Towards a Firthian Notion of Collocation

Sabine Bartsch, Technische Universität Darmstadt
Stefan Evert, Friedrich-Alexander-Universität Erlangen-Nürnberg

1. Introduction

Collocations are pervasive in language. According to Altenberg (1991: 128), “roughly 70% of the running words in the corpus form part of recurrent word combinations of some kind.” The investigation of such word combinations in corpora of authentic language dates back to the earliest studies of collocations by J. R. Firth (1957), who is commonly credited with introducing the concept within British Contextualism. However, serious corpus-based exploration of collocations on a larger scale has only become feasible with the arrival of the computer in the linguist’s workspace in the late 20th century. Since then, a substantial number of corpora of different sizes have become available, opening up new possibilities for collocation studies and many other linguistic applications. Progress has been made in particular by harnessing ever larger corpora, a growing range of statistical measures of association (cf. Evert 2004), and state-of-the-art software tools for automatic linguistic annotation and analysis.

The purpose of the research presented in this paper is to enhance our understanding of the role played by (i) the size and composition of the corpus (ranging from reasonably sized clean, balanced reference corpora to huge, messy Web collections), (ii) automatic linguistic annotation (part-of-speech tagging, syntactic parsing, etc.), and (iii) the mathematical properties of statistical association measures in the automatic extraction of collocations from corpora. In contrast to most prior comparative evaluation studies, which focused on the extraction of lexicalised multiword expressions relevant for traditional paper dictionaries, the present study builds upon a strictly Firthian (1951/1957) definition of collocation as the habitual and recurrent juxtaposition of words with particular other words. By this approach, we hope to complement work towards data acquisition for electronic dictionaries of the future with a closer look at a type of word combinatorics that despite some considerable progress so far has proven quite difficult to grasp.

The research presented in this paper was driven by three questions: Are bigger corpora always better, or are a balanced composition and clean data more important? To what extent does automatic linguistic annotation improve collocation identification and what annotation levels are most beneficial? Can we find evidence for the postulated presence of syntactic relations between collocates (Bartsch 2004: 79), in contrast to the traditional window-based operationalization (Sinclair 1966, 1991) of the Firthian notion of collocation?

2. The notion of collocation revisited

The concept of collocation is most commonly defined as a characteristic co-occurrence of lexical items, although definitions differ in a number of details. The definition advocated by J. R. Firth, who can be credited with systematically establishing the concept in modern linguistics, holds that collocations are to be defined as the habitual and recurrent juxtaposition of semantically related words. Definitions, furthermore, differ in terms of the number of postulated lexical items assumed as constituents of collocations. Hausmann (1985), for example, assumes a binary and directional relation and permits as constituents of collocations content words only to the exclusion of function words. However, function words are subsumed under the notion of collocation in Renouf/Sinclair's (1991: 128ff.) "collocational frameworks" exemplified by constructions such as 'a + … + of' as in 'a pride of lions', 'a pair of scissors' etc.

Halliday/Hasan (1976) deviate in their definition by describing collocations as "semantically related lexical items" which are more commonly interpreted in the sense of semantic field relations such as "doctor – hospital – nurse" co-occurring within the same context. Their definition rests on a tendency of lexical items to occur in the same context because they belong to the same semantic field. The common underlying assumption is that collocations are characteristic co-occurrences of related lexical items, a notion that can also be identified in Eugenio Coseriu's (1967) concept of "lexical solidarity".

Early approaches to the empirical study of collocation rest upon the manual identification of collocations in relatively small amounts of text (cf. Firth 1957), which are necessarily limited in scope and coverage. Since then, new research methods and tools as well as data have become available. The study of collocations has received fresh impetus through new computational approaches and the availability of large electronic text corpora, especially since the early 1990s. With the successively wider availability of ever larger corpora, studies of collocations have become feasible on a previously unknown scale, reaching a wider coverage of empirical data than ever before. Yet, in order to make such corpus-based studies possible and fruitful, the notion of collocation had to be not only defined, but also had to be operationalized. These new developments have brought about a further aspect in the definition of collocation, namely the definition and operationalization in terms of window-based approaches within the constrained context of a typically 3:3 or 5:5 key-word in context concordance window as proposed for example by Sinclair (1966, 1991). This approach has paved the way for the automatic investigation of collocations on the basis of relatively little linguistic pre-processing other than part of speech tagging, and by means of statistical methods modelling the characteristic co-occurrence of lexical items in terms of significance of co-occurrence as well as statistical measures of association. The current state of the art in statistical collocation identification will be discussed in the following section 3.

3. Statistical identification of collocations

Collocations are best studied on the basis of suitably large corpora of authentic language data. Their identification in corpora rests on linguistic hypotheses regarding the nature of collocations in terms of the co-occurrence of their constituents and the qualitative and quantitative relations between them. These hypotheses have to be operationalized so they can be systematically applied to corpus data; furthermore, suitable parameters have to be chosen in order to be able to distinguish instances of genuine collocations from false positives. These parameters are informed by features employed in linguistic definitions of collocations such as frequency of co-occurrence etc. In order to implement a Firthian definition of collocations, the explicitly mentioned parameter of a recurrent co-occurrence of lexical items translates directly into co-occurrence frequency in a corpus, where the context is usually taken to be a collocational span of 3 or 5 words to either side (we refer to this type of context as surface co-occurrence, following Evert 2008). For a meaningful interpretation, observed co-occurrence frequency has to be put in relation to the expected frequency of co-occurrence by chance, which can be computed from the individual frequencies of the two lexical items. For this purpose, a large number of mathematical formulae have been suggested as measures of statistical association.

Questions arising towards a corpus-based and quantitative definition of collocation entail the question of which association measure best captures our intuition of habitual, recurrent word combinations. More explicitly, the question which statistical measures are most suitable for the identification of Firthian collocations in corpora requires an answer. The parameters employed towards an operationalization of a Firthian notion of collocation as described above are thus to be implemented in terms of a suitable search space as exemplified most pervasively in terms of so-called window-based approaches which typically identify collocations as lexical co-occurrences within a 3:3 or 5:5 window; maximally, the sentence boundary is assumed as the upper limit for a collocational relation (textual co-occurrence); some approaches assume syntactic co-occurrence and postulate a syntactic relation as the context of co-occurrence (Bartsch 2004). Statistical measures are employed for the identification of collocations which, put in simple terms, rest on gauging the frequency of the co-occurrence of the collocation in relation to the independent frequencies of the constituent lexical items. A number of statistical measures have come to be used for these purposes over the years, among them widely applied measures such as the log-likelihood ratio, t-score, the Dice coefficient, and mutual information (MI), which is widely used in lexicographic contexts, as well as a number of variations of the MI formula (see Evert 2004 for a more detailed discussion). Despite the widespread use and discussion of statistically based studies of collocation, there has not, to our knowledge, been any systematic large-scale study resting on a Firthian notion of collocation. Studies typically take as their vantage point specific types of multi-word expressions (such as support verb constructions or verb-particle constructions, e.g. Baldwin 2008), or rely completely on intuitions of annotators (e.g. lexicographers' judgements). This study tries to avoid such initial assumptions and aims to evaluate association measures solely based on a Firthian notion of collocation, i.e. it works on the basis of co-occurrence in context without phraseological or lexicographical assumptions guiding the experiment. It studies the effects and suitability of the different association measures in correlation to their performance in different research settings concerning size and composition of the corpora employed, thus challenging the sometimes bluntly put assumption that when it comes to corpora bigger is always better. The research also incorporates an investigation of the effects of different levels of linguistic pre-processing and annotation on collocation identification and extraction quality. These latter sets of factors, corpus size and composition and linguistic pre-processing and annotation are of special relevance in order to gain a better understanding of properties of collocation as entailed in notions of collocation postulating that constituents of collocations must be assumed to be in a direct syntactic relation with one another (e.g. Daille 1994; Bartsch 2004).

The different statistical measures of collocations have, as yet, to be tested in terms of their performance as well as the impact of different factors concerning the corpora under study such as corpus size and composition and different types of corpus pre-processing and annotation ranging from lemmatized and part of speech tagged corpora to corpora that have undergone syntactic parsing and thus allow testing the above mentioned hypothesis of a direct syntactic relation as a constraint on relations obtaining between constituents of collocations.

The next section discusses the ways in which these statistical measures were tested on corpora of different size and composition and what types of linguistic pre-processing and annotation have an impact on the performance of different statistical measures for collocation identification.

4. Research set-up

The present study rests upon a study of collocations in authentic text corpora of different size and composition, which are annotated at different levels of linguistic organisation. The amount and depth of annotation range from entirely unannotated plain text corpora, over corpora with part of speech tagging and lemmatization, to syntactically parsed corpora. The corpora range in size between the 100-million-word British National Corpus and a Web corpus of ca. 2 billion words (ukWaC). Linguistic annotations include part-of-speech tagging, lemmatization and syntactic analysis by means of a dependency parser.

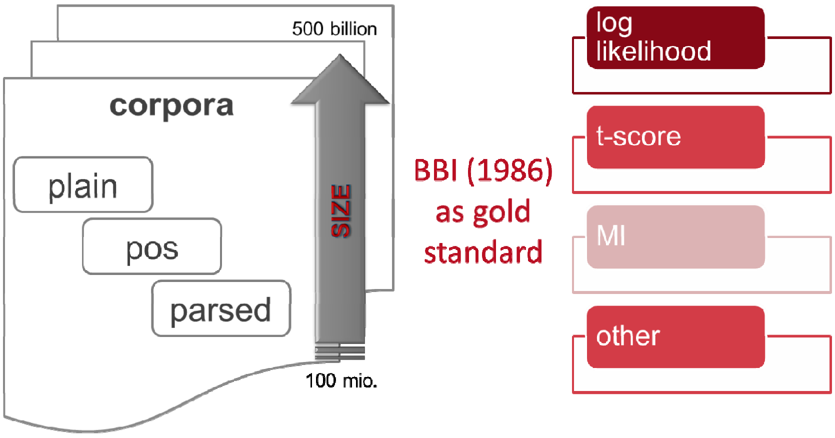

Figure 1: Research scenario

As a gold standard for our evaluation study, we use combinations of lexical words found in the entries of the BBI Combinatory Dictionary (Benson et al. 1986), a pre-corpus collocation dictionary that we consider to come very close to a Firthian definition of collocations. We compare (i) a standard range of well-known association measures (including log-likelihood, t-score, the Dice coefficient, co-occurrence frequency and different variants of mutual information), (ii) corpora of widely different sizes and composition (ranging from the 100-million-word British National Corpus to the ukWaC Web corpus comprising approx. 2 billion words), and (iii) different sizes of co-occurrence context (ranging from direct syntactic relations to co-occurrence within the same sentence) (see figure 1).

British National Corpus (BNC, Aston/Burnard 1998)	100 M
CLAWS tagger, lemmatised, C&C dependency parser (Curran/Clark/Bos 2007)	
Subset of Wackypedia (first 500 words of selected articles)	200 M
TreeTagger POS & lemmatisation (Schmid 1995), MaltParser (Nivre/Hall/Nilsson 2006)	
English Wikipedia @ 2009 (Wackypedia, available from wacky.sslmit.unibo.it)	850 M
TreeTagger POS & lemmatisation (Schmid 1995), MaltParser (Nivre/Hall/Nilsson 2006)	
Web corpus of British English (ukWaC, Baroni et al. 2009)	2,000 M
TreeTagger POS & lemmatisation (Schmid 1995), MaltParser (Nivre/Hall/Nilsson 2006)	

Table 1: Overview of corpora, corpus sizes and pre-processing and annotation

The quantitative task (section 5 below) entails the automatic identification of collocation candidates on the basis of well-known statistical association measures such as log-likelihood (G^2), t-score (t), Mutual Information (MI), and the Dice coefficient to name but some of the most widely used ones (see Evert 2008 for details). The aim of this aspect of the study is a more thorough investigation of the factors influencing results of collocation evaluation tasks as findings remain inconclusive as to which association measure is most useful depending on factors such as language, type of multiword expression as well as corpus size and composition. For example, MI is very popular among computational lexicographers despite a well-known bias towards low-frequency data, while computational linguists prefer measures based on statistical hypothesis tests such as G^2 (following Dunning 1993). Recent versions of the SketchEngine (Kilgarriff et al. 2004) use a variant of the Dice coefficient for collocation identification, even though it has never been identified as a top-performing measure in comparative evaluation studies.

Another issue, which has not been systematically addressed yet, concerns the ideal corpus size and composition for statistical collocation identification. More bluntly put, we ask whether "bigger is better", as the recent trend towards large Web corpora presupposes, or whether representativeness and high-quality data preparation are more important than sheer size.

In terms of corpora suitable for collocation analysis an additional issue concerns the question of the impact of amount and level of corpus annotation, i.e. the question in how far collocation studies can benefit from linguistic annotations. A Firthian notion of collocations assumes mere "habitual" co-occurrence and has usually been implemented for very practical reasons as co-occurrences of lexical items that occur a minimum number of times, usually at least five times, the practical side of this decision also being the reduction of the amount of data coming under study. Most definitions of collocation tacitly assume that collocations are potentially specific to the morpho-syntactic class of the word and that different word forms at least potentially enter into similar sets of collocations, although this latter assumption has been called into question on occasion; in order to study collocations across the entire paradigm of a lexical item, most studies are based on lemmatized and part of speech tagged corpora. At the extreme end of definitions, syntactic relations are posited to obtain between the constituents of a collocation and that these, consequently, are best identified on the basis of suitable syntactic annotations, typically either parse trees or dependency parses.

In this research set-up, we are envisioning all of these possibilities comparatively with the aim and hope that we can not only test the Firthian notion of collocation, but also to check whether definitions assuming more intricate linguistic relations between the constituents of collocations might not actually improve the quality of collocation extraction and thereby help to improve the coverage of collocations in lexicographical works based on automatic and statistically driven approaches.

In this study, we follow the well-established paradigm of Evert/Krenn (2001) for the comparative evaluation of association measures. In this approach, a set of candidate word pairs is ranked according to different association measures. The quality of each ranking is then assessed based on how many true collocations (true positives, TPs) are found among the *n* highest-ranked candidates (an *n*-best list), compared to a gold standard of known collocations. The percentage of TPs in such an *n*-best list is called the *n*-best precision of the ranking. Similarly, *n*-best recall is the percentage of true collocations in the gold standard that are found among the *n* highest-ranked candidates. Since it is not clear in most cases what number *n* of ranked candidates should be considered, evaluation results are usually presented visually by plotting *n*-best precision against *n*-best recall for many different *n*-best lists, producing a precision-recall curve as shown in Fig. 2. For example, the black line (for a ranking based on the log-likelihood measure) shows that if enough highest-ranking candidates are considered to find 10% of the gold standard collocations, the *n*-best precision is 30%. With this intuitive visualization, it is easy to compare the performance of the different association measures with all other factors (corpus size-composition, collocational span, corpus annotation) being equal.

As a gold standard, we use lexical collocations from the BBI Combinatory Dictionary (Benson et al. 1986), which we believe to correspond well to the Firthian definition of collocations and which avoid a bias in favour of a particular corpus or collocation identification method (unlike more recent corpus-based collocation dictionaries). Since the BBI is not available in electronic form, we manually selected 224 entries from the dictionary, based on corpus frequency of the headwords, examples from the literature and previous linguistic analyses of English collocations. All lexical collocates (i.e. nouns, verbs, adjectives and adverbs) listed in these entries were transcribed, resulting in a gold standard of 2,949 lexical collocations. Note that the collocations in the gold standard are directed, consisting of one of the 224 headwords and a lexical collocate. In total, there are 1,849 distinct collocates in the gold standard.

Candidate data were extracted from the four corpora listed in Table 1. In order to ensure a fair comparison between corpora of different size and composition, we prepared a list of 7,711 lexical words that are considered as potential collocates. This list includes all 1,849 collocates from the gold standard and was extended with nouns, verbs, adjectives and adverbs that fall into the same frequency range in the British National Corpus. Collocation candidates are thus all co-occurrences of one of the 224 headwords with one of the 7,711 potential collocates. No frequency thresholds were applied since the goal is to identify as many known collocations from the BBI as possible.

From each of the four corpora, we extracted collocation candidates for five different context settings:

- Surface co-occurrence with a collocational span of 3 words (L3 / R3)
- Surface co-occurrence with a collocational span of 5 words (L5 / R5)
- Surface co-occurrence with a collocational span of 10 words (L10 / R10)
- Textual co-occurrence within sentences
- Syntactic co-occurrence, where the two words must occur in a direct syntactic relation (according to the MaltParser or C&C analyses)

corpus / context	# candidates	coverage
BNC syntactic	201397	91.69%
BNC span 3 words	349372	95.56%
BNC span 5 words	455020	96.54%
BNC span 10 words	584723	97.49%
BNC sentence	735191	98.17%
Wackypedia 200M syntactic	239956	91.66%
Wackypedia 200M span 3 words	378468	94.17%
Wackypedia 200M span 5 words	483649	95.15%
Wackypedia 200M span 10 words	612099	96.41%
Wackypedia 200M sentence	766944	97.25%
Wackypedia syntactic	396504	96.64%
Wackypedia span 3 words	588229	97.73%
Wackypedia span 5 words	717994	98.27%
Wackypedia span 10 words	864365	98.68%
Wackypedia sentence	1034844	98.95%
ukWaC syntactic	544198	98.58%
ukWaC span 3 words	759158	99.15%
ukWaC span 5 words	898030	99.29%
ukWaC span 10 words	1048637	99.39%
ukWaC sentence	1256383	99.49%

Table 2: Number of candidates in each data set and coverage of the BBI-derived gold standard

Table 2 shows the number of collocation candidates found in each setting and the corresponding coverage of the BBI gold standard, i.e. how many of the BBI collocations co-occurred at least once in the respective corpus and type of context. With all coverage values well above 90%, it is possible at least in principle to extract Firthian collocations in the sense of the BBI even from a relatively small corpus such as the BNC.

Not surprisingly, coverage increases for larger corpora and context windows with a coverage of up to > 99% for the ukWaC Web corpus. It likewise comes as no surprise that the syntactic dependency filter reduces the coverage, partly due to collocation candidates that do not form grammatical units after all or for which a direct grammatical relation does not show in the parse tree, partly due to parsing errors. It should be noted that the values shown here include all combinations, even if they just occur once and thus cannot be recognised as collocational by a quantitative analysis.

The statistical association measures tested in this study were selected based on the recommendations of Evert (2008):

- log-likelihood measure (G^2) (Dunning 1993)
- t-score (*t*) (Church et al 1991; *pace* Evert 2004: 82f.)
- Mutual Information (*MI*) (Church / Hanks 1990)
- Dice coefficient (*Dice*) (as employed in the Sketch Engine; Kilgarriff et al. 2004)
- Ranking by co-occurrence frequency (*f*) as a baseline

In addition, we considered several variants of MI that aim to reduce its low-frequency bias by raising the observed co-occurrence frequency to the *k*-th power. From this family of MI^k measures we selected MI^2, since it consistently gave the best results in preliminary experiments.

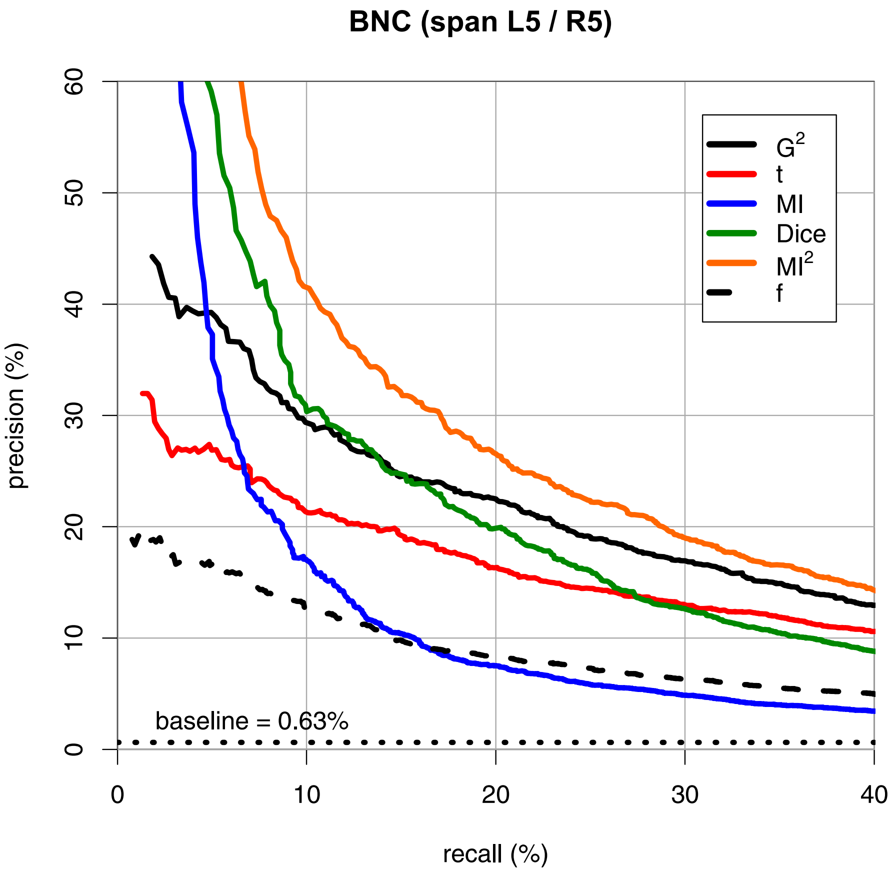

Figure 2: Evaluation of different association measures on British National Corpus with surface context (symmetric span of 5 tokens to the left and right)

Fig. 2 shows evaluation results for the British National Corpus, a “traditional” balanced reference corpus that has been widely used for collocation extraction and other tasks in computational lexicography. Candidates were obtained based on surface co-occurrence in a L5 / R5 window, which is a typical setting for “traditional” collocation studies. This evaluation generally conforms with our expectations and the results of previous studies: G^2 (black) and *t*-score (red) perform significantly better than plain co-occurrence frequency *f* (dashed black line). All association measures achieve much higher precision than the very low baseline corresponding to a random ranking of the candidates. The blue line once again confirms the well-known observation that MI performs poorly without frequency thresholds, due to its low-frequency bias. The Dice coefficient (green) displays surprisingly good performance and achieves higher accuracy than log-likelihood up to 15% recall; its performance drops drastically if a better coverage of the gold standard is required (above 25% recall). This observation explains why Dice was selected by the Sketch Engine developers: lexicographers reading the word sketches will typically focus on a few highly salient collocations. The uniformly best performance is achieved by MI^2 (orange) which outperforms the other association measures across all recall points. Results for other corpora and settings are qualitatively very similar; in almost all cases, MI^2 is the uniformly best association measure.

5. Quantitative insights

Ever since quantitative and statistical approaches have come to be applied to electronic corpora of substantial size, the dictum of "bigger is better" has been treated almost as a natural law. However, there have also always been cautioning voices warning that size might come with a price, especially concerning corpora whose sheer size forbids careful manual intervention to improve the quality of the data and the annotation. Furthermore, the very large corpora available today are typically collected opportunistically from the web and thus, despite all cleaning efforts, are neither balanced nor can the text base be assumed to be as clean as that of smaller, carefully crafted corpora such as the BNC. Nevertheless, there is a certain appeal to corpora that are big enough to promise to fulfil the dream of being representative or at least overcoming the obvious gaps of coverage in corpora of lesser size.

The study of lexical phenomena is a classical case where, due to the highly skewed distribution of lexical items, large corpora are often assumed to be of paramount importance, and the same assumption is consequently made for the study of collocations.

In order to test the effects of corpus size and quality on the accuracy of collocation identification, the different association measures were tested on the two-billion-word Web corpus ukWaC, which is opposite to the BNC at the "large and dirty" end of the spectrum. Figure 3 shows the results of this evaluation.

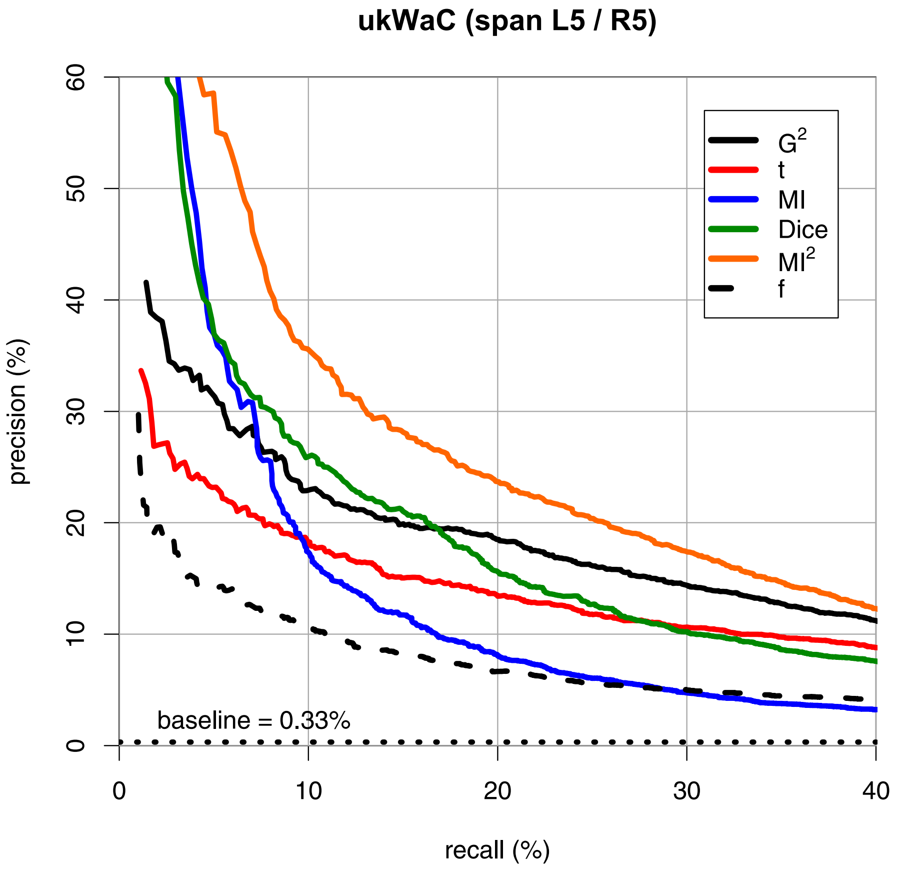

Figure 3: Evaluation of different association measures on the ukWaC Web corpus with surface context (symmetric span of 5 words to the left and right)

It turns out that the precision values are lower overall, but that the different association measures show a very similar pattern as compared to the BNC or, indeed, the other corpora under study. Again, MI^2 is uniformly the best-performing measure. These findings might lead one to modify the expectation of the impact of corpus size and quality to “bigger is worse” or “composition is more important than size”. The following experiments take this modified hypothesis into account and focus on MI^2 as an association measure.

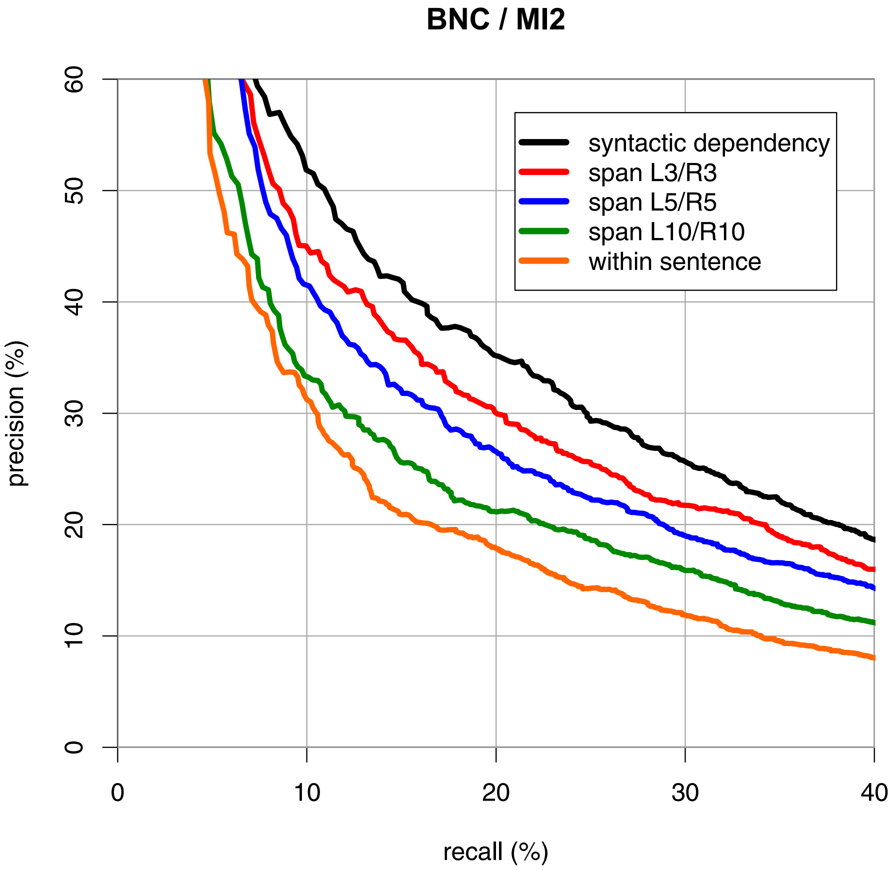

Figure 4: Comparison of different co-occurrence contexts for the British National Corpus and the best-performing association measure MI^2

The next question to be tested is the impact of the context type and size on the quality of the extraction task. In this set of experiments, contexts were tested ranging from narrow and specific (syntactic dependency) via increasing window sizes (L3 / R3, L5 / R5 and L10 / R10) to co-occurrence within a full sentence as the broadest context under consideration. Fig. 4 shows the results of this comparison for the British National Corpus.

These precision-recall graphs confirm that the larger the contexts become, the lower the precision drops. Thus, the smaller L3 / R3 span is better than the commonly used L5 / R5 span. Maybe surprisingly for some, the assumption of a syntactic relation between the constituents of collocations leads to even better results thus confirming the claims made by Bartsch (2004) that the Firthian notion of collocation should be supplemented with the additional criterion of a direct syntactic relation between the constituents.

The same pattern could be confirmed for all corpora, although not all plots are shown and discussed here in detail (due to space constraints): larger contexts lead to lower precision. It can thus be confirmed in turn that the component words of Firthian collocations tend to occur closely together and are usually in a direct syntactic relation. In the following experiment, we therefore focus on syntactic co-occurrence contexts for the comparison of different corpora.

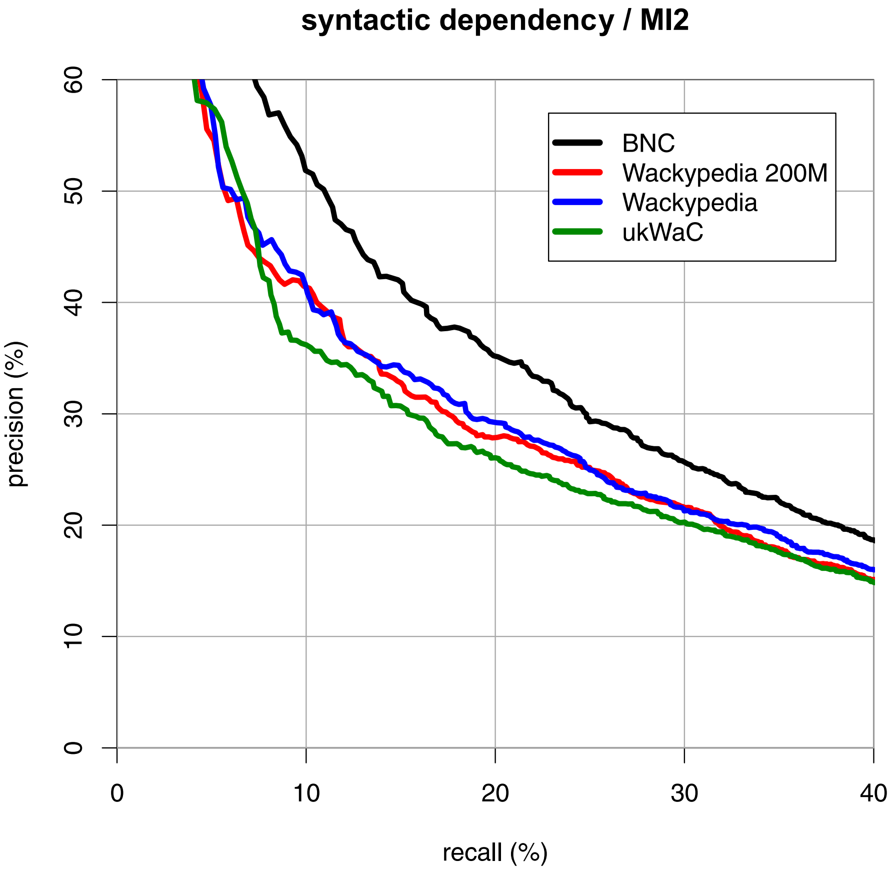

Figure 5: Comparison of different corpora for the best-performing combination of syntactic dependency co-occurrence and association measure MI^2

As already indicated above, initial observations suggest that increasing corpus size does not improve the performance of collocation extraction, especially if the larger corpora are less balanced and "clean". The graphs in Fig. 5 above confirm our new hypothesis that bigger is indeed worse. It appears that the larger and messier the corpora, the lower the precision drops. They also show that it is obviously composition and cleanliness that matter rather than size: the precision-recall curve for the full Wackypedia (approx. 850M words) and a subset of 200M words are practically identical, even though the former corpus is more than four times as large as the latter. The lowest precision is obtained from the large and messy Web corpus ukWaC.

However, there is one confounding factor that needs to be mentioned here and that requires further testing: ukWaC and the Wackypedia corpora were analysed by means of the Malt-Parser (syntactic annotations are included in the official distribution of the corpus) while the BNC was parsed by means of C&C, a sophisticated Combinatory Categorial Grammar (CCG) parser (Clark/Curran 2004; Curran/Clark/Bos 2007). There is thus a possibility that C&C is simply more accurate or covers important direct relations that are not generated by the Malt-Parser. Further testing of annotation quality is thus planned for the near future. In particular, we intend to re-annotate ukWaC and Wackypedia with C&C so that we will be able to gauge the impact of differences between the parsers.

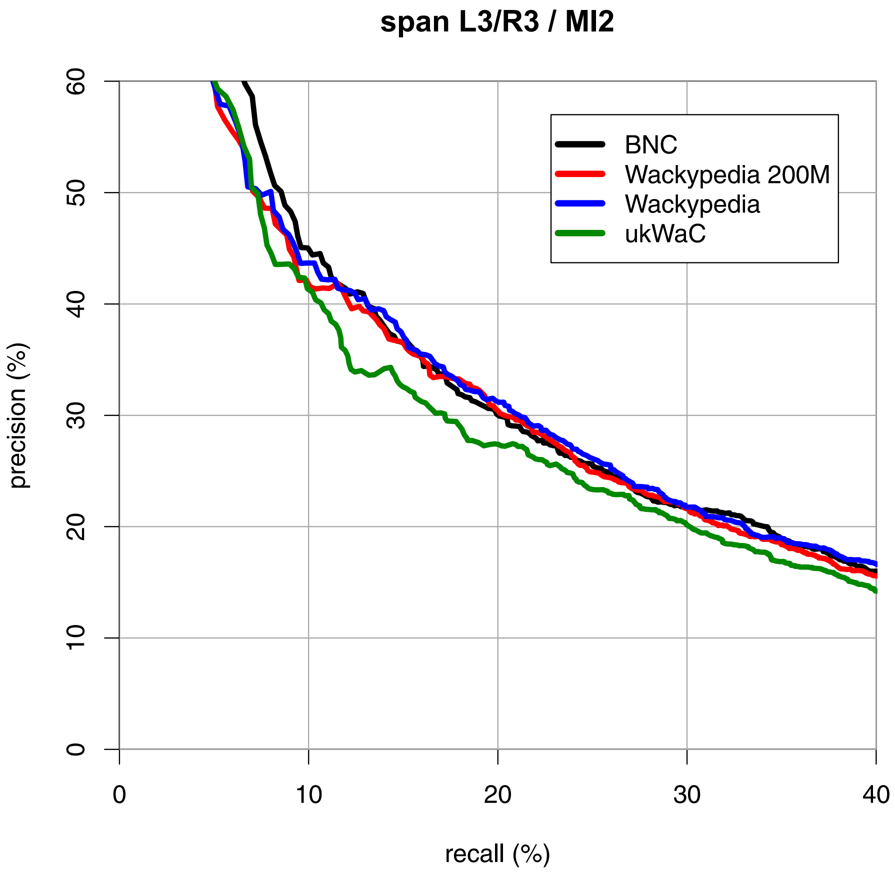

Figure 6: Comparison of different corpora for the best-performing collocational span (L3 / R3) and association measure MI^2

For the time being, in order to exclude the parser as a source of differences, we show a comparison on the basis of the smallest (and best-performing) surface context using a collocational span of 3 words (Fig. 6). In this experiment, there is no discernible difference between BNC, the Wackypedia subset and the full Wackypedia. The web corpus ukWaC even yields slightly worse results, despite its substantially larger size of approx. 2 billion words.

These findings lead to three conclusions: (i) increasing corpus size does not matter at all, at least for the identification of BBI collocations (since corpora ranging from 100M to 850M words yield virtually indistinguishable results); (ii) Wikipedia is a good replacement for the BNC as a reference corpus (at least with regard to Firthian collocations); (iii) the messiness of web data results in lower precision, even though the larger size should improve statistical analyses (and even though a web corpus might be expected to contain a broader range of genres than Wikipedia).

6. Conclusions

Based on the findings reported in this paper, we believe that some tentative conclusions can be drawn. The first one concerns corpus size where it could be shown that larger corpora do not necessarily lead to better results. This goes even for relatively simple statistical analyses (association measures) that should benefit from larger samples.

The second conclusion concerns corpus composition and suggests that composition and cleanness of a corpus are more important than corpus size. This is not to say that Web corpora might not be useful, but it suggests that their usefulness is enhanced if clean and more balanced samples can be obtained without compromising size (i.e. ≥ 2G words).

Collocations have been shown by some studies to tend to form grammatical relations, thus, assuming a syntactic dependency context is optimal for an identification of Firthian collocations. However, the practical benefit of taking this approach depends on the accuracy of the parser and the set of syntactic dependency relations recognized. In our case, it looks as if the linguistically optimised C&C parser is much better suited to the task than the fast off-the-shelf MaltParser used by the WaCky team.

The experiments with recent corpora such as the 200M subset of Wackypedia suggest that it is on par with BNC in collocation studies based on surface context. This result is very encouraging because it suggests that such corpora might be a useful substitute for languages such as German for which no standard general reference corpus like the British National Corpus is publicly available.

7. References

Altenberg, Bengt (1991): Amplifier collocations in Spoken English. In: Johansson, Stig / Stenström, Anna-Brita (eds.): English computer corpora. Selected papers and research guide. Berlin / New York: de Gruyter, p. 127-147.

Aston, Guy / Burnard, Lou (1998): The BNC Handbook. Edinburgh: Edinburgh University Press. http://www.natcorp.ox.ac.uk/ (last visited: 10.05.2013).

Baldwin, Timothy (2008): A resource for evaluating the deep lexical acquisition of English verb-particle constructions. In: Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008). Marrakech, Morocco, p. 1-2. Internet: http://www.lrec-conf.org/proceedings/lrec2008/workshops/W20_Proceedings.pdf (last visited: 10.05.2013).

Baroni, Marco et al. (2009): The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. In: Language Resources and Evaluation 43, 3, p. 209-226.

Bartsch, Sabine (2004): Structural and functional properties of collocations in English. A corpus study of lexical and pragmatic constraints on lexical co-occurrence. Tübingen: Narr.

Benson, Morton / Benson, Evelyn / Ilson, Robert (1986): The BBI Combinatory Dictionary of English: A Guide to Word Combinations. Amsterdam / New York: John Benjamins.

Church, Kenneth W. / Hanks, Patrick (1990): Word association norms, mutual information, and lexicography. In: Computational Linguistics 16, 1, p. 22-29.

Church, Kenneth / Gale, William A. / Hanks, Patrick / Hindle, Donald (1991): Using statistics in lexical analysis. In: Zernick, Uri (ed.): Lexical Acquisition: Using On-line Resources to Build a Lexicon. Hillsdale, NY: Lawrence Erlbaum, p. 115-164.

Clark, Stephen / Curran, James R. (2004): Parsing the WSJ using CCG and Log-Linear Models. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04). Barcelona, Spain 2004, p. 104-111. Internet: http://aclweb.org/anthology-new/P/P04/#1000 (last visited: 10.05.2013).

Coseriu, Eugenio (1967): Lexikalische Solidaritäten. In: Poetica 1, p. 293-203.

Curran, James / Clark, Stephen / Bos, Johan (2007): Linguistically motivated large-scale NLP with C&C and Boxer. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Sessions. Prague, Czech Republic. Madison: ACL, p. 33-36. Internet: http://www.aclweb.org/anthology-new/P/P07/P07-2.pdf (last visited: 10.05.2013).

Daille, Béatrice (1994): Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques. Ph.D. thesis, Université Paris 7. Internet: http://www.bdaille.com/index.php?option=com_docman&task=doc_download&gid=8&Itemid (last visited: 10.05.2013).

Dunning, Ted E. (1993): Accurate methods for the statistics of surprise and coincidence. In: Computational Linguistics 19, 1, p. 61-74.

Evert, Stefan / 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. Toulouse, France, p. 188-195. Internet: http://aclweb.org/anthology-new/P/P01/ (last visited: 10.05.2013).

Evert, Stefan (2004): The statistics of word cooccurrences: word pairs and collocations. Dissertation, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart. Veröffentlicht 2005. URN: urn:nbn:de:bsz:93-opus-23714.

Evert, Stefan (2008): Corpora and collocations. In: Lüdeling, Anke / Kytö, Merja (eds.): Corpus Linguistics. An International Handbook. Chapter 58. Berlin: de Gruyter.

Firth, John R. (1951/1957): Modes of meaning. In: Papers in Linguistics, 1934-1951. Oxford: Oxford University Press.

Halliday, MAK / Hassan, Ruqaiya (1976): Cohesion in English. London: Longman.

Hausmann, Franz Josef (1985): Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels. In: Bergenholtz, Henning / Mugdan, Joachim (eds.): Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch. (= Lexikographica, Series Maior 3). Tübingen: Niemeyer, p. 118-129.

Kilgarriff, Adam et al. (2004): The Sketch Engine. In: Williams, Geoffrey / Vessier, Sandra (eds.): EURALEX 2004 Proceedings. Lorient: UBS, S. 105-116. Internet: http://www.euralex.org/elx_proceedings/Euralex2004/ (last visited: 10.05.2013).

Nivre, Joakim / Hall, Johan / Nilsson, Jens (2006): MaltParser: a data-driven parser-generator for dependency parsing. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC2006). Genoa, Italy, p. 2216-2219.

Renouf, Antoinette / Sinclair, John (1991): Collocational frameworks in English. In: Aijmer, Karin / Altenberg, Bengt (eds.): English corpus linguistics. New York: Longman, p. 128-143.

Schmid, Helmut (1995): Improvements in part-of-speech tagging with an application to German. In: Proceedings of the ACL SIGDAT Workshop. Dublin, p. 47-50. Internet: ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger2.pdf (last visited: 10.05.2013).

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

Sinclair, John (1966): Beginning the study of lexis. In: Bazell, Charles E. et al. (eds.): In Memory of J. R. Firth. London: Longmans, p. 410-430.