A Measure of Syntactic Flexibility for Automatically Identifying Multiword Expressions in Corpora

Colin Bannard
Department of Developmental and Comparative Psychology
Max Planck Institute for Evolutionary Anthropology
Deutscher Platz 6
D-04103 Leipzig
colin.bannard@eva.mpg.de

Abstract

Natural languages contain many multi-word sequences that do not display the variety of syntactic processes we would expect given their phrase type, and consequently must be included in the lexicon as multiword units. This paper describes a method for identifying such items in corpora, focussing on English verb-noun combinations. In an evaluation using a set of dictionary-published MWEs we show that our method achieves greater accuracy than existing MWE extraction methods based on lexical association.

1 Introduction

A multi-word expression (henceforth MWE) is usually taken to be any word combination (adjacent or otherwise) that has some feature (syntactic, semantic or purely statistical) that cannot be predicted on the basis of its component words and/or the combinatorial processes of the language. Such units need to be included in any language description that hopes to account for actual usage. Lexicographers (for both printed dictionaries and NLP systems) therefore require well-motivated ways of automatically identifying units of interest. The work described in this paper is a contribution to this task.

Many linguists have offered classification schemes for MWEs. While these accounts vary in their terminology, they mostly focus on three different phenomena: collocation, non-compositionality and syntactic fixedness. In computational linguistics, a great deal of work has been done on the extraction of collocations in the last decade and a half (see Pecina (2005) for a survey). There have also been a number of papers focusing on the detection of semantic non-compositional items in recent years beginning with the work of Schone and Jurafsky (2001). The task of identifying syntactically-fixed phrases, however, has been much less explored. This third variety is the focus of the present paper. Languages contain many word combinations that do not allow the variation we would expect based solely on their grammatical form. In the most extreme case there are many phrases which seem to allow no syntactic variation whatsoever. These include phrases such as *by and large* and *in short*, which do not allow any morphological variation (**in shortest*) or internal modification (**by and pretty large*). We focus here on phrases that allow some syntactic variation, but do not allow other kinds.

The small amount of previous work on the identification of syntactic fixedness (Wermter and Hahn (2004), Fazly and Stevenson (2006)) has either focused on a single variation variety, or has only been evaluated for combinations of a small preselected list of words, presumably due to noise. In this paper we employ a syntactic parser, thus allowing us to include a wider range of syntactic features in our model. Furthermore we describe a statistical measure of variation that is robust enough to be freely evaluated over the full set of possible word combinations found in the corpus.

The remainder of our paper will be structured as follows. Section 2 will discuss the kinds of fixedness that we observe in our target phrase variety. Sec-

Proceedings of the Workshop on A Broader Perspective on Multiword Expressions, pages 1–8,
Prague, June 2007.

tion 3 will describe our model. Section 4 will evaluate the performance of the method and compare it to some other methods that have been described in the literature. Section 5 will describe some previous work on the problem, and section 6 will review our findings.

2 Syntactic Fixedness in English Verb Phrases

The experiments described here deal with one particular variety of phrase: English verb phrases of the form verb plus noun (e.g. *walk the dog*, *pull teeth*, *take a leaflet*). In a survey of the idiomatic phrases listed in the Collins Cobuild Dictionary of Idioms, Villavicencio and Copestake (2002) found this kind of idiom to account for more of the entries than any other. Riehemann (2001) performed a manual corpus analysis of verb and noun phrase idioms found in the Collins Cobuild Dictionary of Idioms. She found considerable fixedness with some phrases allowing no variation at all.

Based on this literature we identified three important kinds of non-morphological variation that such phrases can undergo, and which crucially have been observed to be restricted for particular combinations. These are as follows:

- Variation, addition or dropping of a determiner so that, for example, *run the show* becomes *run their show*, *make waves* becomes *make more waves*, or *strike a chord* becomes *strike chord* respectively.

- Modification of the noun phrase so that, for example, *break the ice* becomes *break the diplomatic ice*. We refer to this as internal modification.

- The verb phrase passivises so that, for example, *call the shots* is realised as *the shots were called by*.

3 Our Model

We use the written component of the BNC to make observations about the extent to which these variations are permitted by particular verb-noun combinations. In order to do this we need some way to a) identify such combinations, and b) identify when they are displaying a syntactic variation. In order to do both of these we utilise a syntactic parser.

We parse our corpus using the RASP system (Briscoe and Carroll, 2002). The system contains a LR probabilistic parser, based on a tag-sequence grammar. It is particularly suited to this task because unlike many contemporary parsers, it makes use of no significant information about the probability of seeing relationships between particular lexical items. Since we are looking here for cases where the syntactic behaviour of particular word combinations deviates from general grammatical patterns, it is desirable that the analysis we use has not already factored in lexical information. Example output can be seen in figure 1. We extract all verb and nouns pairs connected by an object relation in the parsed corpus. We are interested here in the object relationship between *buy* and *apartment*, and we can use the output to identify the variations that this phrase displays.

The first thing to note is that the phrase is passivised. *Apartment* is described as an object of *buy* by the "obj" relation that appears at the end of the line. Because of the passivisation, *apartment* is also described as a non-clausal subject of *buy* by the "ncmod" relation that appears at the beginning of the line. This presence of a semantic object that appears as a surface subject tells us that we are a dealing with a passive. The "ncmod" relation tells us that the adjective *largest* is a modifier of *apartment*. And finally, the "detmod" relation tells us that *the* is a determiner attached to *apartment*. We make a count over the whole corpus of the number of times each verb-object pair occurs, and the number of times it occurs with each relation of interest.

For passivisation and internal modification, a variation is simply the presence of a particular grammatical relation. The addition, dropping or variation of a determiner is not so straightforward. We are interested in the frequency with which each phrase varies from its dominant determiner status. We need therefore to determine what this dominant status is for each item. A verb and noun object pair where the noun has no determiner relation is recorded as having no determiner. This is one potential determiner status. The other varieties of status are defined by the kind of determiner that is appended. The RASP parser uses the very rich CLAWS-2 tagset. We con-

```
(|ncsubj| |buy+ed:6_VVN| |apartment:3_NN1| |obj|)
(|arg_mod| |by:7_II| |buy+ed:6_VVN| |couple:10_NN1| |subj|)
(|ncmod| _ |apartment:3_NN1| |largest:2_JJT|)
(|detmod| _ |apartment:3_NN1| |The:1_AT|)
(|ncmod| _ |couple:10_NN1| |Swedish:9_JJ|)
(|detmod| _ |couple:10_NN1| |a:8_AT1|)
(|mod| _ |buy+ed:6_VVN| |immediately:5_RR|)
(|aux| _ |buy+ed:6_VVN| |be+ed:4_VBDZ|)
```

Figure 1: RASP parse of sentence *The largest apartment was immediately bought by a Swedish couple*.

sider each of these tags as a different determiner status. Once the determiner status of all occurrences has been recorded, the dominant status for each item is taken to be the status that occurs most frequently. The number of variations is taken to be the number of times that the phrase occurs with any other status.

3.1 Quantifying variation

We are interested here in measuring the degree of syntactic variation allowed by each verb-object pair found in our corpus. Firstly we use the counts that we extracted above to estimate the probability of each variation for each combination, employing a Laplace estimator to deal with zero counts.

A straightforward product of these probabilities would give us the probability of free variation for a given verb-object pair. We need, however, to consider the fact that each phrase has a prior probability of variation derived from the probability of variation of the component words. Take passivisation for example. Some verbs are more prone to passivisation than others. The degree of passivisation of a phrase will therefore depend to a large extent upon the passivisation habits of the component verb.

What we want is an estimate of the extent to which the probability of variation for that combination deviates from the variation we would expect based on the variation we observe for its component words. For this we use conditional pointwise mutual information. Each kind of variation is associated with a single component word. Passivisation is associated with the verb. Internal modification and determiner variation are associated with the object. We calculate the mutual information of the syntactic variation x and the word y given the word z, as seen in equation 1. In the case of passivisation z will be the verb and y will be the object. In the case of internal modification and determiner variation z will be the object.

$$\begin{aligned} I(x;y|z) &= H(x|z) - H(x|y,z) \\ &= -\,log_2\, p(x|z) - [-\log_2\ p(x|y,z)] \\ &= -\,log_2\, p(x|z) + \log_2\ p(x|y,z) \\ &= log_2 \frac{p(x|y,z)}{p(x|z)} \end{aligned} \tag{1}$$

Conditional pointwise mutual information tells us the amount of information in bits that y provides about x (and vice versa) given z (see e.g. MacKay (2003)). If a variation occurs for a given word pair with greater likelihood than we would expect based on the frequency of seeing that same variation with the relevant component word, then the mutual information will be high. We want to find the information that is gained about all the syntactic variations by a particular verb and object combination. We therefore calculate the information gained about all the verb-relevant syntactic variations (passivisation) by the addition of the object, and the information gained about all the object relevant variations (internal modification and determiner dropping, variation or addition) by the addition of the verb. Summing these, as in equation 2 then gives us the total information gained about syntactic variation for the word pair W, and we take this as our measure of the degree of syntactic flexibility for this pair.

$$\begin{aligned} SynVar(W) = &\sum_{i}^{n} I(VerbVar_i; Obj|Verb) \\ + &\sum_{j}^{n} I(ObjVar_j; Verb|Obj) \end{aligned} \tag{2}$$

4 Evaluation

This paper aims to provide a method for highlighting those verb plus noun phrases that are syntactically fixed and consequently need to be included in the lexicon. This is intended as a tool for lexicographers. We hypothesize that in a list that has been inversely ranked with the variability measure valid MWEs will occur at the top.

The evaluation procedure used here (first suggested by Evert and Krenn (2001) for evaluating measures of lexical association) involves producing and evaluating just such a ranking. The RASP parser identifies 979,156 unique verb-noun pairs in the BNC. The measure of syntactic flexibility was used to inverse rank these items (the most fixed first).[1] This ranking was then evaluated using a list of idioms taken from published dictionaries, by observing how many of the gold standard items were found in each top n, and calculating the accuracy score. [2] By reason of the diverse nature of MWEs, these lists can be expected to contain many MWEs that are not syntactically fixed, giving us a very low upper bound. However this seems to us the evaluation that best reflects the application for which the measure is designed. The list of gold standard idioms we used were taken from the Longman Dictionary of English idioms (Long and Summers, 1979) and the SAID Syntactically Annotated Idiom Dataset (Kuiper et al., 2003). Combining the two dictionaries gave us a list of 1109 unique verb-noun pairs, 914 of which were identified in the BNC.

In order to evaluate the performance of our technique it will be useful to compare its results with the ranks of scores that can be obtained by other means. A simple method of sorting items available to the corpus lexicographer that might expected to give reasonable performance is item frequency. We take this as our baseline. In the introduction we referred to multiple varieties of MWE. One such variety is the collocation. Although the collocation is a different variety of MWE, any dictionary will contain collocations as well as syntactically fixed phrases. The collocation has received more attention than any other variety of MWE and it will therefore be useful to compare our measure with these methods as state-of-the-art extraction techniques. We report the performance obtained when we rank our candidate items using all four collocation extraction techniques described in Manning and Schutze (1999) : t-score, mutual information, log likelihood and χ^2.

4.1 Results

Figure 2 provides a plot of the accuracy score each sample obtains when evaluated using the superset of the two dictionaries for all samples from n = 1 to n = 5,000.

Included in figure 2 are the scores obtained when we inverse ranked using the variation score for each individual feature, calculated with equation 1. There is notable divergence in the performance of the different features. The best performing feature is passivisation, followed by internal modification. Determiner variation performs notably worse for all values of n.

We next wanted to look at combinations of these features using equation 2. We saw that the various syntactic variations achieved very different scores when used in isolation, and it was by no means certain that combining all features would the best approach. Nonetheless we found that the best scores were achieved by combining all three - an accuracy of 18%, 14.2 and 5.86% for n of 100, 1000 and 5000 respectively. This can be see in figure 2. The results achieved with frequency ranking can also be seen in the plot.

The accuracy achieved by the four collocation measures can be seen plotted in figure 3. The best performers are the t-score and the log-likelihood ratio, with MI and χ-squared performing much worse. The best score for low values of n is t-score, with log-likelihood overtaking for larger values. The best performing collocation measures often give a performance that is only equal to and often worse than raw frequency. This is consistent with results reported by Evert and Krenn (2001). Our best syntactic variation method outperforms all the collocation extraction techniques.

We can see, then, that our method is outperforming frequency ranking and the various collocation measures in terms of accuracy. A major claim we

[1] Any ties were dealt with by generating a random number for each item and ranking the drawn items using this.

[2] Note that because the number of candidate items in each sample is fixed, the relative performance of any two methods will be the same for recall as it is for precision. In such circumstances the term accuracy is preferred.

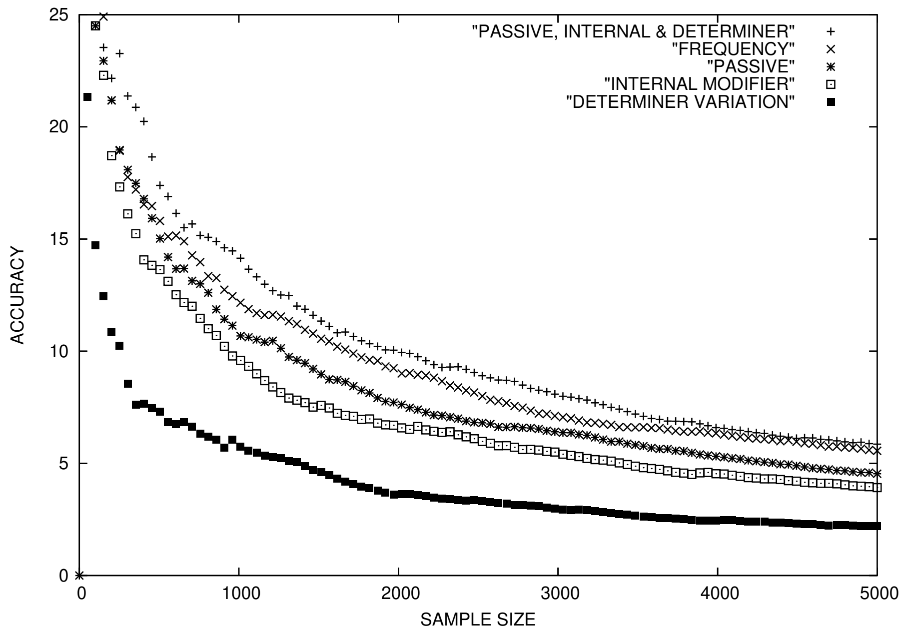

Figure 2: Accuracy by sample size for syntactic variation measures

are making for the method however is that it extracts a different kind of phrase. A close examination tells us that this is the case. Table 1 lists the top 25 verb-noun combinations extracted using our best performing combination of features, and those extracted using frequency ranking. As can be see there is no overlap between these lists. In the top 50 items there is an overlap of 3 between the two lists. Over the top 100 items of the two lists there is only an overlap of 6 items and over the top 1000 there is an overlap of only 98.

This small overlap compares favourably with that found for the collocation scores. While they produce ranks that are different from pure frequency, the collocation measures are still based on relative frequencies. The two high-performing collocation measures, t-score and log-likelihood have overlap with frequency of 795 and 624 out of 1000 respectively. This tells us that the collocation measures are significantly duplicating the information available from frequency ranking. The item overlap between t-score items and those extracted using the the best-performing syntactic variation measure is 116. The overlap between syntactic variation and log-likelihood items is 108. This small overlap tells us that our measure is extracting very different items from the collocation measures.

Given that our measure appears to be pinpointing a different selection of items from those highlighted by frequency ranking or lexical association, we next want looked at combining the two sources of information. We test this by ranking our candidate list using frequency and using the most consistently well-performing syntactic variation measure in two separate runs, and then adding together the two ranks achieved using the two methods for each item. The items are then reranked using the resulting sums. When this ranking is evaluated against the dictionaries it gives the scores plotted in figure 3 - a clearly better performance than syntactic fixedness or frequency alone for samples of 1000 and above.

Having reported all scores we now want to measure whether any of them are beating frequency ranking at a level that is statistically significant.

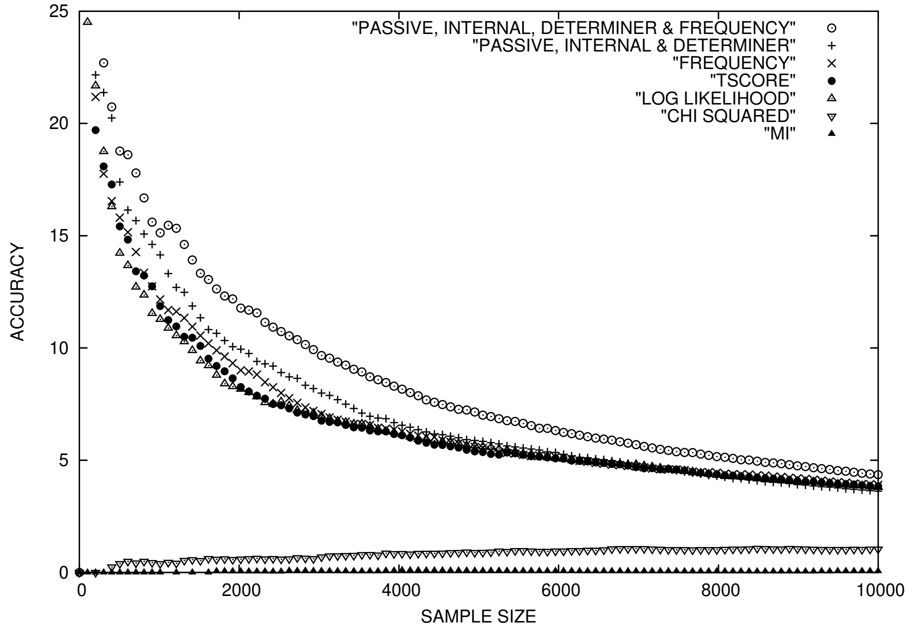

Figure 3: Accuracy by sample size for lexical association measures

In order to do this we pick three values of n (100,1000 and 5000) and examine whether the accuracy achieved by our method are greater than those achieved with frequency ranking at a level that is significantly greater than chance. Conventional significance testing is problematic for this task. Rather than using a significance test that relies upon an assumed distribution, then, we will use a computationally-intensive randomization test of significance called stratified shuffling. This technique works by estimating the difference that might occur between scores by chance through a simulation (see (Cohen, 1995) for details). As is standard we perform 10,000 shuffle iterations.

The results for our three chosen values of n can be seen in table 2. We accept any result of $p < 0.05$ as significant, and scores that achieve this level of significance are shown in bold. As an additional check on performance we also extend our evaluation. In any evaluation against a gold standard resource, there is a risk that the performance of a technique is particular to the lexical resource used and will not generalise. For this reason we will here report results achieved using not only the combined set but also each dictionary in isolation. If the technique is effective then we would expect it to perform well for both resources.

We can see that our syntactic variation measures perform equal to or better than frequency over both dictionaries in isolation for samples of 1000 and 5000. The good performance against two data sets tells us that the performance does generalise beyond a single resource. For the Longman dictionary, the accuracy achieved by the syntactic variation measure employing the three best performing features ("P, I and D") is significantly higher (at a level of $p < 0.05$) than that achieved when ranking with frequency for sample sizes of 1000 and 5000. The ranking achieved using the combination of syntactic fixedness and frequency information produces a result that is significant over all items for samples of 1000 and 5000. By contrast, none of the collocation scores perform significantly better than frequency. [3]

[3] As very low frequency items have been observed to cause

		Syntactic Variation		Collocation			
DICTIONARY	Freq	P,I &D	P,I,D &Freq	t	MI	LLR	χ^2
Top 100 items							
LONGMANS	14	21	15	16	0	13	0
SAID	21	17	17	23	0	17	0
BOTH	28	18	25	32	0	25	0
Top 1000 items							
LONGMANS	6.6	**10.4**	**10.2**	6.3	0	6.5	0.3
SAID	9.1	9	9.9	9	0	8.1	0.2
BOTH	12.2	14.2	**15.2**	12	0	11.4	0.4
Top 5000 items							
LONGMANS	3.24	**4.28**	**4.84**	3.12	0.06	3.44	0.58
SAID	3.86	3.56	**4.54**	3.68	0.04	3.86	0.54
BOTH	5.56	5.86	**7.68**	5.34	0.04	5.66	0.88

Table 2: Accuracy for top 100, 1000 and 5000 items (scores beating frequency at $p < 0.05$ are in bold)

An important issue for future research is how much the performance of our measure is affected by the technology used. In an evalutaion of RASP, Preiss (2003) reports an precision of 85.83 and recall of 78.48 for the direct object relation, 69.45/57.72 for the "ncmod" relation, and 91.15/98.77 for the "detmod" relation. There is clearly some variance here, but it is not easy to see any straightforward relationship with our results. The highest performance relation ("detmod") was our least informative feature. Meanwhile our other two features both rely on the "ncmod" relation. One way to address this issue in future research will be to replicate using multiple parsers.

5 Previous work

Wermter and Hahn (2004) explore one kind of syntactic fixedness: the (non-)modifiability of preposition-noun-verb combinations in German. They extract all preposition-noun-verb combinations from a corpus of German news text, and identify all the supplementary lexical information that occurs between the preposition and the verb. For each phrase they calculate the probability of seeing each piece of supplementary material, and take this as its degree of fixedness. A final score is then calculated by taking the product of this score and the probability of occurrence of the phrase. They then manually evaluated how many true MWEs occurred in the top n items at various values of n. Like us they report that their measure outperformed t-score, log likelihood ratio and frequency.

Fazly and Stevenson (2006) propose a measure for detecting the syntactic fixedness of English verb phrases of the same variety as us. They use a set of regular patterns to identify, for particular word combinations (including one of a chosen set of 28 frequent "basic" verbs), the probability of occurrence in passive voice, with particular determiners and in plural form. They then calculate the relative entropy of this probability distribution for the particular word pair and the probabilities observed over all the word combinations. As we pointed out in section 3.1 a comparison with all verbs is problematic as each verb will have its own probability of variation, and this perhaps explains their focus on a small set of verbs. They use a development set to establish a threshold on what constitutes relative fixedness and calculate the accuracy. This threshhold gives over the set of 200 items, half of which were found in a dictionary and hence considered MWEs and half weren't. They report an accuracy of 70%, against a 50% baseline. While this is promising, their use of a small selection of items of a particular kind in their evaluation makes it somewhat difficult to assess.

problems for collocation measures, we experimented with various cutoffs up to an occurence rate of 5. We found that this did not lead to any significant difference from frequency.

	FREQUENCY	P,I & D
1	take place	follow suit
2	have effect	draw level
3	shake head	give rise
4	have time	part company
5	take part	see chapter
6	do thing	give moment
7	make decision	open fire
8	have idea	run counter
9	play role	take refuge
10	play part	clear throat
11	open door	speak volume
12	do job	please contact
13	do work	leave net
14	make sense	give way
15	have chance	see page
16	make use	catch sight
17	ask question	cite argument
18	spend time	see table
19	take care	check watch
20	have problem	list engagement
21	take step	go bust
22	take time	change subject
23	take action	change hand
24	find way	keep pace
25	have power	see paragraph

Table 1: Top 25 phrases

6 Discussion

Any lexicon must contain multiword units as well as individual words. The linguistic literature contains claims for the inclusion of multiword items in the lexicon on the basis of a number of linguistic dimensions. One of these is syntactic fixedness. This paper has shown that by quantifying the syntactic fixedness of verb-noun phrases we can identify a gold standard set of dictionary MWEs with a greater accuracy than the lexical association measures that have hitherto dominated the literature, and that, perhaps more crucially, we can identify a different set of expressions, not available using existing techniques.

Acknowledgements

Thanks to Tim Baldwin, Francis Bond, Ted Briscoe, Chris Callison-Burch, Mirella Lapata, Alex Lascarides, Andrew Smith, Takaaki Tanaka and two anonymous reviewers for helpful ideas and comments.

References

Ted Briscoe and John Carroll. 2002. Robust accurate statistical annotation of general text. In *Proceedings of LREC-2003*.

P. Cohen. 1995. *Empirical Methods for Artificial Intelligence*. MIT Press.

Stefan Evert and Brigitte Krenn. 2001. Methods for the qualitative evaluation of lexical association measures. In *Proceedings of ACL-2001*.

Afsaneh Fazly and Suzanne Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proceedings of EACL-2006*.

Koenraad Kuiper, Heather McCann, and Heidi Quinn. 2003. A syntactically annotated idiom database (said), v,1.

Thomas H. Long and Della Summers. 1979. *Longman Dictionary of English Idioms*. Longman Dictionaries.

David J.C. MacKay. 2003. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press.

C. Manning and H. Schutze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, USA.

Pavel Pecina. 2005. An extensive empirical study of collocation extraction methods. In *Proceedings of the ACL-2005 Student Research Workshop*.

Judita Preiss. 2003. Using grammatical relations to compare parsers. In *Proceedings of EACL-03*.

Suzanne Riehemann. 2001. *A Constructional Approach to Idioms and Word Formation*. Ph.D. thesis, Stanford University.

Patrick Schone and Dan Jurafsky. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proceedings of EMNLP-2001*.

Aline Villavicencio and Ann Copestake. 2002. On the nature of idioms. *LinGO Working Paper No. 2002-04*.

Joachim Wermter and Udo Hahn. 2004. Collocation extraction based on modifiability statistics. In *Proceedings of COLING-2004*.