	334.79	*look at*	68	389.72
fiscal year	55	605.25	*wage rate*	24	319.65	*stared at*	33	245.89
last night	61	540.72	*home runs*	25	310.36	*look like*	32	244.62
high school	53	495.26	*anode holder*	19	309.42	*depends on*	30	217.03
last week	51	479.03	*living room*	25	302.75	*live in*	54	216.94
great deal	43	475.45	*index words*	24	291.31	*talk about*	27	213.40
dominant stress	31	464.40	*index word*	21	248.62	*went into*	38	206.36
nineteenth century	32	462.73	*hearing officer*	17	244.55	*worry about*	21	201.79
other hand	60	443.24	*chemical name*	18	242.75	*looked like*	28	198.36
old man	56	373.96	*radio emission*	16	236.38	*deal with*	27	195.40
young man	50	370.57	*oxidation pond*	13	234.39	*sat down*	20	185.09
first time	66	357.51	*capita income*	14	229.86	*account for*	24	177.63
foreign policy	30	351.24	*information cell*	18	217.65	*serve as*	28	169.42
few days	42	350.72	*station wagon*	15	212.49	*looks like*	17	156.15
nuclear weapons	23	325.97	*urethane foam*	12	200.74	*cope with*	18	139.63
few years	48	325.93	*wash wheel*	12	196.14	*came from*	39	137.54
real estate	24	320.51	*school districts*	16	193.60	*do with*	70	133.01
few minutes	33	303.53	*urethane foams*	11	192.93	*look for*	37	129.90
electronic switches	18	298.85	*interest rates*	17	181.42	*fall into*	16	129.45

Table 3: Most strongly collocated bigrams in the Brown corpus, categorised into adjective-noun, noun-noun and verb-preposition bigrams (other bigram types are not shown). [*extended manuscript only*]

This example also illustrates a general problem of frequency data obtained from balanced samples like the British National Corpus (cf. Article 10 for details on the BNC composition). Collocations related to the computer-science sense of *bucket* are almost exclusively found in a single document (BNC file FNR), which is concerned with the relation between bucket size and data packing density. These collocations may therefore well be specific to the author or topic of this document, and not characteristic of the typical usage of the noun *bucket* in computer science. Another example of accidental cooccurrence is *bucket of oats*, where 4 out of 6 cooccurrences stem from a text on horse feeding (BNC file ADF).

Observant readers may have noticed that the list of collocations in Table 2 is quite similar to the entry for *bucket* in the Oxford Collocations Dictionary (OCD, Lea 2002). This is not as surprising as it may seem at first, since the OCD is also based on the British National Corpus as its main source of corpus data (Lea 2002, viii). Obviously, collocations were identified with a technique similar to the one used here.

As a second case study, Table 3 shows the most strongly collocated bigrams in the Brown corpus (Francis and Kucera 1964) according to the simple-ll measure. For this BIGRAM data set, pairs of adjacent words were extracted from the Brown corpus, excluding punctuation and other non-word tokens. As an exception, sentence-ending punctuation (*.*, *!* or *?*) was allowed in the second position. A frequency threshold of $f \geq 5$ was applied, and the remaining 24,770 word pairs were ranked according to simple-ll scores. In order to give a better impression of the underlying linguistic phenomena, the bigrams were categorised by part-of-speech combination. Only adjective-noun, noun-noun and verb-preposition bigrams are displayed in Table 3.

3 Cooccurrence and frequency counts

As has already been stated in Section 1.2, the operationalisation of collocations requires a precise definition of the cooccurrence, or “nearness”, of two words (or, more precisely, word *tokens*). Based on this definition, cooccurrence frequency data for each recurrent word pair (or, more precisely, pair of word *types*) can be obtained from a corpus. Association scores as a measure of attraction between words are then calculated from these frequency data. It will be shown in Section 4.1 that *cooccurrence frequency* alone is not sufficient to quantify the strength of attraction.[21] It is also necessary to consider the occurrence frequencies of the individual words, known as *marginal frequencies*,[22] in order to assess whether the observed cooccurrences might have come about by chance. In addition, a measure of corpus size is needed to interpret absolute frequency counts. This measure is referred to as *sample size*, following statistical terminology.

The following notation is used in this article: O for the "observed" cooccurrence frequency in a given corpus (sometimes also denoted by f, especially when specifying frequency thresholds such as $f \geq 5$); f_1 and f_2 for the marginal frequencies of the first and second component of a word pair, respectively; and N for the sample size. These four numbers provide the information needed to quantify the statistical association between two words, and they are called the *frequency signature* of the pair (Evert 2004, 36). Note that a separate frequency signature is computed for every recurrent word pair (w_1, w_2) in the corpus. The set of all such recurrent word pairs together with their frequency signatures is referred to as a *data set*.

Three different approaches to measuring nearness are introduced below and explained with detailed examples: *surface*, *textual* and *syntactic* cooccurrence. For each type of cooccurrence, an appropriate procedure for calculating frequency signatures (O, f_1, f_2, N) is described. The mathematical reasons behind these procedures will become clear in Section 5. The aim of the present section is to clarify the logic of computing cooccurrence frequency data. Practical implementations that can be applied to large corpora use more efficient algorithms, especially for surface cooccurrences (e.g. Gil and Dias 2003; Terra and Clarke 2004).

3.1 Surface cooccurrence

The most common approach in the Firthian tradition defines cooccurrence by surface proximity, i.e. two words are said to cooccur if they appear within a certain distance or *collocational span*, measured by the number of intervening word tokens. Surface cooccurrence is often, though not always combined with a node–collocate view, looking for collocates within the collocational spans around the instances of a given node word.

Span size is the most important choice that has to be made by the researcher. The most common values range from 3 to 5 words (e.g. Sinclair 1991), but many other span sizes can be found in the literature. Some studies in computational linguistics have focused on bigrams of immediately adjacent words, i.e. a span size of 1 (e.g. Choueka 1988; Schone and Jurafsky 2001), while others have used span sizes of dozens or hundreds of words, especially in the context of distributional semantics (Schütze 1998).[23] Other decisions are whether to count only word tokens or all tokens (including punctuation and numbers), how to deal with multiword units (does *out of* count as a single token or as two tokens?), and whether cooccurrences are allowed to cross sentence boundaries.

[21] For example, the bigrams *Rhode Island* and *to which* both occur 100 times in the Brown corpus (based on the BIGRAM data set described in Section 2.2). The former combination is much more predictable, though, in the sense that *Rhode* is more likely to be followed by *Island* (100 out of its 105 occurrences) than *to* by *which* (100 out of 25,000 occurrences). The precise frequency signatures of the two bigrams are $(100, 105, 175, 909768)$ for *Rhode Island* and $(101, 25106, 2766, 909768)$ for *to which*.

[22] See Section 5.1 for an explanation of this term.

[23] Schütze (1998) used symmetric spans with a total size of 50 tokens, i.e. 25 tokens to the left and 25 to the right.

A vast deal of coolness and a peculiar degree of judgement, are ⌊requisite in catching a **hat**⌋. A man must not be precipitate, or he runs over it; he must not rush into the opposite extreme, or he loses it altogether. [...] There was a fine gentle ⌊wind, and Mr. Pickwick's **hat** *rolled* sportively before it⌋. The wind puffed, and Mr. ⌊Pickwick puffed, and the **hat** *rolled* over and over⌋ as merrily as a lively porpoise in a strong tide; and on it might have *rolled*, far beyond Mr. Pickwick's reach, had not its course been providentially stopped, just as that gentleman was on the point of resigning it to its fate.

Figure 1: Illustration of surface cooccurrence for the word pair (*hat*, *roll*).

Figure 1 shows surface cooccurrences between the words *hat* (in bold face, as node) and the collocate *roll* (in italics, as collocate). The span size is 4 words, excluding punctuation and limited by sentence boundaries. Collocational spans around instances of the node word *hat* are indicated by brackets below the text.[24] There are two cooccurrences in this example, in the second and third span, hence $O = 2$. Note that multiple instances of a word in the same span count as multiple cooccurrences, so for *hat* and *over* we would also calculate $O = 2$ (with both cooccurrences in the third span). The marginal frequencies of the two words are given by their overall occurrence counts in the text, i.e. $f_1 = 3$ for *hat* and $f_2 = 3$ for *roll*. The sample size N is simply the total number of tokens in the corpus, counting only tokens that are relevant to the definition of spans. In our example, N is the number of word tokens excluding punctuation, i.e. $N = 111$ for the text shown in Figure 1. If we include punctuation tokens in our distance measurements, the sample size would accordingly be increased to $N = 126$ (9 commas, 4 full stops and 2 semicolons). The complete frequency signature for the pair (*hat*, *roll*) is thus $(2, 3, 3, 111)$. Of course, realistic data will have much larger sample sizes, and the marginal frequencies are usually considerably higher than the cooccurrence frequency.

Collocational spans can also be asymmetric, and are generally written in the form (Lk, Rn) for a span of k tokens to the left of the node word and n tokens to its right. The symmetric spans in the example above would be described as (L4, R4). Asymmetric spans introduce an asymmetry between node word and collocate that is absent from most other approaches to collocations. For a one-sided span (L4, R0) to the left of the node word, there would be 2 cooccurrences of the pair (*roll*, *hat*) in Figure 1, but none of the pair (*hat*, *roll*). A special case are spans of the form (L0, R1), where cooccurrences are ordered pairs of immediately adjacent words, often referred to as *bigrams* in computational linguistics. Thus, *took place* would be a bigram cooccurrence of the lemma pair (*take*, *place*), but neither *place taken* nor *take his place* would count as cooccurrences.

3.2 Textual cooccurrence

A second approach considers words to cooccur if they appear in the same textual unit. Typically, such units are sentences or utterances, but with the recent popularity of Google searches and the Web as corpus (see Article 18), cooccurrence within (Web) documents has found more widespread use.

One criticism against surface cooccurrence is the arbitrary choice of the span size. For a span size of 3, *throw a birthday party* would be accepted as a cooccurrence of (*throw*, *party*), but *throw a huge birthday party* would not. This is particularly counterintuitive for languages with relatively free word order, where closely related words can be far apart at the surface.[25] In such languages, textual cooccurrence within the same sentence may

[24] All text samples in this section have been adapted from the novel *The Pickwick Papers* by Charles Dickens.

[25] Consider the German collocation *(einen) Blick zuwerfen*, which cooccurs at a distance of 16 words in the

provide a more appropriate collocational span. Textual cooccurrence also captures weaker dependencies, in particular those caused by paradigmatic semantic relations. For example, if an English sentence contains the noun *bucket*, it is quite likely to contain the noun *mop* as well (although the connection is far weaker than for *water* or *spade*), but the two nouns will not necessarily be near each other in the sentence.

A vast deal of coolness and a peculiar degree of judgement, are requisite in catching a hat.	hat	—
A man must not be precipitate, or he runs *over* it ;	—	over
he must not rush into the opposite extreme, or he loses it altogether.	—	—
There was a fine gentle wind, and Mr. Pickwick's hat rolled sportively before it.	hat	—
The wind puffed, and Mr. Pickwick puffed, and the hat rolled *over* and *over* as merrily as a lively porpoise in a strong tide ;	hat	over

Figure 2: Illustration of textual cooccurrence for the word pair (*hat*, *over*).

The definition of textual cooccurrence and the appropriate procedure for computing frequency signatures are illustrated in Figure 2, for the word pair (*hat*, *over*) and sentences as textual segments. There is one cooccurrence of *hat* and *over* in the last sentence of this text sample, hence $O = 1$. In contrast to surface cooccurrence, the count is 1 even though there are two instances of *over* in the sentence. Similarly, the marginal frequencies are given by the number of sentences containing each word, ignoring multiple occurrences in the same sentence: hence $f_1 = 3$ and $f_2 = 2$ (although there are three instances each of *hat* and *over* in the text sample). The sample size $N = 5$ is the number of sentences in this case. The complete frequency signature of (*hat*, *over*) is thus $(1, 3, 2, 5)$, whereas for surface cooccurrence within the spans shown in Figure 1 it would have been $(2, 3, 3, 79)$.

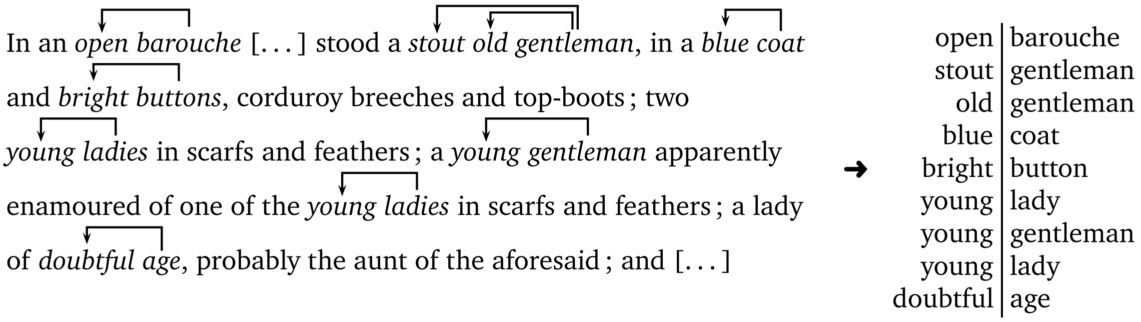
The most intuitive procedure for calculating frequency signatures is the method shown in Figure 2. Each sentence is written on a separate line and marked for occurrences of the two words in question. The marginal frequencies and the cooccurrence frequency can then be read off directly from this table of yes/no marks (shown to the right of the vertical line). In principle, the same procedure has to be repeated for every word pair of interest, but more efficient implementations pass through the corpus only once, generating frequency signatures for all recurrent word pairs in parallel.

3.3 Syntactic cooccurrence

In this more restrictive approach, words are only considered to be near each other if there is a direct syntactic relation between them. Examples are a verb and its object (or subject) noun, prenominal adjectives (in English and German) and nominal modifiers (the pattern N *of* N in English, genitive noun phrases in German). Sometimes, indirect relations might also be of interest, e.g. a verb and the adjectival modifier of its object noun, or a noun and the adjective modifying a postnominal *of*-NP. The latter pattern

sentence *Der Blick voll inniger Liebe und Treue, fast möchte ich sagen Hundetreue, welchen er mir dabei zaghaft zuwarf, drang mir tief zu Herzen.* (Karl May, *In den Schluchten des Balkan*).

accounts for several surface collocations of the noun *bucket* such as *a bucket of iced, cold, steaming water* (cf. Table 2). Collocations for different types of syntactic relations are usually treated separately. From a given corpus, one might extract a data set of verbs and their object nouns, another data set of verbs and subject nouns, a data set of adjectives modifying nouns, etc. Syntactic cooccurrence is particularly appropriate if there may be long-distance dependencies between collocates: unlike surface cooccurrence, it does not set an arbitrary distance limit, but at the same time introduces less “noise” than textual cooccurrence.[26] Syntactic cooccurrence is often used for multiword extraction, since many types of lexicalised multiword expressions tend to appear in specific syntactic patterns such as verb + object noun, adjective + noun, adverb + verb, verb + predicated adjective, delexical verb + noun, etc. (see Bartsch 2004, 11).[27]



In an *open barouche* [...] stood a *stout old gentleman*, in a *blue coat* and *bright buttons*, corduroy breeches and top-boots; two *young ladies* in scarfs and feathers; a *young gentleman* apparently enamoured of one of the *young ladies* in scarfs and feathers; a lady of *doubtful age*, probably the aunt of the aforesaid; and [...]

➜

open	barouche
stout	gentleman
old	gentleman
blue	coat
bright	button
young	lady
young	gentleman
young	lady
doubtful	age

Figure 3: Illustration of syntactic cooccurrence (nouns modified by prenominal adjectives).

Frequency signatures for syntactic cooccurrence are obtained in a more indirect way, illustrated in Figure 3. First, all instances of the desired syntactic relation are identified, in this case modification of nouns by prenominal adjectives. Then the corresponding arguments are compiled into a list with one entry for each instance of the syntactic relation (shown on the right of Figure 3). Note that the list entries are lemmatised here, but e.g. case-folded word forms could have been used as well. Just as the original corpus is understood as a sample of language, the list items constitute a sample of the targeted syntactic relation, and Evert (2004) refers to them as “pair tokens”. Cooccurrence frequency data are computed from this sample, while all word tokens that do not occur in the relation of interest are disregarded. For the word pair (*young*, *gentleman*), we find one cooccurrence in the list of pair tokens, i.e. $O = 1$. The marginal frequencies are given by the total numbers of entries containing one of the component words, $f_1 = 3$ and $f_2 = 3$, and the sample size is the total number of list entries, $N = 9$. The frequency signature of (*young*, *gentleman*) as a syntactic adjective-noun cooccurrence is thus $(1, 3, 3, 9)$.

---

[26] In the German example *Der Blick voll inniger Liebe und Treue, fast möchte ich sagen Hundetreue, welchen er mir dabei zaghaft zuwarf, drang mir tief zu Herzen.*, syntactic verb-object cooccurrence would identify the word pair (*Blick*, *zuwerfen*) correctly, without introducing spurious cooccurrences between all nouns and verbs in the sentence, i.e. (*Liebe*, *zuwerfen*), (*Liebe*, *sagen*), (*Treue*, *zuwerfen*), (*Treue*, *sagen*), etc.

[27] In her working definition of collocations, which combines aspects of both empricial collocations and multiword expressions, Bartsch explicitly requires collocates to stand in a direct syntactic relation (Bartsch 2004, 70f).

Note that all counts are based on instances of syntactic relations. Thus, the marginal frequency of *gentleman* is 3 although the word occurs only twice in the text: the first occurrences enters into two syntactic relations with the adjectives *stout* and *old*. Conversely, the marginal frequency of *lady* is 2 because its third occurrence (in the last sentence) is not modified by an adjective. Depending on the goal of a collocation study, it might make sense to include such non-modified nouns by introducing entries with "null modifiers" into the list of pair tokens. Some nouns might then be found to collocate strongly with the null modifier.

Bigrams can also be seen as syntactic cooccurrences, where the relation between the words is immediate precedence at the surface (i.e., linear ordering is considered to be part of the syntax of a language). Frequency signatures of bigrams according to syntactic cooccurrence are very similar to those calculated according to the procedure for (L0, R1) surface cooccurrence.

3.4 Comparison

Collocations according to surface cooccurrence have proven useful in corpus-linguistic and lexicographic research (cf. Sinclair 1991). They strike a balance between the restricted notion of syntactic cooccurrence (esp. when only a single type of syntactic relation is considered) and the very broad notion of textual cooccurrence. The number of recurrent word pairs extracted from a corpus is also more manageable than for textual cooccurrence. In this respect, syntactic cooccurrence is even more practical. A popular application of surface cooccurrence in computational linguistics are word space models of distributional semantics (Schütze 1998; Sahlgren 2006). As an alternative to the surface approach, Kilgarriff *et al.* (2004) collect syntactic collocates from different types of syntactic relations and display them as a *word sketch* of the node word.

Textual cooccurrence is easier to implement than surface cooccurrence, and more robust against certain types of non-randomness such as term clustering, especially when the textual units used are entire documents (cf. the discussion of non-randomness in Article 36). However, it tends to create huge data sets of recurrent word pairs that can be challenging even for powerful modern computers.

Syntactic cooccurrence separates collocations of different syntactic types, which are overlaid in frequency data according to surface cooccurrence, and discards many indirect and accidental cooccurrences. It should thus be easier to find suitable association measures to quantify the collocativity of word pairs. Evert (2004, 19) speculates that different measures might be appropriate for different types of syntactic relations. Syntactic cooccurrence is arguably most useful for the identification of multiword expressions, which are typically categorised according to their syntactic structure. However, it requires an accurate syntactic analysis of the source corpus, which will have to be performed with automatic tools in most cases. For prenominal adjectives, the analysis is fairly easy in English and German (Evert and Kermes 2003), while for German verb-object relations, it is extremely difficult to achieve satisfactory results: recent syntactic parsers achieve dependency F-scores of 70%–75% (Schiehlen 2004).[28] Outspoken advocates of syntactic cooccurrence include Daille (1994), Goldman *et al.* (2001), Bartsch (2004) and Evert (2004).

Leaving such practical and philosophical considerations aside, frequency signatures computed according to the different types of cooccurrence can disagree substantially for

[28] As an additional compliciation, it would often be more appropriate to consider the "logical" (or "deep") objects of verbs, i.e. grammatical subjects for verbs in passive voice and grammatical objects for verbs in active voice. Both *eat humble pie* and *much humble pie had to be eaten* should be identified as a verb-object cooccurrence of (*eat*, *humble pie*).

the same word pair. For examples, the frequency signatures of (*short*, *time*) in the Brown corpus are: $(16, 135, 457, 59710)$ for syntactic cooccurrence (of prenominal adjectives), $(27, 213, 1600, 1170811)$ for (L5, R5) surface cooccurrence, and $(32, 210, 1523, 52108)$ for textual cooccurrence within sentences.[29]

4 Simple association measures

4.1 Expected frequency

It might seem natural to use the cooccurrence frequency O as an association measure to quantify the strength of collocativity (e.g. Choueka 1988). This is not sufficient, however; the marginal frequencies of the individual words also have to be taken into account. To illustrate this point, consider the following example. In the Brown corpus, the bigram *is to* is highly recurrent. With $O = 260$ cooccurrences it is one of the most frequent bigrams in the corpus. However, both components are frequent words themselves: *is* occurs roughly 10,000 times and *to* roughly 26,000 times among 1 million word tokens.[30] If the words in this corpus were rearranged in completely random order, thereby removing all associations between cooccurring words, we would still expect to see the sequence *is to* approx. 260 times. The high cooccurrence frequency of *is to* does therefore not constitute evidence for a collocation; on the contrary, it indicates that *is* and *to* are not attracted to each other at all. The expected number of cooccurrences for a completely "uncollocational" word pair has been derived by the following reasoning: *to* occurs 26 times every 1,000 words on average. If there is no association between *is* and *to*, then each of the 10,000 instances of *is* in the Brown corpus has a chance of $26/1{,}000$ to be followed by *to*.[31] Therefore, we expect around $10{,}000 \times (26/1{,}000) = 260$ occurrences of the bigram *is to*, provided that there is indeed no association between the words. Of course, even in a perfectly randomised corpus there need not be exactly 260 cooccurrences: statistical calculations compute averages across large numbers of samples (formally called *expectations*), while the precise value in a corpus is subject to unpredictable random variation (see Article 36).[32]

The complete absence of association, as between words in a randomly shuffled corpus, is called *independence* in mathematical statistics. What we have calculated above is the *expected value* for the number of cooccurrences in a corpus of 1 million words, under the *null hypothesis* that *is* and *to* are independent. In analogy to the *observed frequency* O of a word pair, the expected value under the null hypothesis of independence is denoted E and referred to as the *expected frequency* of the word pair. Expected frequency serves as a reference point for the interpretation of O: the pair is only considered collocational if the observed cooccurrence frequency is substantially greater than the expected frequency,

[29] Prenominal adjectives were identified with the CQP query `[pos="JJ.*"] (",|and|or"? [pos="RB.*"]* [pos="JJ.*"]+)* [pos="NN.*"]` (in "traditional" matching mode), which gives a reasonably good approximation of a full syntactic analysis. The surface distance measure included all tokens (also punctuation etc.), so that the frequency counts could easily be obtained from CQP.

[30] The precise marginal frequencies in the BIGRAM data set are 9,775 for *is* and 24,814 for *to*, with a sample size of $N = 909{,}768$ tokens. This results in an expected frequency of $E = 266.6$ chance cooccurrences, slightly higher than the observed frequency $O = 260$.

[31] This is particularly clear if one assumes that the words have been rearranged in random order, as we have done above. In this case, each instance of *is* is followed by a random word, which will be *to* in 26 out of 1,000 cases.

[32] As a point of interest, we could also have calculated the expectation the other way round: each of the 26,000 instances of *to* has a chance of $10/1{,}000$ to be preceded by *is*, resulting in the same expectation of 260 cooccurrences.

$$\mathrm{MI} = \log_2 \frac{O}{E} \qquad \mathrm{MI}^k = \log_2 \frac{O^k}{E} \qquad \text{local-MI} = O \cdot \log_2 \frac{O}{E}$$

$$\text{z-score} = \frac{O - E}{\sqrt{E}} \qquad \text{t-score} = \frac{O - E}{\sqrt{O}} \qquad \text{simple-ll} = 2 \left(O \cdot \log \frac{O}{E} - (O - E) \right)$$

Figure 4: A selection of simple association measures.

$O \gg E$. Using the formal notation of Section 3, the marginal frequencies of (*is*, *to*) are $f_1 = 10{,}000$ and $f_2 = 26{,}000$. The sample size is $N = 1{,}000{,}000$ tokens, and the observed frequency is $O = 260$. Expected frequency is thus given by the equation $E = f_1 \cdot (f_2/N) = \frac{f_1 f_2}{N} = 260$. While the precise calculation of expected frequency is different for each type of cooccurrence, it always follows the basic scheme $f_1 f_2/N$.[33]

For textual and syntactic cooccurrence, the standard formula $E = f_1 f_2/N$ can be used directly. For surface cooccurrence, an additional factor k represents the total span size, i.e. $E = k f_1 f_2/N$. This factor is $k = 10$ for a symmetric span of 5 words (L5, R5), $k = 4$ for a span (L3, R1), and $k = 1$ for simple bigrams (L0, R1).

Intuitively, for every instance of the node w_1 there are k "slots" in which w_2 might cooccur with w_1. Under the null hypothesis, there is a chance of f_2/N to find w_2 in each one of these slots. With a total of $k f_1$ slots, we expect $k f_1 \cdot (f_2/N)$ cooccurrences. Note that according to this equation, expected frequency E is symmetric, i.e. it will be the same for (w_1, w_2) and (w_2, w_1), while O may be different for asymmetric spans.

This equation for E is only correct if (i) spans are not limited by sentence boundaries and (ii) the spans of different instances of w_1 do not overlap. Otherwise, the total number of slots for cooccurrences with w_2 is smaller than $k f_1$, and it would be necessary to determine its precise value by scanning the corpus. This procedure has to be repeated for every distinct first component w_1 among the word pairs in a data set. Fortunately, the error introduced by our approximation is usually small and unproblematic unless very large span sizes are chosen, so the simple equation above can be used in most cases.

4.2 Essential association measures

A *simple association measure* interprets observed cooccurrence frequency O by comparison with the expected frequency E, and calculates an *association score* as a quantitative measure for the attraction between two words. The most important and widely-used simple association measures are shown in Figure 4. In the following paragraphs, their mathematical background and some important properties will be explained.

The most straightforward and intuitive way to relate O and E is to use the ratio O/E as an association measure. For instance, $O/E = 10$ means that the word pair cooccurs 10 times more often than would be expected by chance, indicating a certain degree of collocativity.[34] Since the value of O/E can become extremely high for large sample size (because $E \ll 1$ for many word pairs), it is convenient and sensible to measure association on a (base-2) logarithmic scale. This measure can also be derived from information theory,

[33] Some of the differences have already been accounted for by computing frequency signatures in an appropriate way, as described in Section 3.

[34] Taking the difference $O - E$ might seem equally well justified at first, but turns out to be much less intuitive than the ratio measure: a word pair with $O = 100$ and $E = 10$ would be assigned a much higher score than a pair with $O = 10$ and $E = 1$, in contrast to the intuitively appealing view that both are 10 times more frequent than expected by chance.

where it is interpreted as the number of bits of "shared information" between two words and known as *(pointwise) mutual information* or simply *MI* (Church and Hanks 1990, 23). A MI value of 0 bits corresponds to a word pair that cooccurs just as often as expected by chance ($O = E$); 1 bit means twice as often ($O = 2E$), 2 bits mean 4 times as often, 10 bits about 1000 times as often, etc. A negative MI value indicates that a word pair cooccurs less often than expected by chance: half as often for -1 bit, a quarter as often for -2 bits, etc. Thus, negative MI values constitute evidence for a "repulsion" between two words, the pair forming an *anti-collocation*.[35]

The MI measure exemplifies two general conventions for association scores that all association measures should adhere to. (i) Higher scores indicate stronger attraction between words, i.e. a greater degree of collocativity. In particular, repulsion, i.e. $O < E$, should result in very low association scores. (ii) Ideally, an association measure should distinguish between *positive* association ($O > E$) and negative association ($O < E$), assigning positive and negative scores, respectively. A strong negative association would thus be indicated by a large negative value. As a consequence, the null hypothesis of independence corresponds to a score of 0 for such association measures. It is easy to see that MI satisfies both conventions: the more O exceeds E, the larger the association score will be; for $O = E$, the MI value is $\log_2 1 = 0$. Most, though not all association measures follow at least the first convention (we will shortly look at an important exception in the form of the simple-ll measure).

In practical applications, MI was found to have a tendency to assign inflated scores to low-frequency word pairs with $E \ll 1$, especially for data from large corpora.[36] Thus, even a single cooccurrence of two word types might result in a fairly high association score. In order to counterbalance this low-frequency bias of MI, various heuristic modifications have been suggested. The most popular one multiplies the denominator with O in order to increase the influence of observed cooccurrence frequency compared to the expected frequency, resulting in the formula $\log_2 \left(O^2 / E\right)$. Multiplication with O can be repeated to strengthen the counterbalancing effect, leading to an entire family of measures MI^k with $k \geq 1$, as shown in Figure 4. Common choices for the exponent are $k = 2$ and $k = 3$. Daille (1994) has systematically tested values $k = 2, \dots, 10$ and found $k = 3$ to work best for her purposes. An alternative way to reduce the low-frequency bias of MI is to multiply the entire formula with O, resulting in the *local-MI* measure. Unlike the purely heuristic MI^k family, local-MI can be justified by an information-theoretic argument (Evert 2004, 89) and its value can be interpreted as bits of information. Although not immediately obvious from its equation, local-MI fails to satisfy the first convention for association scores in the case of strong negative association: for fixed expected frequency E, the score reaches a minimum at $O = E / \exp(1)$ and then increases for smaller O. Local-MI distinguishes between positive and negative association, though, and satisfies both conventions if only word pairs with positive association are considered. The measures MI^k satisfy the first convention, but violate the second convention for all $k > 1$.[37]

It has been pointed out above that MI assigns high association scores whenever O exceeds E by a large amount, even if the absolute cooccurrence frequency is as low as $O = 1$ (and $E \ll 1$). In other words, MI only looks at what is known as *effect size* in statistics and does not take into account how much *evidence* the observed data provide. We will return to the distinction between effect-size measures and evidence-based measures in Section 6. Here, we introduce three simple association measures from the latter group.

[35] See the remarks on anti-collocations in Section 7.1.

[36] Imagine a word pair (w_1, w_2) where both words occur 10 times in the corpus. For a sample size of $N = 1{,}000$, the expected frequency is $E = 0.1$, for a sample size of $N = 1{,}000{,}000$, it is only $E = 0.0001$.

[37] In the case of independence, i.e. $O = E$, MI^k assigns the score $(k - 1) \cdot \log_2 O$.

A *z-score* is a standardised measure for the amount of evidence provided by a sample against a simple null hypothesis such as $O = E$ (see Article 36). In our case, the general rule for calculating z-scores leads to the equation shown in Figure 4.[38] Z-scores were first used by Dennis (1965, 69) as an association measure, and later by Berry-Rogghe (1973, 104). They distinguish between positive and negative association: $O > E$ leads to $z > 0$ and $O < E$ to $z < 0$. Z-scores can be interpreted by comparison with a standard normal distribution, providing theoretically motivated cut-off thresholds for the identification of "true collocations". An absolute value $|z| > 1.96$ is generally considered sufficient to reject the null hypothesis, i.e. to provide significant evidence for a (positive or negative) association; a more conservative threshold is $|z| > 3.29$. When used as an association measure, z-score tends to yield much larger values, though, and most word pairs in a typical data set are highly significant. For instance, 80% of all distinct word bigrams in the Brown corpus have $|z| > 1.96$, and almost 70% have $|z| > 3.29$.[39] Recent studies avoid standard thresholds and use z-scores only to rank word pairs or select n-best lists.

Authors using theoretical thresholds sometimes speak of "significant collocation" (Sinclair 1966, 418) or "significant word pairs" (e.g. Zinsmeister and Heid 2003). Since only a small proportion of the recurrent word pairs is rejected by such tests, the traditional concept of significance obviously has little meaning for collocation identification tasks. Therefore, use of these terms is strongly discouraged. Note that we still refer to z-score and related association measures as "significance measures" because of their mathematical background.

A fundamental problem of the z-score measure is the normal approximation used in its mathematical derivation, which is valid only for sufficiently high expected frequency E. While there is no clearly defined limit value,[40] the approximation becomes very inaccurate if $E < 1$, which is often the case for large sample sizes (e.g., 89% of all bigrams in the Brown corpus have $E < 1$).[41] Violation of the normality assumption leads to highly inflated z-scores and a low-frequency bias similar to the MI measure.[42] In order to avoid this low-frequency bias, various other significance measures have been suggested, based on more "robust" statistical tests. One possibility is the *t-score* measure, which replaces E in the denominator of z-score by O.[43] This measure has been widely used in computational

[38] For the mathematically inclined, and those who have read Article 36 carefully, this equation assumes that the observed frequency O is a binomially distributed random variable with sample size N and success probability $p = E/N$ under the null hypothesis (this ensures that the expected value of O equals E). The binomial distribution is then approximated by a normal distribution with mean $\mu = Np = E$ and variance $\sigma^2 = Np(1-p) \approx E$. For a normally distributed random variable, the z-score corresponding to an observed value O is given by $z = (O - \mu)/\sigma = (O - E)/\sqrt{E}$.

[39] Bigrams of adjacent words, using case-folded word forms and excluding punctuation except for sentence-ending punctuation as second component of the bigram. No frequency threshold was applied. Out of 368,210 distinct bigrams of this type found in the Brown corpus, 296,320 (= 80.5%) have $|z| > 1.96$ and 251,590 (= 68.3%) have $|z| > 3.29$. If data are restricted to the 93,205 recurrent bigrams ($f \geq 2$), there are still 73,431 items (= 78.8%) with $|z| > 1.96$ and 59,748 items (= 64.1%) with $|z| > 3.29$. In most cases, there is significant evidence for a positive association: 71,790 (= 77.0%) with $z > 1.96$ and 58,923 (= 63.2%) with $z > 3.29$.

[40] Article 36 suggests a very conservative threshold of $E > 9$, while many other authors feel that the normal approximation is sufficiently accurate for expected frequencies as small as $E = 1$.

[41] Bigrams of adjacent words as described above, with 327,998 out of 368,210 bigrams (= 89.1%) satisfying $E < 1$. For recurrent bigrams ($f \geq 2$), 63,551 out of 93,205 bigrams (= 68.2%) have $E < 1$.

[42] The low-frequency bias of the z-score measure is in part due to the fact that word pairs with higher cooccurrence frequency are less likely to satisfy $E < 1$ and hence generate inflated z-scores. Out of 10,339 bigrams with cooccurrence frequency $O \geq 10$ in the Brown corpus, only 1,997 (= 19.3%) have $E < 1$.

[43] Intuitively, it is straightforward to see how t-score reduces the low-frequency bias. For $E < 1$, the denominator of z-score becomes less than 1 so that the difference $O - E$ is inflated, while the denominator of t-score is always greater than (or equal to) 1.

lexicography following its introduction into the field by Church *et al.* (1991, Sec. 2.2). See Evert (2004, 82–83) for a criticism of its derivation from the statistical t test, which is entirely inappropriate for corpus frequency data.

Dunning (1993) advocated the use of likelihood-ratio tests, which are also more robust against low expected frequencies than z-score. For a simple measure comparing O and E, the likelihood-ratio procedure leads to the *simple-ll* equation in Figure 4.[44] It can be shown that simple-ll scores are always non-negative and violate both conventions for association scores. Because the underlying likelihood-ratio test is a *two-sided* test, the measure does not distinguish between $O \gg E$ and $O \ll E$, assigning high positive scores in both cases. This detail is rarely mentioned in publications and textbooks and may easily be overlooked.[45] A general procedure can be applied to convert a two-sided association measure like simple-ll into a one-sided measure that satisfies both conventions: association scores are calculated in the normal way and then multiplied with -1 for all word pairs with $O < E$. This procedure is applicable if association scores of the two-sided measure are always non-negative and high scores are assigned to strong negative associations. For the resulting transformed measure, significance is indicated by the absolute value of an association score, while positive and negative association are distinguished by its sign.[46]

Similar to the z-score measure, simple-ll measures significance (i.e. the amount of evidence against the null hypothesis) on a standardised scale, known as a chi-squared distribution with one degree of freedom, or χ_1^2 for short. Theoretically motivated cut-off thresholds corresponding to those for z-scores are $|ll| > 3.84$ and $|ll| > 10.83$, but the same reservations apply: many word pairs achieve scores far above these thresholds, so that they are not a meaningful criterion for the identification of "true collocations".

Article 36 gives detailed explanations of statistical concepts such as *significance*, *effect size*, *hypothesis test*, *one-sided* vs. *two-sided* test, *z-score* and *normal distribution* that have been used in this section.

4.3 Simple association measures in a nutshell

The preceding section has introduced a basic selection of simple association measures. These measures quantify the "attraction" between two words, i.e. their statistical association, by comparing observed cooccurrence frequency O against E, the expected frequency under the null hypothesis of independence (i.e. complete absence of association). E is important as a reference point for the interpretation of O, since two frequent words might cooccur quite often purely by chance. Most association measures follow the convention that higher association scores indicate stronger (positive) association. Many measures also differentiate between positive association ($O > E$), indicated by positives scores, and negative association ($O < E$), indicated by negative scores. Two-sided measures fail to make any distinction between positive and negative association, but can be converted into one-sided measures with an explicit test for $O > E$.

[44] The derivation of the simple-ll measure assumes O to follow a Poisson distribution with expected value E, a close approximation to the correct binomial distribution for large samples. See Appendix A.1 for details. Note the striking similarity to local-MI: apart from the additional term $(O - E)$, simple-ll uses a natural logarithm (log) instead of the base-2 logarithm ($\log_2$). This, and the constant factor of 2 are important for the interpretation of simple-ll scores according to a standardised scale of significance (viz., the χ_1^2 distribution).

[45] A typical example is Manning and Schütze (1999, 173), who also fail to mention the simple explicit form of log-likelihood for contingency tables shown in Figure 9.

[46] Note that the term "two-sided measure" is reserved for association measures derived from two-sided statistical tests. For instance, simple-MI is not a two-sided measure in this sense (although it fails to satisfy the first convention, too), and the general procedure for conversion to a one-sided measure cannot be applied.

The association measures listed in Figure 4 offer a number of different angles on collocativity that are sufficient for many purposes. Except for the heuristic MI^k family, all measures have theoretical motivations, allowing a meaningful interpretation of the computed association scores. As has been exemplified with the standard z-score thresholds, one should not put too much weight on such interpretations, though. Cooccurrence data do not always satisfy the assumptions made by statistical hypothesis tests, and heuristic measures may be just as appropriate.

Association measures can be divided into two general groups: measures of *effect size* (MI and MI^k) and measures of *significance* (z-score, t-score and simple-ll). The former ask the question "how strongly are the words attracted to each other?" (operationalised as "how much does observed cooccurrence frequency exceed expected frequency?"), while the latter ask "how much evidence is there for a positive association between the words, no matter how small effect size is?" (operationalised as "how unlikely is the null hypothesis that the words are independent?"). The two approaches to measuring association are not entirely unrelated: a word pair with large "true" effect size is also more likely to show significant evidence against the null hypothesis in a sample. However, there is an important difference between the two groups. Effect-size measures typically fail to account for sampling variation and are prone to a low-frequency bias (small E easily leads to spuriously high effect size estimates, even for $O = 1$ or $O = 2$), while significance measures are often prone to a high-frequency bias (if O is sufficiently large, even a small relative difference between O and E, i.e. a small effect size, can be highly significant).

Of the significance measures shown in Figure 4, simple-ll is the most accurate and robust choice. Z-score has a strong low-frequency bias because the approximations used in its derivation are not valid for $E < 1$, while t-score has been derived from an inappropriate hypothesis test. Nonetheless, t-score has proven useful for certain applications, especially the identification of certain types of multiword expressions (see Section 6.2). It has to be kept in mind that simple-ll is a two-sided measure and assigns high scores both to positive and negative associations. If only positive associations are of interest (as is the case for most studies), then word pairs with $O < E$ should be discarded. Alternatively, simple-ll can be transformed into a one-sided measure that satisfies both conventions for association scores (by multiplying scores with -1 if a word pair has $O < E$).

Association measures with a background in information theory take a different approach, which at first sight seems appropriate for the interpretation of collocations as mutually predictable word combinations (e.g. Sinclair 1966, 414). They ask the question to what extent the occurrences of a word w_1 determine the occurrences of another word w_2, and vice versa, based on the information-theoretic notion of mutual information (MI). Interestingly, different variants of MI lead to measures with entirely different properties: pointwise MI is a measure of effect size, while local-MI is very similar to simple-ll and hence has to be considered a measure of significance.

It is probably impossible to choose a single most appropriate association measure (cf. the discussion in Section 6). The recommended strategy is therefore to apply simple-ll, t-score and MI as proven association measures with well-understood mathematical properties, in order to obtain three entirely different perspectives on the cooccurrence data. MI should always be combined with a frequency threshold to counteract its low-frequency bias. As an example, and to illustrate the different properties of these association measures, Table 4 shows the collocates of *bucket* in the British National Corpus (following the case study in Section 2.2), according to simple-ll, t-score, MI without frequency threshold, and MI with an additional frequency threshold of $f \geq 5$. Table 5 gives a second example

collocate	f	f_2	simple-ll
water	184	37012	1083.18
a	590	2164246	449.30
spade	31	465	342.31
plastic	36	4375	247.65
size	42	14448	203.36
slop	17	166	202.30
mop	20	536	197.68
throw	38	11308	194.66
fill	37	10722	191.44
with	196	658584	171.78

collocate	f	f_2	t-score
a	590	2164246	15.53
water	184	37012	13.30
and	479	2616723	10.14
with	196	658584	9.38
of	497	3040670	8.89
the	832	6041238	8.26
into	87	157565	7.67
size	42	14448	6.26
in	298	1937966	6.23
record	43	29404	6.12

collocate	f	f_2	MI
fourteen-record	4	4	13.31
ten-record	3	3	13.31
multi-record	2	2	13.31
two-record	2	2	13.31
a-row	1	1	13.31
anti-sweat	1	1	13.31
axe-blade	1	1	13.31
bastarding	1	1	13.31
dippermouth	1	1	13.31
Dok	1	1	13.31

collocate	$f \geq 5$	f_2	MI
single-record	5	8	12.63
randomize	10	57	10.80
slop	17	166	10.03
spade	31	465	9.41
mop	20	536	8.57
oats	7	286	7.96
shovel	8	358	7.83
rhino	7	326	7.77
synonym	7	363	7.62
bucket	18	1356	7.08

Table 4: Collocates of *bucket* in the BNC according to the association measures simple-ll, t-score, MI, and MI with frequency threshold $f \geq 5$.

bigram	$f \geq 10$	f_1	f_2	simple-ll
of the	9702	34036	58451	13879.8
in the	6018	19615	58451	9302.3
it is	1482	8409	9415	5612.9
on the	2459	5990	58451	4972.9
United States	395	480	600	4842.6
it was	1338	8409	9339	4831.2
to be	1715	25106	6275	4781.1
had been	760	5107	2460	4599.8
have been	650	3884	2460	4084.0
has been	567	2407	2460	3944.9

bigram	$f \geq 10$	f_1	f_2	t-score
of the	9702	34036	58451	76.30
in the	6018	19615	58451	61.33
on the	2459	5990	58451	41.83
to be	1715	25106	6275	37.23
it is	1482	8409	9415	36.24
it was	1338	8409	9339	34.22
at the	1654	5032	58451	32.72
to the	3478	25106	58451	31.62
from the	1410	4024	58451	30.66
he was	1110	9740	9339	30.32

bigram	$f \geq 10$	f_1	f_2	MI
Hong Kong	11	11	11	16.34
gon na	16	16	16	15.80
Viet Nam	14	16	14	15.80
Simms Purdew	12	16	12	15.80
Pathet Lao	10	10	17	15.71
El Paso	10	19	11	15.41
Lo Shu	21	21	21	15.40
Puerto Rico	21	24	21	15.21
unwed mothers	10	12	26	14.83
carbon tetrachloride	18	30	19	14.81

bigram	$f \geq 50$	f_1	f_2	MI
Los Angeles	50	51	50	14.12
Rhode Island	100	105	175	12.27
Peace Corps	55	171	109	11.39
per cent	146	371	155	11.17
United States	395	480	600	10.29
President Kennedy	54	374	156	9.72
years ago	138	793	246	9.33
fiscal year	58	118	701	9.32
New York	303	1598	309	9.12
United Nations	51	480	175	9.11

Table 5: Most strongly collocated bigrams in the Brown corpus according to the association measures simple-ll, t-score, MI with frequency threshold $f \geq 10$, and MI with frequency threshold $f \geq 50$.

for word bigrams in the Brown corpus (excluding punctuation).[47] Obviously, simple-ll and especially t-score focus on frequent grammatical patterns like *of the* or *to be*. More interesting bigrams can only be found if separate lists are generated for each part-of-speech combination.[48] The top collocations according to MI, on the other hand, tend to be proper names and other fixed combinations. Their cooccurrence frequency is often close to the applied frequency threshold.

5 Statistical association measures

The simple association measures introduced in Section 4 are convenient and offer a range of different perspectives on collocativity. However, two serious shortcomings make this approach unsatisfactory from a theoretical point of view and may be problematic for certain types of applications. The first of these problems is most easily explained with a worked example. In a corpus of about a million words, you might find that the bigrams A = *the Iliad* and B = *must also* both occur $O = 10$ times, with the same expected frequency $E = 1$. Therefore, any simple measure will assign the same association score to both bigrams. However, bigram A is a combination of a very frequent word (*the* with, say, $f_1 = 100{,}000$) and an infrequent word (*Iliad* with $f_2 = 10$), while B combines two words of intermediate frequency (*must* and *also* with $f_1 = f_2 = 1{,}000$). Using the formula $E = f_1 f_2 / N$ from Section 4.1, you can easily check that the expected frequency is indeed $E = 1$ for both bigrams.[49] While O exceeds E by the same amount for *the Iliad* as for *must also*, it is intuitively obvious that bigram A is much more strongly connected than bigram B. In particular, $O = 10$ is the highest cooccurrence frequency that can possibly be observed for these two words (since $O \leq f_1, f_2$): every instance of *Iliad* in the corpus is preceded by an instance of *the*. For bigram B, on the other hand, the words *must* and *also* could have cooccurred much more often than 10 times. One might argue that A should therefore obtain a higher association score than B, at least for certain applications.

> This example points to a property of collocations that most current approaches do not take into account: they may be asymmetric, i.e. word w_2 might be strongly predicted by w_1, but w_1 only weakly by w_2 (or vice versa, as in the example of *the Iliad*). Such asymmetries play an important role in the node–collocate view: *Iliad* should not rank highly as a collocate of *the*, while *the* is clearly a strong collocate of *Iliad*. These considerations also explain to a certain extent why measures with a high-frequency bias like simple-ll seem to produce more "satisfactory" lists of collocations in Table 4 than measures with a low-frequency bias like MI: simple-ll will rank *the* fairly highly as a collocate for *Iliad* since there can be no collocates that cooccur more often, but it will prefer more frequent words than *Iliad* as collocates for *the*. It is less clear to what extent asymmetries are relevant for the unit view of collocations, and whether association should be as closely linked to predictability in this case. Here, we will follow the mainstream approach of symmetric association measures, but see the remarks on asymmetric measures in Section 7.1.

The second limitation of simple association measures is of a more theoretical nature. We made use of statistical concepts and methods to define measures with a meaningful

[47] This example is based on the BIGRAM data set introduced in Section 2.2.

[48] As has been done for simple-ll in Table 3.

[49] While this is an invented example where the precise frequency counts have been kept artificially simple, it is by no means unrealisitic. In fact, the example is based on the BIGRAM data set extracted from the Brown corpus, with a sample size of $N = 909{,}768$. The frequency signatures of the two bigrams in this data set are: $f = 14, f_1 = 69{,}349, f_2 = 14 \to E = 1.07$ for *the Iliad* vs. $f = 13, f_1 = 1{,}000, f_2 = 936 \to E = 1.03$ for *must also*. Another "balanced" bigram with the same observed and expected frequency as *the Iliad* is *can say*: $f = 14, f_1 = 1{,}997, f_2 = 465 \to E = 1.02$.

	w_2	$\neg w_2$	
w_1	O_{11}	O_{12}	$= R_1$
$\neg w_1$	O_{21}	O_{22}	$= R_2$
	$= C_1$	$= C_2$	$= N$

	w_2	$\neg w_2$
w_1	$E_{11} = \frac{R_1 C_1}{N}$	$E_{12} = \frac{R_1 C_2}{N}$
$\neg w_1$	$E_{21} = \frac{R_2 C_1}{N}$	$E_{22} = \frac{R_2 C_2}{N}$

Figure 5: General form of the contingency table of observed frequencies with row and column marginals (left panel), and contingency table of expected frequencies under the null hypothesis of independence (right panel).

interpretation, but did not apply the procedures with full mathematical rigour.[50] In statistical theory, measures of association and tests for the independence of events are always based on a cross-classification of a random sample of certain items. An appropriate representation of cooccurrence frequency data in the form of *contingency tables* is described in Section 5.1, with different rules for each type of cooccurrence. Then several widely used statistical association measures are introduced in Section 5.2.

We will see in Section 6 that simple association measures often give close approximations to the more sophisticated association measures introduced below. Therefore, they are sufficient for many applications, so that the computational and mathematical complexities of the rigorous statistical approach can be avoided.

5.1 Contingency tables

A rigorous statistical approach to measuring association is based on contingency tables representing the cross-classification of a set of items. Such tables naturally take marginal frequencies into account, unlike a simple comparison of O against E. As a first step, we have to define the set of cooccurrence items in a meaningful way, which is different for each type of cooccurrence. Then a separate contingency table is calculated for every word pair (w_1, w_2), using the presence of w_1 and w_2 in each cooccurrence item as factors for the cross-classification.

The resulting contingency table (left panel of Figure 5) has four *cells*, representing the items containing both w_1 and w_2 (O_{11}, equivalent to the observed cooccurrence frequency O), the items containing w_1 but not w_2 (O_{12}), the items containing w_2 but not w_1 (O_{21}), and the items containing neither of the two words (O_{22}). These *observed frequencies* add up to the total number of items or *sample size*, since every item has to be classified into exactly one cell of the table. The row and column sums, also called *marginal frequencies* (as they are written in the margins of the table), play an important role in the statistical analysis of contingency tables. The first row sum R_1 corresponds to the number of cooccurrence items containing w_1, and is therefore usually equal to f_1 (except for surface cooccurrence, see below), while the first column sum C_1 is equal to f_2. This equivalence explains the name "marginal frequencies" for f_1 and f_2.

[50] In particular, the sampling variation of the expected frequency E, which is a sample estimate rather than the "true" population value, is neglected. Moreover, we did not specify a precise sampling model, simply assuming O to have a binomial distribution with expectation E under the null hypothesis.

	$*\mid w_2$	$*\mid\neg w_2$	
$w_1\mid *$	O_{11}	O_{12}	$= f_1$
$\neg w_1\mid *$	O_{21}	O_{22}	
	$= f_2$		$= N$

	$*\mid$gent.	$*\mid\neg$gent.	
young$\mid *$	1	2	= 3
$\neg$young$\mid *$	2	4	
	= 3		= 9

Figure 6: Contingency table of observed frequencies for syntactic cooccurrence, with concrete example for the word pair (*young*, *gentleman*) and the data in Figure 3 (right panel).

As in the case of simple association measures, the statistical analysis of contingency tables is based on a comparison of the observed frequencies O_{ij} with expected frequencies under the null hypothesis that the factors defining rows and columns of the table are statistically independent (which is the mathematical equivalent of the intuitive notion of independence between w_1 and w_2 introduced in Section 4.1). In contrast to the simple approach, we are not only interested in the expected number of cooccurrences of w_1 and w_2, but have to compute expected frequencies for all four cells of the contingency table, according to the equations shown in the right panel of Figure 5. Note that $O_{11} = O$ and $E_{11} = E$, so statistical contingency tables are a genuine extension of the previous approach.[51] The statistical association measures introduced in Section 5.2 below are formulated in terms of observed frequencies O_{ij} and expected frequencies E_{ij}, the marginals R_i and C_j, and the sample size N. This standard notation follows Evert (2004) and allows equations to be expressed in a clean and readable form.

The definition of appropriate contingency tables is most straightforward for syntactic cooccurrence. The pair tokens on the right of Figure 3 can naturally be interpreted as a set of cooccurrence items. If the first word is w_1, an item is classified into the first row of the contingency table for the pair (w_1, w_2), otherwise it is classified into the second row. Likewise, the item is classified into the first column if the second word is w_2 and into the second column if it is not. This procedure is illustrated in the left panel of Figure 6. The first row sum R_1 equals the total number of cooccurrence items containing w_1 as first element, and the first column sum equals the number of items containing w_2 as second element. This corresponds to the definition of f_1 and f_2 for syntactic cooccurrence in Section 3.3. The example in the right panel of Figure 6 shows a contingency table for the word pair (*young*, *gentleman*) obtained from the sample of adjective-noun cooccurrences in Figure 3. Since there are nine instances of adjectival modification of nouns in this toy corpus, the sample size is $N = 9$. There is one cooccurrence of *young* and *gentleman* ($O_{11} = 1$), two items where *gentleman* is modified by another adjective ($O_{12} = 2$), two items where *young* modifies another noun ($O_{21} = 2$), and four items that contain neither the adjective *young* nor the noun *gentleman* ($O_{22} = 4$).

For textual cooccurrence, Figure 2 motivates the definition of cooccurrence items as instances of textual units. In this example, each item corresponds to a sentence of the corpus. The sentence is classified into the first row of the contingency table if it contains one or more instances of w_1 and into the second row otherwise; it is classified into the first column if it contains one or more instances of w_2 and into the second column otherwise (see Figure 7). Note that no distinction is made between single and multiple occurrence

[51] With the equalities $R_1 = f_1$ and $C_1 = f_2$, we find that $E_{11} = R_1C_1/N = f_1f_2/N = E$. Figure 8 shows that the approximate equality of E_{11} and E also holds for surface cooccurrence: $E_{11} = R_1C_1/N \approx kf_1f_2/N \approx E$.

	$w_2 \in S$	$w_2 \notin S$	
$w_1 \in S$	O_{11}	O_{12}	$= f_1$
$w_1 \notin S$	O_{21}	O_{22}	
	$= f_2$		$= N$

	over $\in S$	over $\notin S$	
hat $\in S$	1	2	= 3
hat $\notin S$	1	1	
	= 2		= 5

Figure 7: Contingency table of observed frequencies for textual cooccurrence, with concrete example for the word pair (*hat*, *over*) and the data in Figure 2 (right panel).

	w_2	$\neg w_2$	
$near(w_1)$	O_{11}	O_{12}	$\approx k \cdot f_1$
$\neg\, near(w_1)$	O_{21}	O_{22}	
	$= f_2$		$= N - f_1$

	roll	¬roll	
near(hat)	2	18	= 20
¬ *near*(hat)	1	87	
	= 3		= 108

Figure 8: Contingency table of observed frequencies for surface cooccurrence, with concrete example for *roll* as a collocate of the node *hat* according to Figure 1 (right panel).

of w_1 or w_2 in the same sentence. Again, the first row and column sums correspond to the marginal frequencies f_1 and f_2 as defined in Section 3.2. The right panel of Figure 7 shows a contingency table for the word pair (*hat*, *over*), based on the example in Figure 2. With five sentences in the toy corpus,[52] sample size is $N = 5$. One of the sentences contains both *hat* and *over* ($O_{11} = 1$), two sentences contain *hat* but not *over* ($O_{12} = 2$), one sentence contains *over* but not *hat* ($O_{21} = 1$), and one sentence contains neither of the two words ($O_{22} = 1$).[53]

The statistical interpretation of surface cooccurrence is less straightforward than for the other two types. The most sensible definition identifies cooccurrence items with the relevant word tokens in the corpus, but excluding instances of the node word w_1, for which no meaningful cross-classification is possible.[54] Each item, i.e. word token, is then classified into the first row of the contingency table if it cooccurs with the node word w_1, i.e. if it falls into one of the collocational spans around the instances of w_1; it is classified into the second row otherwise. The item is classified into the first column of the table if it is an instance of the targeted collocate w_2, and into the second column otherwise. The procedure is illustrated in Figure 8, with a concrete example for the data of Figure 1 shown in the right panel. This toy corpus consists of 111 word tokens (excluding punctuation). Subtracting the three instances of the node word *hat*, we obtain a sample size of $N = 108$. Of the 108 cooccurrence items, 20 fall into the collocational spans around instances of *hat*,

[52] Recall that semicolons were interpreted as sentence boundaries.

[53] The marginal frequency of *over* was defined as $f_2 = 2$ even though there are three instances in the text sample, since *over* occurs only in two of the five sentences. This corresponds to the equal status of single and multiple occurrences in the cross-classification procedure.

[54] Note that the set of cooccurrence items differs slightly between word pairs (more precisely, between word pairs with different first components). This is in contrast to syntactic and textual cooccurrence, where a fixed set of items is used for all word pairs and only the cross-classifying factors are different.

so that the first row sum is $R_1 = 20$. Two of these items are cooccurrences of *hat* and *roll* ($O_{11} = 2$), and the remaining 18 items are classified into the second cell ($O_{12} = 18$). The 88 items outside the collocational spans are classified analogously: there is one instance of the collocate *roll* ($O_{21} = 1$), and all other items are assigned to the last cell of the table ($O_{22} = 87$).

> Note that the first row sum R_1 does not correspond to the occurrence frequency f_1 of the node word w_1, but rather to the total number of word tokens that cooccur with w_1, i.e. the number of tokens in the collocational spans around w_1. This value is approximately equal to $k \cdot f_1$, where k is the "standard" size of a single span, but R_1 will be smaller if different spans overlap or spans are truncated by sentence boundaries. In the example shown in Figure 8, $R_1 = 20$ is quite different from the approximation $k \cdot f_1 = 8 \cdot 3 = 24$ because the first span is truncated by a sentence boundary on the right-hand side.
>
> If we had calculated a contingency table for *hat* and *over* based on the same text sample, the two cooccurrences would also have resulted in $O_{11} = 2$, even though they are in the same collocational span. This is in marked contrast to textual cooccurrence, while for syntactic cooccurrence, the issue of multiple occurrences within a single "unit" does not arise.
>
> An important property of surface cooccurrence is that the resulting contingency tables are asymmetric: the table for w_1 as a collocate of the node w_2 (i.e. the word pair (w_2, w_1)) may be substantially different from the one for w_2 as a collocate of w_1 (i.e. the word pair (w_1, w_2)). This effect can be very pronounced if the marginal frequencies of w_1 and w_2 differ considerably. Simple association measures, which only take the top left cell of the table into account (i.e. O_{11} and E_{11}), gloss over the difference between the two contingency tables, since E_{11} is the same for both tables (except for minor differences due to overlaps and truncations of the collocational spans). Statistical association measures, on the other hand, will be affected to varying degrees.

For syntactic and textual cooccurrence, the contingency tables can be calculated directly from frequency signatures (O, f_1, f_2, N) that have been obtained as described in Sections 3.2 and 3.3, using the following transformation equalities:

$$\begin{aligned} O_{11} &= O & \qquad O_{12} &= f_1 - f \\ O_{21} &= f_2 - f & \qquad O_{22} &= N - f_1 - f_2 + f \end{aligned}$$

The use of frequency signatures in combination with the equalities above is usually the most practical and convenient implementation of contingency tables.[55] Tables for surface cooccurrence cannot be simplified in the same way, and it is recommended to calculate them by the explicit cross-classification procedure explained above.

> Since surface cooccurrence is most often combined with a node–collocate view of collocations, the implementation can be optimised by calculating the row sums for the fixed node word w_1 first and then filling in the cells O_{ij} based on cooccurrence frequency $O = O_{11}$ and marginal frequency $f_2 = C_1$ for each collocate w_2. A rough approximation to the correct contingency tables, corresponding to $E \approx kf_1f_2/N$ in Section 4.1, can be obtained by using the equalities above and replacing f_1 with $k \cdot f_1$, where k is the total size of the collocational span (see Section 4.1). This approximation can be quite inaccurate if spans overlap or are truncated by sentence boundaries, though.

[55] In particular, marginal frequencies can be shared by different word pairs with the same first or second component and do not have to be recomputed for every word pair.

5.2 Selected measures

Statistical association measures assume that the set of cooccurrence items is a random sample from a large population (representing an extensional definition of language as the set of all utterances that have been or can be produced, cf. Article 36) and attempt to draw inferences about this population. Like simple measures, they can be divided into the general groups of effect-size and significance measures.

Effect-size measures aim to quantify how strongly the words in a pair are attracted to each other, i.e. they measure statistical association between the cross-classifying factors in the contingency table. Liebetrau (1983) gives a comprehensive survey of such association coefficients and Evert (2004, 54–58) discusses their mathematical properties. Coefficients describe properties of a population without taking sampling variation into account. They can be used as association measures in a straightforward way if this fact is ignored and the observed frequencies are taken as direct estimates[56] for the corresponding population probabilities. As a result, effect-size measures tend to be unreliable especially for low-frequency data.

MI is the most intuitive association coefficient, comparing observed cooccurrence frequency against the value expected under the null hypothesis of independence. The equation shown in Figure 4 is also meaningful as a statistical association measure, where it should more precisely be written $\log_2(O_{11}\,/\,E_{11})$. Two other association coefficients are the (logarithmic) *odds ratio* (Blaheta and Johnson 2001, 56) and the *Dice coefficient* (Smadja *et al.* 1996), shown in Figure 9. The odds-ratio measure satisfies both conventions for association scores, with a value of 0 corresponding to independence and high positive values indicating strong positive association. Its interpretation is less intuitive than that of MI, though, and it has rarely been applied to collocations. The Dice coefficient does not adhere to the second convention, as it does not assume a well-defined value in the case of independence. It cannot be used to identify word pairs with strong negative association, but is well-suited for rigid combinations such as fixed multiword units (Smadja *et al.* 1996; Dias *et al.* 1999).

The odds ratio quantifies association strength by comparing the column ratios $O_{11}\,/\,O_{21}$ and $O_{12}\,/\,O_{22}$ in the contingency table (or, equivalently, the row ratios). From the right panel of Figure 5 it is obvious that the two ratios should be equal under the null hypothesis of independence: $E_{11}\,/\,E_{21} = E_{12}\,/\,E_{22} = R_1\,/\,R_2$. In the case of a positive association, i.e. $O_{11} > E_{11}$, we will find that the first column ratio $O_{11}\,/\,O_{21}$ is larger than the second ratio $O_{12}\,/\,O_{22}$, while it is smaller for a negative association. This motivates the fraction $\frac{O_{11}/O_{21}}{O_{12}/O_{22}} = \frac{O_{11}O_{22}}{O_{12}O_{21}}$ as an association coefficient. It is known as “odds ratio” because the column ratios are sometimes called “odds”, as in gambling. The logarithm of the odds ratio satisfies both conventions for association scores, with a value of 0 in the case of independence.

The odds-ratio equation is often amended by adding $\frac{1}{2}$ to each observed frequency, in order to avoid undefined values for contingency tables with zero entries and improve its mathematical properties as a statistical estimator (Agresti 2002, 71). Figure 9 shows the amended version of the odds-ratio measure. The odds ratio does not have the same intuitive interpretation as MI and is mostly appreciated for its statistical properties, which fit in well with the random sampling model for contingency tables. Consequently, it has rarely been applied to collocations, with Blaheta and Johnson (2001) as a notable exception.

[56] Such direct estimates are called *maximum-likelihood estimates* in mathematical terminology.

$$\text{chi-squared} = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad \text{chi-squared}_{\text{corr}} = \frac{N\big(|O_{11}O_{22} - O_{12}O_{21}| - N/2\big)^2}{R_1 R_2 C_1 C_2}$$

$$\text{log-likelihood} = 2 \sum_{ij} O_{ij} \log \frac{O_{ij}}{E_{ij}} \qquad \text{average-MI} = \sum_{ij} O_{ij} \cdot \log_2 \frac{O_{ij}}{E_{ij}}$$

$$\text{Dice} = \frac{2O_{11}}{R_1 + C_1} \qquad \text{odds-ratio} = \log \frac{\big(O_{11} + \frac{1}{2}\big)\big(O_{22} + \frac{1}{2}\big)}{\big(O_{12} + \frac{1}{2}\big)\big(O_{21} + \frac{1}{2}\big)}$$

Figure 9: Some widely used statistical association measures.

The Dice coefficient focuses on cases of very strong association rather than the comparison with independence. It can be interpreted as a measure of predictability, based on the ratios O_{11}/R_1 (the proportion of instances of w_1 that cooccur with w_2) and O_{11}/C_1 (the proportion of instances of w_2 that cooccur with w_1). The two ratios are averaged by calculating their *harmonic mean*, leading to the equation in Figure 9. Unlike the more familiar arithmetic mean, the harmonic mean only assumes a value close to 1 (the largest possible Dice score) if there is a strong prediction in both directions, from w_1 to w_2 and vice versa. The association score will be much lower if the relation between the two words is asymmetric.

Statistical significance measures are based on the same types of hypothesis tests as the simple measures in Section 4.2, viz. chi-squared tests (as a generalisation of z-scores) and likelihood-ratio tests. Unsurprisingly, there is no counterpart for t-score, which was based on an inappropriate test and hence cannot be translated into a rigorous statistical measure. The *chi-squared* measure adds up squared z-scores for all cells of the contingency table ($\sum_{ij}$ indicates summation over all four cells, i.e. over indices $ij = 11, 12, 21, 22$).[57] The normal approximation implicit in the z-scores becomes inaccurate if any of the expected frequencies E_{ij} are small, and chi-squared exhibits a low-frequency bias similar to the z-score measure. A better approximation is obtained by applying *Yates' continuity correction* (cf. DeGroot and Schervish 2002, Sec. 5.8). The continuity-corrected version is often written in the compact form shown as chi-squared$_{\text{corr}}$ in Figure 9, without explicit reference to expected frequencies E_{ij}. Chi-squared is a two-sided measure because the squared values are always positive. It can be transformed into a one-sided measure using the general procedure introduced in Section 4.2. Chi-squared is often abbreviated X^2, the symbol used for the chi-squared test statistic in mathematical statistics.

The *log-likelihood* measure (Dunning 1993) is a straightforward extension of simple-ll, replacing the term $O - E$ by a summation over the remaining three cells of the contingency table. It is a two-sided measure and is sometimes abbreviated G^2 in analogy to X^2. Interestingly, the association scores of log-likelihood, simple-ll and chi-squared are all interpreted against the same scale, a χ^2_1 distribution (cf. Section 4.2).[58] Mathematicians generally agree that the most appropriate significance test for contingency tables is *Fisher's exact test* (Agresti 2002, 91–93), which was put forward by Pedersen (1996) as

[57] Note that adding raw z-scores does not make sense, as positive and negative values would cancel out.

[58] The values of the one-sided z-score measure are on a comparable scale as well. Under the null hypothesis of independence, squared z-scores follow the same χ^2_1 distribution as the other three measures. Although all these scores measure the same quantity (viz., the amount of evidence against the null hypothesis of independence) on comparable scales *in principle*, they often compute strikingly different values in practice, as will be demonstrated in Section 6.1.

an alternative to the log-likelihood measure. Unlike chi-squared and likelihood-ratio tests, this exact test does not rely on approximations that may be invalid for low-frequency data. Fisher's test can be applied as a one-sided or two-sided measure and provides a useful reference point for the discussion of other significance measures. However, it is computationally expensive and a sophisticated implementation is necessary to avoid numerical instabilities (Evert 2004, 93). Section 6.1 shows that log-likelihood provides an excellent approximation to association scores computed by Fisher's test, so there is little reason to use the complicated and technically demanding exact test. The information-theoretic measure *average-MI* is identical to log-likelihood (except for a constant factor) and need not be discussed further here.

The appropriate information-theoretic measure for the statistical cross-classification model is *average-MI*, which can also be understood as an extension of local-MI to the full contingency table. It has repeatedly been noted as a curiosity that the average-MI formula is almost identical to that of log-likelihood (Dunning 1998, 75–76), but a thorough discussion of the relation between information theory and likelihood-ratio tests is far beyond the scope of this article. Pointwise MI can also be defined in a meaningful way in the cross-classification model and leads to the familiar MI measure that has been interpreted as an association coefficient above.

It is important to understand that although MI and average-MI are based on the same information-theoretic concepts, they measure entirely different aspects of association. Pointwise MI tells us how much information each individual occurrence of w_1 provides about nearby occurrences of w_2, and vice versa, making it a measure of effect size (how "tightly linked" w_1 and w_2 are). Average-MI, on the other hand, tells us how much information the distribution of w_1 in the corpus provides about the distribution of w_2, and vice versa. It is thus related to significance measures, since "shared" information between the distributions corresponds to deviation from the null hypothesis of independence. While the precise mathematical reasons for such interconnections may be difficult to grasp, they are obvious from the equations of the association measures in Figures 4 and 9 (by comparison with MI as an effect-size measure and log-likelihood as a significance measure).

In this section, the most commonly used statistical association measures have been presented and are summarised in Figure 9. Log-likelihood is by far the most popular significance measure and has found widespread use especially in the field of computational linguistics. The chi-squared measure is known to have a low-frequency bias similar to z-score and MI, which can be reduced (but not eliminated completely) by applying Yates' continuity correction. Of the effect-size measures, MI is the most well-known one and has a very intuitive interpretation. Its formula is identical to the simple association measure MI in Figure 4. The Dice coefficient enjoys a certain popularity, too, especially for the identification of rigid multiword expressions. The odds-ratio measure offers certain mathematical advantages, but its scores are difficult to interpret and it has rarely been used as an association measure (but see the remarks on merging effect-size and significance approaches in Section 7.1).

Note that simple association measures can also be computed from the full contingency tables, replacing O by O_{11} and E by E_{11} in the equations given in Figure 4. This shows clearly that many simple measures can be understood as a simplified version (or approximation) of a corresponding statistical measure. A more comprehensive list of association measures with further explanations can be found in Evert (2004, Sec. 3) and online at

`http://www.collocations.de/AM/`

Both resources describe simple as well as statistical association measures, using the notation for contingency tables introduced in this section and summarised in Figure 5.

6 Finding the right measure

The twelve equations in Figures 4 and 9 represent just a small selection of the many association measures that have been suggested and used over the years. Evert (2004) discusses more than 30 different measures, Pecina (2005) lists 57 measures, and new measures and variants are constantly being invented. While some measures have been established as de-facto standards, e.g. log-likelihood in computational linguistics, t-score and MI in computational lexicography, there is no ideal association measure for all purposes. Different measures highlight different aspects of collocativity and will hence be more or less appropriate for different tasks: the n-best lists in Tables 4 and 5 are a case in point.

> The suitability of an association measure also depends on many other parameters such as cooccurrence type, frequency threshold, language, genre, domain, corpus size, etc. Therefore, one should never become so daunted by the multitude of options as to fall back on a standard choice.

The goal of this section is to help researchers choose a suitable association measure (or set of measures) for their study. While the primary focus is on understanding the characteristic properties of the measures presented in this article and the differences between them, the methods introduced below can also be applied to other association measures, allowing researchers to make an informed choice from the full range of options.

6.1 Mathematical arguments

Theoretical discussions of association measures are usually concerned with their mathematical derivation:[59] the assumptions of the underlying model, the theoretical quantity to be measured, the validity and accuracy of the procedures used (especially if approximations are involved), and general mathematical properties of the measures (such as a bias towards low- or high-frequency word pairs). A first step in such discussions is to collect association measures with the same theoretical basis into groups. Measures within each group can often be compared directly with respect to their mathematical properties (since ideally they should measure the same theoretical quantity and hence lead to the same results), while different groups can only be compared at a general and rather philosophical level (does it make more sense to measure effect size or significance of association?).

As has already been mentioned in Sections 4 and 5, the association measures introduced in this article fall into two major groups: *effect-size measures* (MI, Dice, odds-ratio) and *significance measures* (z-score, t-score, simple-ll, chi-squared, log-likelihood). The choice between these two groups is largely a philosophical issue: one cannot be considered "better" than the other. Instead, they highlight different aspects of collocativity and are plagued by different types of mathematical problems.

> While the information-theoretic measures are based on a distinct mathematical theory, they do not form a separate group: MI is identical to the effect-size measure of the same name, while average-MI is fully equivalent to the significance measure log-likelihood (and local-MI is strikingly similar to simple-ll). Therefore, we will not consider these measures in the following discussion. The heuristic MI^k family of measures is difficult to place, lacking a well-defined theoretical background. However, their similarity to MI puts them at least close to the group of effect-size measures.

[59] Such discussions are thus limited to measures with sound statistical underpinnings. They cannot be applied to heuristic measures or, e.g., the invalid derivation of the t-score measure from t tests by Church *et al.* (1991).

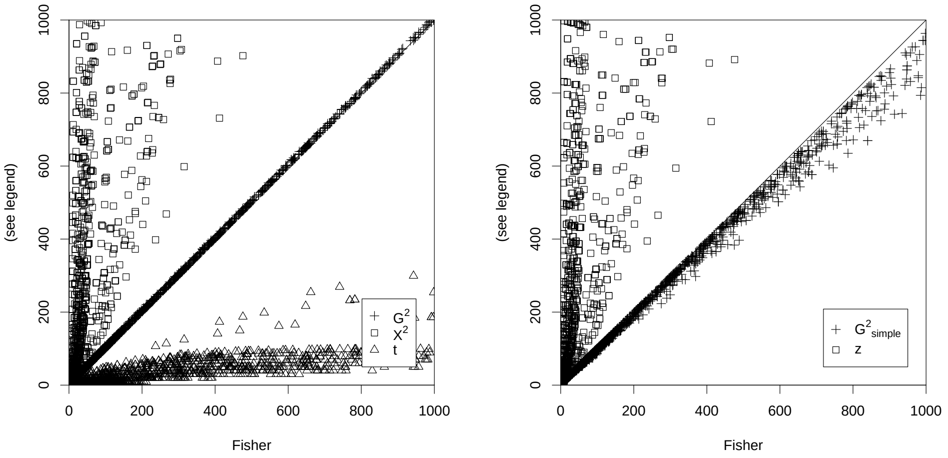


Figure 10: Direct comparison of association scores on synthetic data set, using Fisher's exact test as reference point (scores are transformed to χ_1^2 scale).

Significance measures are particularly amenable to mathematical discussions, since in principle they attempt to measure the same theoretical quantity: the amount of evidence provided by a sample against the null hypothesis of independence. Moreover, chi-squared, log-likelihood and simple-ll use the same scale (the χ_1^2 distribution), so that their scores are immediately comparable. While z-score and t-score use a scale based on the normal distribution, their scores can easily be transformed to the χ_1^2 scale. The long-standing debate in mathematical statistics over appropriate significance tests for contingency tables has not completely been resolved yet (see Yates 1984), but most researchers consider Fisher's exact test to be the most sensible and accurate measure of significance (Yates 1984, 446). We will therefore use it as a reference point for the comparison of association measures in the significance group. Fisher's test calculates so-called p-values (cf. Article 36), which are also transformed to the χ_1^2 scale for the comparison. The scatterplots in Figure 10 compare association scores calculated by various significance measures with those of Fisher's exact test, using a synthetic data set in which cooccurrence and marginal frequencies have been varied systematically.[60] The log-likelihood measure (G^2) and to some extent also simple-ll (G^2_{simple}) give an excellent approximation to Fisher's test, as all data points are close to the diagonal. Chi-squared and z-score overestimate significance drastically (points far above diagonal), while t-score underestimates significance to a similar degree (points far below diagonal).[61]

For effect-size measures, there is no well-defined theoretical quantity that would allow a direct comparison of their scores (e.g. with scatterplots as in Figure 10). Numerous coefficients have been suggested as measures of association strength in the population, but statisticians do not agree on a theoretically satisfactory choice (see e.g. Liebetrau 1983).[62] A common mathematical property of effect-size measures is the use of direct estimates

[60] The data points of an ideal significance measure, which calculates the same scores as Fisher's test, should lie on or close to the main diagonal in such plots.

[61] More detailed analyses show that the overestimation is particularly strong for low-frequency data, explaining the observed low-frequency bias of z-score and chi-squared.

[62] Evert (2004, 54–58) compares several of these coefficients by calculating their (asymptotic) values under certain special conditions such as independence, total association where w_1 is always accompanied by w_2, minimal deviation from independence, etc. The comparison provides support for MI if the focus is on deviation from independence and for Dice if the focus is on total association, but once again no definitive answer can be given.

that do not take sampling variation into account. As a result, association scores tend to become unreliable for low-frequency data. This effect is particularly severe for MI, odds-ratio and similar measures that compare observed and expected frequency, since $E_{11} \ll 1$ for many low-frequency word pairs.[63] Extending effect-size measures with a correction for sampling variation is a current topic of research and is expected to bridge the gap between the effect-size and significance groups (see Section 7.1).

It should be emphasised that despite their mathematical shortcomings, measures such as chi-squared and t-score may have linguistic merits that justify their use as heuristic measures for collocation identification. While clearly not satisfactory as measures of significance, they must not completely be excluded from the following discussion, which focuses on empirical and intuitive properties of association measures.

To summarise the mathematical discussion, association measures can be divided into effect-size and significance measures (except for some heuristic equations that are difficult to place). These two major groups highlight different aspects of collocativity, and a decision for one or the other cannot be made on purely mathematical grounds. Within the significance group, mathematical theory and direct comparison of association scores clearly identify log-likelihood as the most appropriate and convenient measure, with simple-ll as a good approximation that does not require full contingency tables to be computed. Within the effect-size group, no clear-cut recommendation can be made, as measures tend to focus on different aspects of collocativity. In particular, MI seems appropriate for relatively weak associations that are compared to the independence baseline, while Dice identifies rigid word combinations with almost total association.

6.2 Collocations and multiword extraction

In those cases where mathematical theory does not help us choose between association measures, we can study their empirical properties independent of the underlying statistical reasoning. In this section, we specifically address empirical *linguistic* properties, i.e. we ask what kinds of word pairs are identified as collocations by the different association measures. A simple approach is to look at n-best lists as shown in Tables 4 and 5, which give a good impression of the different linguistic aspects of collocativity that the association measures capture. For instance, Table 4 indicates that simple-ll is a useful measure for identifying typical and intuitively plausible collocates of a node word. Without a frequency threshold, MI brings up highly specialised terms (**-record bucket*), but also many obviously accidental cooccurrences (such as *dippermouth* or *Dok*).[64] A more thorough and systematic study along these lines has been carried out by Stubbs (1995).

[63] Recall the low-frequency bias of MI that has already been observed in Section 4.2.

[64] These cooccurrences were found in the following sentences (node and collocate have been highlighted):

- *Accordingly, the five bodies are baptized with the names of famous Blues songs: '**Dippermouth**', 'Gut **Bucket**', 'Potato Head', 'Tin Roof', and 'Really'.* [G1N: 696]
- *By planting two seedlings of the variety '**Dok** Elgon' per 5-litre **bucket** of peat in August, they have harvested good-quality curds in time for Christmas.* [A0G: 2173]

It is also obvious that MI is highly sensitive to frequency thresholds: if only word pairs with cooccurrence frequency $f \geq 5$ are considered, the MI list looks much more reasonable, although it still includes a broad mixture of linguistic phenomena. A frequency threshold of $f \geq 3$ or $f \geq 20$ would produce entirely different lists once again. This instability, which is shared by all measures with a low-frequency bias, makes it necessary to carefully balance frequency thresholds against corpus size (and also span size for surface cooccurrence) in order to obtain an interpretable set of collocates. Researchers might thus prefer to use more robust measures without a low-frequency bias. Measures such as simple-ll and t-score, which have an explicit high-frequency bias, are not sensitive to frequency thresholds: there are no low-frequency cooccurrences among the highest-ranking collocates.

One has to keep in mind, though, that these are merely impressionistic case studies with serious shortcomings. (i) They take only a small number of highest-ranking collocations into account, and a quite different picture might emerge if one were to look at a list of several hundred collocations. (ii) Only a small number of association measures are typically considered. If more measures are included (especially very similiar measures like simple-ll and local-MI), it becomes increasingly difficult to make general statements about the distinct properties of individual measures. (iii) Impressions and conclusions from these case studies cannot easily be generalised to other data sets. In addition to the problem of frequency thresholds discussed above, results depend very much on the size and content of the corpus being used, preprocessing (such as automatic lemmatisation or part-of-speech filters), whether a node–collocate or a unit view is adopted, and especially on the type of cooccurrence (surface, textual or syntactic). When applied to adjacent bigrams in the Brown corpus, simple-ll ranks much less interesting patterns of functions words at the top of the list, while MI identifies proper names and noun compounds with high accuracy and is less susceptible to different frequency thresholds than for the node–collocate data (see Table 5).

More precise empirical statements than such impressionistic case studies can be made if there is a well-defined goal or application for the identified collocations. A particularly profitable setting is the use of association scores for multiword extraction, where the goal usually is to identify a particular subtype of multiword expressions, e.g. compounds (Schone and Jurafsky 2001), technical terminology (Daille 1994) or lexical collocations (Krenn 2000). Evert and Krenn (2001, 2005) suggest an evaluation methodology for such tasks that allows fine-grained quantitative comparisons between a large number of association measures. The evaluation follows the standard procedure for semi-automatic multiword extraction, where recurrent word pairs are obtained from a corpus, optionally filtered by frequency or other criteria, and ranked according to a selected association measure. Since there are no meaningful absolute thresholds for association scores (cf. Section 2.1), it is standard practice to select an n-best list of the 500, 1000 or 2000 highest-ranking collocations as candidate multiword expressions. The candidates are then validated by an expert, e.g. a professional lexicographer or terminologist.

In the evaluation setting, candidates in the n-best list are manually annotated as *true positives* (i.e. multiword expressions of the desired type) and *false positives*. These annotations are used to calculate the *precision* of the n-best list, i.e. the proportion of true positives among the n multiword candidates, and sometimes also *recall*, i.e. how many of all suitable multiword expressions that could have been extracted from the corpus are actually found in the n-best list. The precision values of different association measures can then be compared: the higher the precision of a measure, the better it is suited for identifying the relevant type of multiword expressions. Such evaluation experiments could be used, e.g., to confirm our impression that MI reliably identifies multiword proper names among adjacent bigrams (Table 5).

Instead of large and confusing tables listing precision values for various association measures and n-best lists, evaluation results can be presented in a more intuitive graphical

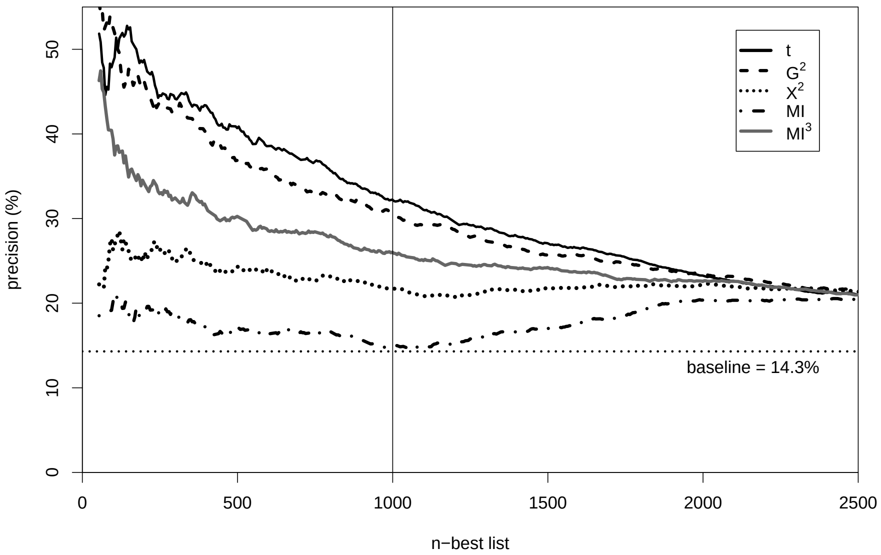


Figure 11: Comparative evaluation of association measures t-score (t), log-likelihood (G^2), chi-squared (X^2), MI and MI^3 on the data set of Krenn (2000).

form as precision plots. Figure 11 illustrates this evaluation methodology for the data set of Krenn (2000), who uses PP-verb cooccurrences from an 8 million word subset of the German *Frankfurter Rundschau* newspaper corpus to identify lexical collocations between prepositional phrases and verbs (including support verb constructions and figurative expressions).[65] The lines in Figure 11 summarise the precision values of five different association measures for arbitrary n-best lists. The precision for a particular n-best list can easily be read off from the graph, as indicated by the thin vertical line for $n = 1{,}000$: the solid line at the top shows that t-score achieves a precision of approx. 32% on the 1000-best list, while log-likelihood (the dashed line below) achieves only 30.5%. The precision of chi-squared (dotted line) is much lower at 21.5%. Looking at the full lines, we see that log-likelihood performs much better than chi-squared for all n-best lists, as predicted by the mathematical discussion in Section 6.1. Despite the frequency threshold, MI performs worse than all other measures and is close to the baseline precision (dotted horizontal line) corresponding to a random selection of candidates among all recurrent word pairs.[66] Evaluation results always have to be interpreted in comparison to the baseline, and an association measure can only be considered useful if it achieves substantially better precision. The most striking result is that t-score outperforms all other measures, despite its mathematical shortcomings. This illustrates the limitations of a purely theoretical discussion: empirically, t-score is the best indicator for lexical PP-verb collocations among all association measures.[67]

[65] In order to improve the performance of measures with a low-frequency bias, an additional frequency threshold $f \geq 5$ has been applied (the original study used a threshold of $f \geq 3$), leaving 4,489 candidate pairs to be ranked by the association measures.

[66] It is interesting to see that the heuristic variant MI^3 leads to a considerable improvement, suggesting that the poor performance of MI might indeed be connected to its low-frequency bias.

[67] Note that the precision values of all measures are virtually identical for $n \geq 2{,}000$. This is hardly surprising, as almost half of the candidates are included in the n-best list at this point.

Figure 12 compares the statistical association measures log-likelihood and chi-squared$_{\text{corr}}$ to their simple counterparts simple-ll and z-score. Their scores have already turned out to be very similar on the synthetic data set in Section 6.1, and this similarity is confirmed by the evaluation results: the performance of the statistical and the simple measure is virtually identical for each pair, although simple-ll is marginally worse than log-likelihood (far away from any significant difference). This shows that for the purpose of multiword extraction, the more convenient simple association measures can be used without reservation, an insight that has been confirmed (though not explicitly stated) by many other studies.

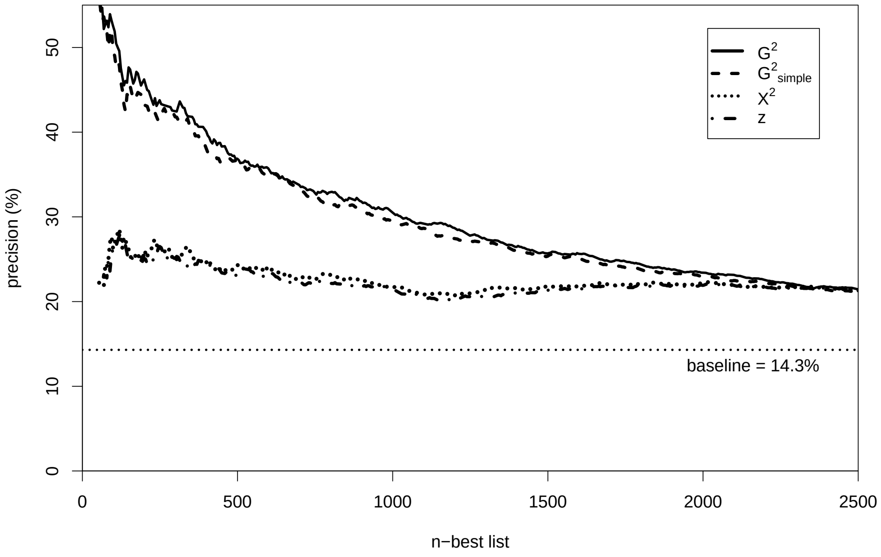


Figure 12: Comparison of simple and statistical association measures: log-likelihood (G^2) vs. simple-ll (G^2_{simple}) and chi-squared (X^2) vs. z-score (z). [*extended manuscript only*]

Similar evaluations have been carried out by a number of authors, but have not led to conclusive and generally valid answers.

Daille (1994) considers log-likelihood, MI^3 and the *Coefficient de Fager et McGowan* useful for the extraction of French terminological compound nouns. After a detailed qualitative evaluation, she singles out log-likelihood as the most appropriate association measure (Daille 1994, 173). This assessment is confirmed by Lezius (1999) in a small case study on German multiword expressions.

The results of Krenn (2000) remain on the whole inconclusive. MI and Dice seem to be the most suitable association measures for high-frequency data, while log-likelihood achieves better performance for medium- to low-frequency data. A lexical filter based on a list of typical “support verbs” (Breidt 1993) improves the identification of support-verb constructions. In many cases, the distributional *PP-entropy* measure yields better precision than association measures.

Schone and Jurafsky (2001) report that z-score, chi-squared and Dice (followed by MI) greatly outperform log-likelihood and t-score. These results may not be directly comparable to other studies, however, since the evaluation was carried out on n-grams of variable length. The results of Pearce (2002) are difficult to interpret because of the extremely low precision values obtained and because many standard measures were not included in the evaluation.

Pecina and Schlesinger (2006) find large groups of measures with nearly identical performance. Surprisingly, chi-squared and z-score are among the best measures. Precision can be further improved by combining multiple association measures with the help of machine-learning techniques. Evert and Krenn (2005) also report significantly better precision for chi-squared than for log-likelihood in a lexicographic application.

Some studies use different kinds of reference data than manually annotated true and false positives. For instance, Lapata *et al.* (1999) compare association scores to human plausibility judgements. The human ratings correlate better with cooccurrence frequency than with any association measure, but there is also a significant correlation with log-likelihood.

6.3 An intuitive geometrical model

In the previous section, we have looked at "linguistic" properties of association measures, viz. how accurately they can identify a particular type of multiword expressions or one of the other linguistic phenomena behind collocativity (see Section 2.2). If we take a pre-theoretic view of collocations as an observable property of language, though, the purpose of association scores is to measure this property in an appropriate way, not to match theoretical linguistic concepts. In this context, evaluation studies that depend on a theoretical or intuitive definition of true positives seem less appropriate. Instead, our goal should be to understand which quantitative aspects of collocativity each association measure singles out: we are interested in empirical mathematical properties of the measures.

The theoretical discussion in Section 6.1 already gives a partial answer: effect-size measures that do not take sampling variation into account may produce unreliable scores, especially for low-frequency data where they can easily overestimate association to a large degree. The same holds for other measures with a low-frequency bias such as z-score and chi-squared. These measures become much more useful if a sufficiently high frequency threshold is applied to the cooccurrence data. Measures with a high-frequency bias (e.g. simple-ll and log-likelihood) may put undue emphasis on very frequent word pairs, on the other hand. The t-score measure appears to be a special case: despite its unsatisfactory mathematical derivation, it achieves better performance than well-founded statistical measures in the multiword extraction task of Krenn (2000), cf. Figure 11. These findings suggest that some mathematical properties of the association measures cannot be captured by theoretical discussions alone.

Evert (2004, Sec. 3.4) proposes a geometric visualisation technique in order to reach the desired intuitive understanding of association measures. This technique works well for simple measures that require only two real numbers, O and E, to calculate an association score for a given word pair.[68] Interpreting the numbers (O, E) as two-dimensional coordinates, we can thus represent each word pair in a data set by a point in the real Euclidean plane. The left panel of Figure 13 illustrates this "point cloud" view for adjacent bigrams in the Brown corpus.[69] The data point representing the bigram *New York* (with $O = 303$ and $E \approx 0.54$) is marked with a circle. Its expected frequency $E \approx 0.5$ can be read off the x-axis, and its observed frequency $O = 303$ off the y-axis, as indicated by the thin horizontal and vertical lines. Note that both axes are on logarithmic scales in order to accommodate the wide range of observed and expected frequencies found in a typical data set. The frequency threshold $f \geq 10$ applied to the data set is clearly visible in the

[68] From the measure's point of view, this is all the relevant information about the word pair.

[69] The data set has been thinned by a factor of 10 to improve the visualisation. In addition, a technique called *jittering* has been applied, moving each point a small distance in a random direction, in order to avoid banding due to the integer quantisation of O and to avoid overplotting word pairs with identical frequency signatures.

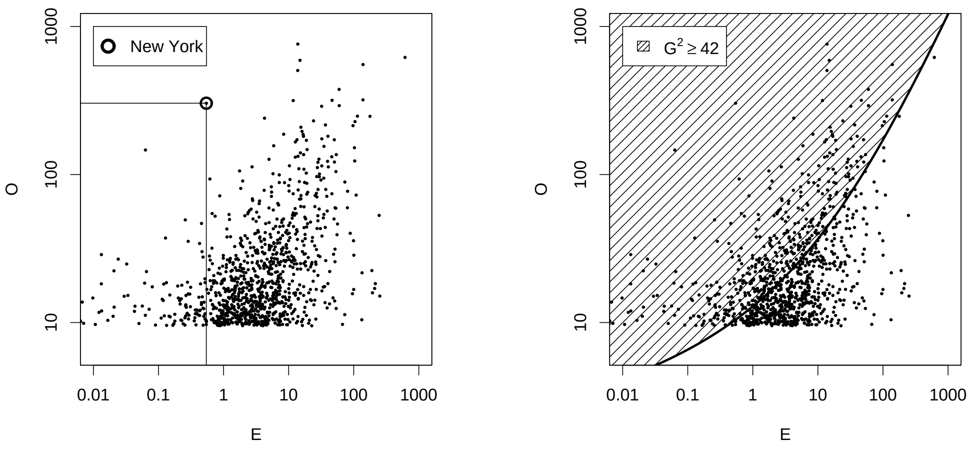


Figure 13: Geometric visualisation of cooccurrence frequency data (left panel) and an acceptance region of the simple-ll association measure (right panel).

graph.

Association scores are usually compared against a cutoff threshold, whose value is either established in advance (in a threshold approach) or determined interactively (for n-best lists). In terms of the geometric model, the point cloud representing a data set is divided into *accepted* and *rejected* points by such a cutoff threshold. For any given association measure and cutoff threshold, this decision only depends on the coordinates of a point in the Euclidean plane, not on the word pair represented by the point. It is therefore possible to determine for any hypothetical point in the plane whether it would be accepted or rejected, i.e. whether the association score would be higher than the threshold or not. The right panel of Figure 13 shows an illustration for the simple-ll measure and a cutoff threshold of 42. Any data point in the shaded region will be assigned a score $G^2 \geq 42$, and any point outside the region a score $G^2 < 42$.

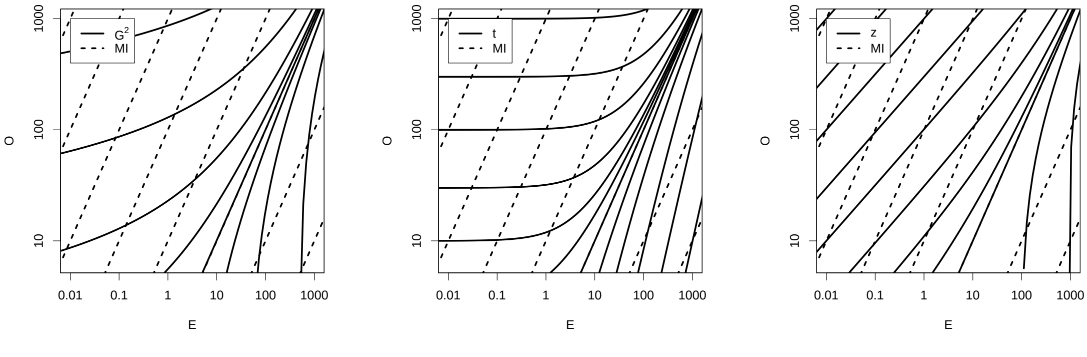


Figure 14: Intuitive comparison of simple association measures represented by contour plots. The three panels compare simple-ll (G^2, left panel), t-score (centre panel) and z-score (right panel) against MI (dashed lines).

It can be shown that for most association measures[70] the set of accepted hypothetical points forms a simple connected acceptance region. The region is bounded below by a single increasing line referred to as a contour of the association measure. All points on a contour line have the same association score according to this measure; in our example, a

[70] A notable exception is local-MI because it does not adhere to the conventions for associations scores.

simple-ll score of 42.[71] Every simple association measure is uniquely characterised by its contour lines for different threshold values. We can thus visualise and compare measures in the form of contour plots as shown in Figure 14. Each panel overlays the contour plots of two different association measures. Comparing the shapes of the contour lines, we can identify the characteristic mathematical properties of the measures and understand the differences between them. Reading contour plots takes some practice: keep in mind that contours connect points with the same association scores, just as the contour lines of a topographic map connect points of the same elevation.

MI only considers the ratio between O and E, even for very low observed frequency O. Hence its dashed contours in Figure 14 are straight lines.[72] These straight lines of constant ratio O/E also provide a grid for the interpretation of other contour plots.[73] A significance measure such as simple-ll (left panel) is sensitive to the smaller amount of evidence provided by low-frequency data. Therefore, a higher ratio between O and E is required to achieve the same score, and the contour lines have a left curvature. There is a single straight contour line, which marks the null hypothesis of independence ($O = E$) and coincides with the corresponding contour line of MI. Contours for positive association are located above and to the left of the independence line. Contours for negative association show a right curvature and are located below and to the right of the independence line.

It is important to keep in mind that the precise distances between contour lines are irrelevant, since there is no absolute scale for association scores (except for special cases such as significance measures that can be transformed to a common scale, cf. Section 6.1). For most applications, only the ranking of data points is important. While the contour lines of simple-ll seem to be further apart on average than those of MI, this does not tell us anything about the differences between the measures. An important observation, though, is that the contour lines of simple-ll are closer together for high-frequency data (in the top right corner of the plot) and spread out for low-frequency data (towards the bottom left corner), while the contour lines of MI are parallel across all frequency ranges.

The centre panel of Figure 14 shows a contour plot for the t-score measure. Again, independence is marked by a straight line that coincides with the MI contour. For positive association, the t-score contours have a left curvature similar to simple-ll, but much more pronounced. For very small expected frequencies, they flatten out to horizontal lines, creating an implicit frequency threshold effect.[74] We may speculate that this implicit threshold is responsible for the good performance of t-score in some evaluation studies, especially if low-frequency word pairs are not discarded in advance. Interestingly, the contour lines for negative association are nearly parallel and do not seem to take random variation into account, in contrast to simple-ll.

Finally, the right panel shows contour lines for z-score. Despite its mathematical background as a significance measure, z-score fails to discount low-frequency data. The contour lines for positive association are nearly parallel, although their slope is less steep than for MI. Thus, even data points with low observed frequency O can easily achieve high association scores, explaining the low-frequency bias of z-score that has been noted repeatedly.

[71] Note that contour lines always have non-negative slope, but not necessarily leftward curvature, as shown by the examples in Figure 14.

[72] They are parallel lines with the same slope (rather than lines with different slopes intersecting at the origin) because of the logarithmic scale of the plots.

[73] In Figure 14, the MI contour lines are chosen to correspond to ratios O/E that are powers of ten, i.e. $O = E$ for independence, $O = 10 \cdot E$, $O = 100 \cdot E$, etc. for positive association, and $O = E/10$, $O = E/100$, etc. for negative association.

[74] Because of this implicit frequency threshold, high association scores can only be achieved by word pairs with relatively high O, no matter how small E becomes.

Interestingly, z-score seems to work well as a measure of significance for negative association, where its contour lines are very similar to those of simple-ll.

The visualisation technique presented in this section can be extended to statistical association measures, but the geometric interpretation is more difficult and requires three-dimensional plots. See Evert (2004, Sec. 3.4) for details and sample plots.

7 Summary and conclusion

In this article, we have been concerned with the empirical Firthian notion of *collocations* as observations on the combinatorics of words in a language, which have to be distinguished clearly from lexicalised *multiword expressions* as pre-fabricated units, and in particular from *lexical collocations*, a subtype of multiword expressions. From the perspective of theoretical linguistics, collocations are often understood as an *epiphenomenon*, the surface reflections of compounds, idioms, lexical collocations and other types of multiword expressions, selectional preferences, semantic restrictions, cultural stereotypes, and to a considerable extent also conceptual knowledge ("facts of life").

Introduced as an intuitively appealing, but fuzzy and pre-theoretical notion by Firth (1957), collocativity can be operationalised in terms of *cooccurrence* frequencies and quantified by mathematical *association measures*. High association scores indicate strong attraction between two words, but there is no standard scale of measurement to draw a clear distinction between collocations and non-collocations. Association measures and collocations have many uses, ranging from technical applications in computational linguistics to lexicographic and linguistic studies, where they provide descriptive generalisations about word usage. Collocations are closely related to lexicalised multiword expressions, and association measures are central to the task of automatic multiword extraction from corpora.

In order to identify and score collocations from a given corpus, the following steps have to be performed: (1) Choose an appropriate *type of cooccurrence* (surface, textual or syntactic). (2) Determine *frequency signatures* (i.e. cooccurrence frequency f and the marginal frequencies f_1 and f_2 in the corpus) for all relevant word pairs (w_1, w_2) as described in Section 3 (Figures 1, 2 and 3 serve as a reminder), as well as sample size N. (3) Filter the cooccurrence data set by applying a *frequency threshold*. Theoretical considerations suggest a minimal threshold of $f \geq 3$ or $f \geq 5$, but higher thresholds often lead to even better results in practice. (4) Calculate the *expected frequencies* of the word pairs, using the general equation $E = f_1 f_2 / N$ for textual and syntactic cooccurrence, and the approximation $E = k f_1 f_2 / N$ for surface cooccurrence, where k is the total span size. (5) Apply one of the *simple association measures* shown in Figure 4, or produce multiple tables according to different measures. Recall that the cooccurrence frequency f is denoted by O (for *observed frequency*) in these equations. (5) If collocations are treated as units, *rank* the word pairs by association score, or select a threshold to distinguish between collocations and non-collocations (or "strong" and "weak" collocations). In the node–collocate view, collocates w_2 are ranked separately for each node word w_1.

If the data include word pairs with highly skewed marginal frequencies and you suspect that this may have distorted the results of the collocation analysis, you may want to apply *statistical association measures* instead of the simple measures. In order to do so, you have to compute a full 2×2 contingency table for each word pair, as well as a corresponding table of expected frequencies (see Figure 5). The precise calculation procedure depends on the type of cooccurrence and is detailed in Section 5.1 (Figures 6, 7 and 8 serve as quick reminders). Then, one or more of the statistical measures in Figure 9 can be applied. Many further measures are found in (Evert 2004) as well as online at `http://www.collocations.de/AM/` (both resources use the same notation as in this article).

The resulting set or ranking of collocations depends on many parameters, including the size and composition of the corpus, pre-processing (such as lemmatisation), application of frequency thresholds, the definition of cooccurrence used, and the choice of association measure. It is up to the researcher to find a suitable and meaningful combination of parameters, or to draw on results from multiple parameter settings in order to highlight different aspects of collocativity. While a particular type of cooccurrence is often dictated by the theoretical background of a study or practical restrictions (e.g., syntactic cooccurrence requires a sufficiently accurate software for automatic syntactic analysis, or a pre-parsed corpus), other parameter values are more difficult to choose (e.g. span size for surface cooccurrence, or the frequency threshold).

A crucial step, of course, is to select one of well over 50 different association measures that are currently available (or to invent yet another measure). At this point, no definitive recommendation can be made. It is perhaps better to apply several measures with well-understood and distinct properties than attempt to find a single optimal choice. In any case, a thorough understanding of the characteristic properties of association measures and the differences between them is essential for a meaningful interpretation of the extracted collocations and their rankings. In Section 6, various theoretical and empirical techniques have been introduced for this purpose, and the properties of several widely used measures have been discussed.

7.1 Open questions and extensions

The goal of this article was to present the current state of the art with regard to collocations and association measures. The focus has therefore been on established results rather than unsolved problems, open research questions, or extensions beyond simple word pairs. The following paragraphs give an overview of important topics of current research.

Like all statistical approaches in corpus linguistics, association measures suffer from the fact that the assumptions of their underlying statistical models are usually not met by corpus data. In addition to the general question whether any finite corpus can be representative of a language (which is a precondition for the validity of statistical generalisations), *non-randomness* of corpus frequency data is a particularly serious problem for all statistical models based on random samples. A thorough discussion of this problem and possible solutions can be found in Article 36 and in (Evert 2006).

In addition to these common issues, cooccurrence data pose two specific difficulties. First, the null hypothesis of independence is extremely unrealistic. Words are never combined at random in natural language, being subject to a variety of syntactic, semantic and lexical restrictions. For a large corpus, even a small deviation from the null hypothesis may lead to highly significant rejection and inflated association scores calculated by significance measures. Effect-size measures are also subject to this problem and will produce inflated scores, e.g. for two rare words that always occur near each other (such as *déjà* and *vu*). A possible solution would be to specify a more realistic null hypothesis that takes some of the restrictions on word combinatorics into account, but research along these lines is still at a very early stage.

Second, word frequency distributions are highly skewed, with few very frequent types and a large number of extremely rare types. This property of natural language, often referred to as *Zipf's law* (see Articles 37 and 41), is even more pronounced for cooccurrence data. In combination with the quantisation of observed frequencies (it is impossible to observe $O = 0.7$ cooccurrences), Zipf's law invalidates statistical corrections for sampling variation to the extent that accidental cooccurrences between low-frequency words may achieve very high association scores. An extensive study of this effect has resulted in the recommendation to apply a frequency threshold of $f \geq 5$ in order to weed out potentially spurious collocations (Evert 2004, Ch. 4). Non-randomness effects may exacerbate the situation and necessitate even higher thresholds. Current research based on more sophisticated models of Zipfian frequency distributions aims to develop better correction techniques that are less drastic than a simple frequency threshold.

Intuitively, “mutual expectancies” often hold between more than two words. This is particularly obvious in the case of multiword expressions: *kick ... bucket* is always accompanied by the definite article *the*, *humble pie* usually occurs with *eat*, and the bigram *New York* is often followed by *City*.[75] Applying association measures to word pairs will only bring up fragments of such larger collocations, and the missing pieces have to be filled in from the intuition of a linguist or lexicographer. It is therefore desirable to develop suitable measures for word triples and larger *n*-tuples. First attempts to formulate such measures are straightforward generalisations of the equations of MI (Lin 1998), log-likelihood (Zinsmeister and Heid 2003), or the Dice coefficient (da Silva and Lopes 1999). Obviously, a deep understanding of the mathematical properties of association measures for word pairs as well as their shortcomings is essential for a successful extension.

A key problem lies in the fact that the null hypothesis of independence becomes even less realistic for a combination of three or more words, leading to extremely small expected frequencies.[76] Simple association measures become virtually meaningless under these circumstances. The full cooccurrence frequency data for a combination of n words can be summarised in an n-dimensional contingency table with 2^n cells. For larger values of n, these tables suffer from an acute data sparseness problem, with many empty or low-frequency cells. Appropriate measures of significance and effect size for the association in n-dimensional contingency tables are poorly understood, and more sophisticated statistical models may be required (Agresti 2002). In addition, it makes little sense to consider word pairs, triples, quadruples, etc. separately from each other. Automatic methods are needed to determine how many words form part of a collocation and to distinguish e.g. between genuine three-word collocations (*New York City*), nested collocations ({*eat* {*humble pie*}}), overlapping two-word collocations (*sip black coffee*, where *sip coffee* and *black coffee* are “independent” collocations), and accidental cooccurrences of a two-word collocation with a third word (cf. Zinsmeister and Heid 2003).

With the extension to n-word collocations, regular patterns become more noticeable: in addition to the well-known collocation *carry emotional baggage*, we also find *carry cultural, historical, ideological, intellectual, political, ... baggage* (some of them even more frequent than *emotional baggage*).[77] This evidence suggests a productive collocational pattern of the form *carry* Adj *baggage*, with additional semantic restrictions on the adjective. Many instances of such patterns are too rare to be identified in corpora by statistical means, but would intuitively be considered as collocations by human speakers (think of *carry phraseo-*

[75] Similar patterns can also be observed for collocations that do not reflect lexicalisation phenomena. For instance, the collocation of *bucket* with *throw* is very frequently accompanied by the noun *water*. Out of 36 instances of *bucket* that cooccur with *throw*, more than half, viz. 20 instances also cooccur with *water*; while overall, only 13.5% of the instances of *bucket* cooccur with *water* (183 of 1,356).

[76] For example, the expected frequency of the bigram *New York* in the Brown corpus is $E = 0.5$; the expected frequency of the trigram *New York City* is $E = 0.00025$. Recall that words are case-folded here: if we only considered uppercase variants, the expected frequencies would be even smaller ($E = 0.2$ and $E = 0.000029$).

[77] These examples were extracted from a 2.1 billion word Web corpus of British English, compiled by the WaCky initiative in 2007. Frequency counts for the different adjectives in the construction *carry* Adj *baggage* are: 15×*cultural*, 13×*emotional*, 6×*historical*, 5×*ideological*, 4×*political* and 3×*intellectual*. In the BNC, *ideological baggage* (9×) is overall much more frequent than *emotional baggage* (3×).

logical baggage, for instance). There has been little systematic research on the productivity of collocations so far, notable exceptions being Lüdeling and Bosch (2003) and Stevenson *et al.* (2004).

Many collocations are intuitively felt to be *asymmetric*. For instance, in the bigram *the Iliad*, *the* is a more important collocate for *Iliad* than *Iliad* is for *the*. In the terminology of Kjellmer (1991), the bigram is left-predictive, but not right-predictive.[78] Although such asymmetries are often reflected in skewed marginal frequencies (the collocation being more important for the less-frequent word), hardly any of the known association measures make use of this information.[79] Preliminary research suggests that measures of *directed association* could be based on the ratios O/f_1 and O/f_2 (as estimators for the conditional probability that w_1 is accompanied by w_2 and vice versa), or could be formulated by putting the association score of a word pair (w_1, w_2) in relation to the scores of all collocates of w_1 and w_2, respectively (Michelbacher *et al.* 2007).

Collocation studies are usually interested in positive associations between words. Sometimes, however, words seem to repel each other and cooccur less often than would be expected by chance. In fact, many possible word pairs are never found to cooccur even in billion-word corpora. Such repulsion leads to a negative association between the words ($O \ll E$), which may then be termed *anti-collocations* (Pearce 2001). Most anti-collocations are probably a consequence of semantic, pragmatic or stylistic incompatibilities, while others result from competition e.g. with established lexical collocations (you can *brush* or *clean* you *teeth*, but you will rarely *wash* or *scrub* them; you *give a talk* but do not *deliver* it, while you can *deliver* a *speech, sermon* or *verdict*).

The automatic identification and quantification of anti-collocations faces two difficult problems. First, the unrealistic null hypothesis of independence results in an artificially low expected frequency E, so that it is hard to find word pairs that cooccur significantly less often than expected by chance. Word pairs that appear to be statistically independent might in fact be anti-collocations with respect to an appropriate null hypothesis. Second, evidence for negative association is much less reliable than evidence for positive association. Even if a word pair does not cooccur at all in a large corpus (the most clear-cut indication of negative association), one cannot be certain that it would not do so in a different corpus. Statistical significance is only reached if the marginal frequencies are high enough so that $E \gg 1$. Recall that according to the contour plots in Figure 14, simple-ll and z-score are more appropriate measures of negative association than t-score.

Although many association measures are available, there is still room for improvement and it would be desirable to develop measures with novel properties. Most existing measures fall into one of two major groups, viz. effect-size and significance measures. Both groups have their strengths and weaknesses: effect-size measures do not correct for sampling variation, while significance measures are biased towards high-frequency word pairs with small effect sizes (which tend to be uninteresting from a linguistic point of view). New association measures might be able to combine aspects of effect-size and significance measures, striking a balance between the low-frequency bias of the former and the high-frequency bias of the latter. First steps in this direction are summarised by Evert (2004, Sec. 3.1.8), but have not led to satisfactory results yet.

[78] John Sinclair has used the terms *upward* and *downward* collocation (Sinclair *et al.* 2004, xxiii).

[79] Simple association measures, which use only the expected frequency E but not the marginals f_1, f_2, treat all collocations as symmetric units. Statistical association measures have access to information from the full contingency table and would in principle be able to calculate directed association. However, all measures presented in this article are symmetric, i.e. they calculate the same scores for (w_2, w_1) as for (w_1, w_2).

7.2 Further reading

Evert (2004) gives a more detailed account of statistical models for association in contingency tables and their limitations, together with a comprehensive inventory of association measures and methods for the comparison and evaluation of different measures. An online version of the inventory can be found at `http://www.collocations.de/AM/`. Contingency tables and the statistical tests that form the basis of many association measures are explained in standard textbooks on mathematical statistics (e.g. DeGroot and Schervish 2002). Advanced books (e.g. Agresti 2002) introduce more sophisticated models for the analysis of contingency tables. Although these models have not found widespread use as association measures yet, they may become important for the development of novel measures and their extension beyond simple word pairs.

Bartsch (2004) offers an insightful theoretical discussion of collocations and their properties, as well as an excellent overview of the various empirical and phraseological definitions of the term. Exemplary proponents of the two views are Sinclair (1991) and Sinclair *et al.* (2004) on the empirical side, and standard textbooks (e.g. Burger *et al.* 1982) for the phraseological view. Current research on collocations and multiword expressions is collected in the proceedings of ACL Workshops on Multiword Expressions (2001, 2003, 2004, 2006, 2007) and in Grossmann and Tutin (2003).

Relevant articles in this volume are Article 24 (on word segmentation and part-of-speech tagging), Article 25 (on lemmatisation), Article 26 (on word sense disambiguation) and Article 28 (on automatic syntactic annotation), as well as Article 10 (on text corpora). Article 36 is a general introduction to the statistical analysis of corpus frequency data, including most of the techniques on which association measures are based. Important applications of collocations can be found in the articles on computational lexicography (Article 8) and word meaning (Article 45).

We have followed a traditional view of collocations as simple word pairs here, but association measures and related techniques can equally well be applied to cooccurrences of other linguistic units (e.g. lexical items and syntactic constructions in Article 43).

A Appendix

This appendix does not appear in the published article.

A.1 Derivation of the simple-ll measure

In analogy to the two-sample likelihood-ratio test suggested by Dunning (1993), the simple log-likelihood (*simple-ll*) measure is the test statistic of a one-sample likelihood-ratio test for the null hypothesis $H_0 : \mathrm{E}[X] = E$, where X is a random variable describing the cooccurrence frequency of a word pair, O is the value of X observed for the particular sample under consideration, and E is the expected frequency assuming independence. This test statistic is given by

$$-2 \log \lambda = -2 \log \frac{\Pr_0(X = O)}{\Pr(X = O)} \tag{1}$$

where λ is the ratio between the maximum likelihood of the observation $X = O$ under H_0, written $\Pr_0(X = O)$, and its unconstrained maximum likelihood, written $\Pr(X = O)$. If H_0 holds, $-2 \log \lambda$ has an asymptotic chi-squared distribution with df $= 1$. In particular, simple-ll is a two-sided association measure that does not distinguish between positive and negative association. It can be converted into a one-sided measure by the standard procedure of multiplying the association scores with -1 if $O < E$; the signed square root of this one-sided statistic has an asymptotic normal distribution under H_0.

Eq. (1) can be expanded in two different ways, depending on whether X is assumed to have a *binomial* distribution ($X \sim B(N, p)$, with $H_0 : p = E/N$) or a *Poisson* distribution ($X \sim P(\alpha)$, with $H_0 : \alpha = E$). The difference is that the binomial distribution is conditioned on a given fixed sample size N, while the Poisson distribution is not. See Evert (2004) for an extensive discussion of binomial vs. Poisson sampling. Only the Poisson distribution leads to a simple association measure, since the equations for the binomial distribution involve the sample size N in addition to O and E.

For the Poisson distribution, we have

$$\Pr(X = O) = e^{-O} \frac{O^O}{O!} \tag{2}$$

with the unconstrained MLE $\alpha = O$ and

$$\Pr_0(X = O) = e^{-E} \frac{E^O}{O!} \tag{3}$$

with $H_0 : \alpha = E$. Inserting (2) and (3) in Eq. (1), we obtain

$$\lambda = e^{O-E} \left(\frac{E}{O} \right)^O \tag{4}$$

and hence

$$-2 \log \lambda = 2 \left((E - O) + O \cdot \log \frac{O}{E} \right) = 2 \left(O \cdot \log \frac{O}{E} - (O - E) \right). \tag{5}$$

Using the binomial distribution instead of Poisson, we have

$$\Pr(X = O) = \binom{N}{O} \left(\frac{O}{N} \right)^O \left(1 - \frac{O}{N} \right)^{N-O} \tag{6}$$

with the unconstrained MLE $p = O/N$ and

$$\Pr_0(X = O) = \binom{N}{O} \left(\frac{E}{N}\right)^{O} \left(1 - \frac{E}{N}\right)^{N-O} \tag{7}$$

with $H_0 : p = E/N$. Therefore,

$$\lambda = \frac{E^O \cdot (N-E)^{N-O}}{O^O \cdot (N-O)^{N-O}} = \left(\frac{E}{O}\right)^{O} \cdot \left(\frac{N-E}{N-O}\right)^{N-O} \tag{8}$$

and

$$-2\log\lambda = 2\left(O \cdot \log\frac{O}{E} + (N-O) \cdot \log\frac{N-O}{N-E}\right). \tag{9}$$

Using the series expansion

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{x^5}{5} - \frac{x^6}{6} + \dots \tag{10}$$

and

$$\frac{N-E}{N-O} = \frac{(N-O) + O - E}{N-O} = 1 + \frac{O-E}{N-O} \tag{11}$$

we find that

$$\begin{aligned} (N-O) \cdot \log\frac{N-O}{N-E} &= -(N-O) \cdot \log\frac{N-E}{N-O} \\ &= -(N-O) \cdot \left(\frac{O-E}{N-O} - \frac{(O-E)^2}{2(N-O)^2} + \dots\right) \\ &= -(O-E) + \frac{(O-E)^2}{2(N-O)} - \dots . \end{aligned} \tag{12}$$

Inserting (12) into (9), we see that the Poisson form (5) and the binomial form (9) of simple-ll are asymptotically equivalent for $O \ll N$. The binomial form (9) shows more clearly that log-likelihood is a straightforward extension of simple-ll.

A.2 BNC examples of *kick the bucket*

This appendix lists all instances of the collocation (*kick*, *bucket*) in the British National Corpus. A simple corpus search by surface distance finds 20 cooccurrences, all of which are indeed verb-object constructions. The lower frequency of 12 reported in Table 2 is explained by two instances of *kick* mistagged as a noun, and the fact that the category filter used for data preparation rejected 6 instances of *kick* with ambiguity tags.

Of the 20 corpus examples found, 8 refer to the literal meaning (marked **L** below). Only 3 are direct uses of the idiom *kick the bucket* (marked **I**), all in reported speech, while the remaining 9 talk *about* the expression (marked **M** for meta-discussion). Among the latter, there are also two cases where *kick the bucket* is repeated within the same sentence.

[AC4]	**L**	*Jinny was so startled that she nearly kicked the bucket over.*
[ATE]	**I**	*“Did you think I’d kicked the bucket, Ma?”*
[C8P]	**L**	*You can also invent little games, such as kicking a ball in a bucket or bowl of water.*
[CA0]	**L**	*Umberto, snoring in the tack room, barricaded against ghoul and hobgoblin by one of the feedbins, was woken to a punishing hangover by the increasingly irritated din of muzzled horses kicking their water buckets.*

[CK9] **L** *When Laddie kicked over a bucket of mash, she said, "You bloody silly donkey."*

[FAC] **M** *It has long been recognised that expressions such as to pull someone's leg, to have a bee in one's bonnet, to kick the bucket, to cook someone's goose, to be off one's rocker, round the bend, up the creek, etc. are semantically peculiar.*

[FAC] **M** *For instance, the reason that to pull someone's left leg and to kick the large bucket have no normal idiomatic interpretation is that leg and bucket carry no meaning in the idiom, so there is nothing for left and large to carry out their normal modifying functions on (in general, a modifier needs a semantic constituent to modify).*

[FAC] **M** *Thus, The aspidistra kicked the bucket exemplifies inappropriateness because replacing kick the bucket with its cognitive synonym die removes the dissonance.*

[G0P] **L** *It was as if God had kicked a bucket of water over.*

[GW8] **L** *Suddenly Clare jumped up, leaving his bucket to be kicked over by the cow, went quickly towards her, and, kneeling down beside her, took her in his arms.*

[HE0] **M** *When an idiom is just something that has the form of, has a certain apparent grammatical form but actually occurs just as a single unit of a fixed meaning, so it has no genuine semantic structure from which you can determine its meaning, for example kick the bucket means die and you don't get that in the meaning of kick the bucket.*

[HE0] **M** *but notice kick the bucket appears as a verb phrase and eat humble pie, get your knickers in a twist and so on.*

[HE0] **M** *So like I said kick the bucket, the meaning of that idiomatically is just die, sorry die.*

[HE0] **M** *So, although in all these three, kick the bucket, eat humble pie, get your knickers in a twist er all look like fairly complex transitive constructions.*

[HH1] **L** *Two of the men flung themselves down on a bench, scabbards clattering, while the third strode forward, kicking Isabel's abandoned bucket out of his path.*

[HTG] **I** *"Chatterton and Fagg and a few more like them who've since kicked the bucket. . . . "*

[JXU] **I** *"It's just that Uncle was a cautious old devil and—" he looked away "—he got the impression I was a bit of a spendthrift because—well, because I used to get through my allowance pretty rapidly when I was away at school, and . . . oh, hell, he wanted to make sure I was going to be dull and sensible about all that money when he finally kicked the bucket. . . . "*

[KC8] **L** *He just kicked the bucket.*

References

Agresti, Alan (2002). *Categorical Data Analysis*. John Wiley & Sons, Hoboken, 2nd edition.

Aston, Guy and Burnard, Lou (1998). *The BNC Handbook*. Edinburgh University Press, Edinburgh. See also the BNC homepage at `http://www.natcorp.ox.ac.uk/`.

Bartsch, Sabine (2004). *Structural and Functional Properties of Collocations in English*. Narr, Tübingen.

Berry-Rogghe, Godelieve L. M. (1973). The computation of collocations and their relevance to lexical studies. In A. J. Aitken, R. W. Bailey, and N. Hamilton-Smith (eds.), *The Computer and Literary Studies*, pages 103–112. Edinburgh.

Blaheta, Don and Johnson, Mark (2001). Unsupervised learning of multi-word verbs. In *Proceedings of the ACL Workshop on Collocations*, pages 54–60, Toulouse, France.

Blumenthal, Peter; Diwersy, Sascha; Mielebacher, Jörg (2005). Kombinatorische Wortprofile und Profilkontraste. Berechnungsverfahren und Anwendungen. *Zeitschrift für romanische Philologie*, **121**(1), 49–83.

Breidt, Elisabeth (1993). Extraction of N-V-collocations from text corpora: A feasibility study for German. In *Proceedings of the 1st ACL Workshop on Very Large Corpora*, Columbus, Ohio. (a revised version is available from `http://arxiv.org/abs/cmp-lg/9603006`).

Burger, Harald; Buhofer, Annelies; Sialm, Ambros (1982). *Handbuch der Phraseologie*. de Gruyter, Berlin, New York.

Choueka, Yaacov (1988). Looking for needles in a haystack. In *Proceedings of RIAO '88*, pages 609–623.

Church, Kenneth; Gale, William A.; Hanks, Patrick; Hindle, Donald (1991). Using statistics in lexical analysis. In *Lexical Acquisition: Using On-line Resources to Build a Lexicon*, pages 115–164. Lawrence Erlbaum.

Church, Kenneth W. and Hanks, Patrick (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, **16**(1), 22–29.

da Silva, Joaquim Ferreira and Lopes, Gabriel Pereira (1999). A local maxima method and a fair dispersion normalization for extracting multi-word units from corpora. In *6th Meeting on the Mathematics of Language*, pages 369–381, Orlando, FL.

Daille, Béatrice (1994). *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. Ph.D. thesis, Université Paris 7.

DeGroot, Morris H. and Schervish, Mark J. (2002). *Probability and Statistics*. Addison Wesley, Boston, 3rd edition.

Dennis, Sally F. (1965). The construction of a thesaurus automatically from a sample of text. In M. E. Stevens, V. E. Giuliano, and L. B. Heilprin (eds.), *Proceedings of the Symposium on Statistical Association Methods For Mechanized Documentation*, volume 269 of *National Bureau of Standards Miscellaneous Publication*, pages 61–148, Washington, DC.

Dias, Gaël; Guilloré, Sylvie; Lopes, José G. P. (1999). Language independent automatic acquisition of rigid multiword units from unrestricted text corpora. In *Proceedings of Traitement Automatique des Langues Naturelles (TALN)*, Cargèse, France.

Dunning, Ted E. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**(1), 61–74.

Dunning, Ted E. (1998). *Finding Structure in Text, Genome and Other Symbolic Sequences*. Ph.D. thesis, Department of Computer Science, University of Sheffield.

Evert, Stefan (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart. Published in 2005, URN urn:nbn:de:bsz:93-opus-23714. Available from `http://www.collocations.de/phd.html`.

Evert, Stefan (2006). How random is a corpus? The library metaphor. *Zeitschrift für Anglistik und Amerikanistik*, **54**(2), 177–190.

Evert, Stefan and Kermes, Hannah (2003). Experiments on candidate data for collocation extraction. In *Companion Volume to the Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 83–86.

Evert, Stefan and Krenn, Brigitte (2001). Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 188–195, Toulouse, France.

Evert, Stefan and Krenn, Brigitte (2005). Using small random samples for the manual evaluation of statistical association measures. *Computer Speech and Language*, **19**(4), 450–466.

Firth, J. R. (1957). A synopsis of linguistic theory 1930–55. In *Studies in linguistic analysis*, pages 1–32. The Philological Society, Oxford. Reprinted in Palmer (1968), pages 168–205.

Francis, W. N. and Kucera, H. (1964). Manual of information to accompany a standard sample of present-day edited American english, for use with digital computers. Technical report, Department of Linguistics, Brown University, Providence, RI. Revised ed. 1971; revised and augmented 1979.

Gil, Alexandre and Dias, Gaël (2003). Using masks, suffix array-based data structures and multidimensional arrays to compute positional ngram statistics from corpora. In *Proceedings of the ACL Workshop on Multiword Expressions*, Sapporo, Japan.

Goldman, Jean-Philippe; Nerima, Luka; Wehrli, Eric (2001). Collocation extraction using a syntactic parser. In *Proceedings of the ACL Workshop on Collocations*, pages 61–66, Toulouse, France.

Grossmann, Francis and Tutin, Agnès (eds.) (2003). *Les Collocations: analyse et traitement*. De Werelt, Amsterdam.

Hausmann, Franz Josef (1989). Le dictionnaire de collocations. In *Wörterbücher, Dictionaries, Dictionnaires. Ein internationales Handbuch*, pages 1010–1019. de Gruyter, Berlin.

Hausmann, Franz Josef (2004). Was sind eigentlich Kollokationen? In K. Steyer (ed.), *Wortverbindungen – mehr oder weniger fest*, Jahrbuch des Instituts für Deutsche Sprache 2003, pages 309–334. de Gruyter, Berlin.

Heid, Ulrich; Evert, Stefan; Docherty, Vincent; Worsch, Wolfgang; Wermke, Matthias (2000). A data collection for semi-automatic corpus-based updating of dictionaries. In U. Heid, S. Evert, E. Lehmann, and C. Rohrer (eds.), *Proceedings of the 9th EURALEX International Congress*, pages 183 – 195.

Kilgarriff, Adam; Rychly, Pavel; Smrz, Pavel; Tugwell, David (2004). The sketch engine. In *Proceedings of the 11th EURALEX International Congress*, Lorient, France.

Kjellmer, Göran (1991). A mint of phrases. In K. Aijmer and B. Altenberg (eds.), *English Corpus Linguistics*, pages 111–127. Longman, London.

Krenn, Brigitte (2000). *The Usual Suspects: Data-Oriented Models for the Identification and Representation of Lexical Collocations*, volume 7 of *Saarbrücken Dissertations in Computational Linguistics and Language Technology*. DFKI & Universität des Saarlandes, Saarbrücken, Germany.

Lapata, Maria; McDonald, Scott; Keller, Frank (1999). Determinants of adjective-noun plausibility. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL 1999)*, pages 30–36, Bergen, Norway.

Lea, Diana (ed.) (2002). *Oxford Collocations Dictionary for students of English*. Oxford University Press, Oxford, New York.

Lezius, Wolfgang (1999). Automatische Extrahierung idiomatischer Bigramme aus Textkorpora. In *Tagungsband des 34. Linguistischen Kolloquiums*, Germersheim, Germany.

Liebetrau, Albert M. (1983). *Measures of Association*. Number 32 in Sage University Papers Series on Quantitative Applications in the Social Sciences. Sage, Newbury Park.

Lin, Dekang (1998). Extracting collocations from text corpora. In *Proceedings of the First Workshop on Computational Terminology*, pages 57–63, Montreal, Canada.

Lüdeling, Anke and Bosch, Peter (2003). Identification of productive collocations. In *Proceedings of the 8th International Symposium on Social Communication*, Santiago de Cuba.

Manning, Christopher D. and Schütze, Hinrich (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Michelbacher, Lukas; Evert, Stefan; Schütze, Hinrich (2007). Asymmetric association measures. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2007)*, Borovets, Bulgaria.

Palmer, F. R. (ed.) (1968). *Selected Papers of J. R. Firth 1952–59*. Longmans, London.

Pearce, Darren (2001). Synonymy in collocation extraction. In *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources*, Pittsburgh, PA.

Pearce, Darren (2002). A comparative evaluation of collocation extraction techniques. In *Third International Conference on Language Resources and Evaluation (LREC)*, pages 1530–1536, Las Palmas, Spain.

Pecina, Pavel (2005). An extensive empirical study of collocation extraction methods. In *Proceedings of the ACL Student Research Workshop*, pages 13–18, Ann Arbor, MI.

Pecina, Pavel and Schlesinger, Pavel (2006). Combining association measures for collocation extraction. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006), Poster Sessions*, pages 651–658, Sydney, Australia. ACL.

Pedersen, Ted (1996). Fishing for exactness. In *Proceedings of the South-Central SAS Users Group Conference*, Austin, TX.

Sag, Ivan A.; Baldwin, Timothy; Bond, Francis; Copestake, Ann; Flickinger, Dan (2002). Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2002)*, pages 1–15, Mexico City, Mexico.

Sahlgren, Magnus (2006). *The Word Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, Department of Linguistics, Stockholm University.

Schiehlen, Michael (2004). Annotation strategies for probabilistic parsing in German. In *Proceedings of COLING 2004*, pages 390–396, Geneva, Switzerland.

Schone, Patrick and Jurafsky, Daniel (2001). Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 100–108, Pittsburgh, PA.

Schütze, Hinrich (1998). Automatic word sense discrimination. *Computational Linguistics*, **24**(1), 97–123.

Sinclair, John (1991). *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.

Sinclair, John (ed.) (1995). *Collins COBUILD English Dictionary*. Harper Collins, London. New edition, completely revised.

Sinclair, John; Jones, Susan; Daley, Robert; Krishnamurthy, Ramesh (2004). *English Collocation Studies: The OSTI Report*. Continuum Books, London and New York. Originally written in 1970 (unpublished).

Sinclair, John McH. (1966). Beginning the study of lexis. In C. E. Bazell, J. C. Catford, M. A. K. Halliday, and R. H. Robins (eds.), *In Memory of J. R. Firth*, pages 410–430. Longmans, London.

Smadja, Frank (1993). Retrieving collocations from text: Xtract. *Computational Linguistics*, **19**(1), 143–177.

Smadja, Frank; McKeown, Kathleen R.; Hatzivassiloglou, Vasileios (1996). Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, **22**(1), 1–38.

Stevenson, Suzanne; Fazly, Afsaneh; North, Ryan (2004). Statistical measures of the semi-productivity of light verb constructions. In *Proceedings of the ACL 2004 Workshop on Multiword Expressions: Integrating Processing*, pages 1–8, Barcelona, Spain.

Stubbs, Michael (1995). Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language*, **1**, 23–55.

Terra, Egidio and Clarke, Charles L. A. (2004). Fast computation of lexical affinity models. In *Proceedings of COLING 2004*, Geneva, Switzerland.

Williams, Geoffrey (2003). Les collocations et l'école contextualiste britannique. In F. Grossmann and A. Tutin (eds.), *Les Collocations: analyse et traitement*, pages 33–44. De Werelt, Amsterdam.

Yates, F. (1984). Tests of significance for 2 × 2 contingency tables. *Journal of the Royal Statistical Society, Series A*, **147**(3), 426–463.

Zinsmeister, Heike and Heid, Ulrich (2003). Significant triples: Adjective+noun+verb combinations. In *Proceedings of the 7th Conference on Computational Lexicography and Text Research (COMPLEX 2003)*, pages 92–101, Budapest, Hungary.