Validation and Evaluation of Automatically Acquired Multiword Expressions for Grammar Engineering

Aline Villavicencio♣♠, Valia Kordoni♢, Yi Zhang♢, Marco Idiart♡ and Carlos Ramisch♣

♣Institute of Informatics, Federal University of Rio Grande do Sul (Brazil)
♠Department of Computer Sciences, Bath University (UK)
♢Department of Computational Linguistics, Saarland University, and DFKI GmbH (Germany)
♡Institute of Physics, Federal University of Rio Grande do Sul (Brazil)
avillavicencio@inf.ufrgs.br, {yzhang,kordoni}@coli.uni-sb.de
idiart@if.ufrgs.br, ceramisch@inf.ufrgs.br

Abstract

This paper focuses on the evaluation of methods for the automatic acquisition of Multiword Expressions (MWEs) for robust grammar engineering. First we investigate the hypothesis that MWEs can be detected by the distinct statistical properties of their component words, regardless of their type, comparing 3 statistical measures: mutual information (MI), χ^2 and permutation entropy (PE). Our overall conclusion is that at least two measures, MI and PE, seem to differentiate MWEs from non-MWEs. We then investigate the influence of the size and quality of different corpora, using the BNC and the Web search engines Google and Yahoo. We conclude that, in terms of language usage, web generated corpora are fairly similar to more carefully built corpora, like the BNC, indicating that the lack of control and balance of these corpora are probably compensated by their size. Finally, we show a qualitative evaluation of the results of automatically adding extracted MWEs to existing linguistic resources. We argue that such a process improves qualitatively, if a more compositional approach to grammar/lexicon automated extension is adopted.

1 Introduction

The task of automatically identifying Multiword Expressions (MWEs) like phrasal verbs (*break down*) and compound nouns (*coffee machine*) using statistical measures has been the focus of considerable investigative effort, (e.g. Pearce (2002), Evert and Krenn (2005) and Zhang et al. (2006)). Given the heterogeneousness of the different phenomena that are considered to be MWEs, there is no consensus about which method is best suited for which type of MWE, and if there is a single method that can be successfully used for any kind of MWE.

Another difficulty for work on MWE identification is that of the evaluation of the results obtained (Pearce, 2002; Evert and Krenn, 2005), starting from the lack of consensus about a precise definition for MWEs (Villavicencio et al., 2005).

In this paper we investigate some of the issues involved in the evaluation of automatically extracted MWEs, from their extraction to their subsequent use in an NLP task. In order to do that, we present a discussion of different statistical measures, and the influence that the size and quality of different data sources have. We then perform a comparison of these measures and discuss whether there is a single measure that has good overall performance for MWEs in general, regardless of their type. Finally, we perform a qualitative evaluation of the results of adding automatically extracted MWEs to a linguistic resource, taking as basis for the evaluation the approach proposed by Zhang et al. (2006). We argue that such results can improve in quality if a more compositional approach to MWE encoding is adopted for the grammar extension. Having more accurate means of deciding for an appropriate method for identifying and incorporating MWEs is critical for maintaining the quality of linguistic resources for precise NLP.

This paper starts with a discussion of MWEs (§ 2), of their coverage in linguistic resources (§ 3), and of some methods proposed for automatically identifying them (§ 4). This is followed by a detailed investigation and comparison of measures for MWE identification (§ 5).

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1034–1043, Prague, June 2007.

After that we present an approach for predicting appropriate lexico-syntactic categories for their inclusion in a linguistic resource, and an evaluation of the results in a parsing task(§ 7). We finish with some conclusions and discussion of future work.

2 Multiword Expressions

The term Multiword Expressions has been used to describe expressions for which the syntactic or semantic properties of the whole expression cannot be derived from its parts (Sag et al., 2002), including a large number of related but distinct phenomena, such as phrasal verbs (e.g. *come along*), nominal compounds (e.g. *frying pan*), institutionalised phrases (e.g. *bread and butter*), and many others. Jackendoff (1997) estimates the number of MWEs in a speaker's lexicon to be comparable to the number of single words. However, due to their heterogeneous characteristics, MWEs present a tough challenge for both linguistic and computational work (Sag et al., 2002). For instance, some MWEs are fixed, and do not present internal variation, such as *ad hoc*, while others allow different degrees of internal variability and modification, such as *spill beans* (*spill several/musical/mountains of beans*).

Sag et al. (2002) discuss two main approaches commonly employed in NLP for treating MWEs: the *words-with-spaces* approach models an MWE as a single lexical entry and it can adequately capture fixed MWEs like *by and large*. A *compositional* approach treats MWEs by general and compositional methods of linguistic analysis, being able to capture more syntactically flexible MWEs, like *rock boat*, which cannot be satisfactorily captured by a words-with-spaces approach, since it would require lexical entries to be added for all the possible variations of an MWE (e.g. *rock/rocks/rocking this/that/his... boat*). Therefore, to provide a unified account for the detection and encoding of these distinct but related phenomena is a real challenge for NLP systems.

3 Grammar and Lexicon Coverage in Deep Processing

Many NLP tasks and applications, like Parsing and Machine Translation, depend on large-scale linguistic resources, such as electronic dictionaries and grammars for precise results. Several substantial resources exist: e.g., hand-crafted large-scale grammars like the English Resource Grammar (ERG - Flickinger (2000)) and the Dutch Alpino Grammar (Bouma et al., 2001).

Unfortunately, the construction of these resources is the manual result of human efforts and therefore likely to contain errors of omission and commission (Briscoe and Carroll, 1997). Furthermore, due to the open-ended and dynamic nature of languages, such linguistic resources are likely to be incomplete, and manual encoding of new entries and constructions is labour-intensive and costly.

Take, for instance, the coverage test results for the ERG (a broad-coverage precision HPSG grammar for English) on the British National Corpus (BNC). Baldwin et al. (2004), among many others, have investigated the main causes of parse failure, parsing a random sample of 20,000 strings from the written component of the BNC using the ERG. They have found that the large majority of failures is caused by missing lexical entries, with 40% of the cases, and missing constructions, with 39%, where missing MWEs accounted for 8% of total errors. That is, even by a margin, the lexical coverage is lower than the grammar construction coverage.

This indicates the acute need for robust (semi-)automated ways of acquiring lexical information for MWEs, and this is the one of the goals of this work. In the next section we discuss some approaches that have been developed in recent years to (semi-)automatically detect and/or repair lexical and grammar errors in linguistic grammars and/or extend their coverage.

4 Acquiring MWEs

The automatic acquisition of specific types of MWE has attracted much interest (Pearce, 2002; Baldwin and Villavicencio, 2002; Evert and Krenn, 2005; Villavicencio, 2005; van der

Beek, 2005; Nicholson and Baldwin, 2006). For instance, Baldwin and Villavicencio (2002) proposed a combination of methods to extract Verb-Particle Constructions (VPCs) from unannotated corpora, that in an evaluation on the Wall Street Journal achieved 85.9% precision and 87.1% recall. Nicholson and Baldwin (2006) investigated the prediction of the inherent semantic relation of a given compound nominalization using as statistical measure the confidence interval.

On the other hand, Zhang et al. (2006) looked at MWEs in general investigating the semi-automated detection of MWE candidates in texts using error mining techniques and validating them using a combination of the World Wide Web as a corpus and some statistical measures. 6248 sentences were then extracted from the BNC; these contained at least one of the 311 MWE candidates verified with World Wide Web in the way described in Zhang et al. (2006). For each occurrence of the MWE candidates in this set of sentences, the lexical type predictor proposed in Zhang and Kordoni (2006) predicted a lexical entry candidate. This resulted in 373 additional MWE lexical entries for the ERG grammar using a words-with-spaces approach. As reported in Zhang et al. (2006), this addition to the grammar resulted in a significant increase in grammar coverage of 14.4%. However, no further evaluation was done of the results of the measures used on the identification of MWEs or of the resulting grammar, as not all MWEs can be correctly handled by the simple words-with-spaces approach (Sag et al., 2002). And these are the starting points of the work we are reporting on here.

5 Evaluation of the Identification of MWEs

One way of viewing the MWE identification task is, given a list of sequences of words, to distinguish those that are genuine MWEs (e.g. *in the red*), from those that are just sequences of words that do not form any kind of meaningful unit (e.g. *of alcohol and*). In order to do that, one commonly used approach is to employ statistical measures (e.g. Pearce (2002) for collocations and Zhang et al. (2006) for MWEs in general). When dealing with statistical analysis there are two important statistical questions that should be addressed: *How reliable is the corpus used?* and *How precise is the chosen statistical measure to distinguish the phenomena studied?*.

In this section we look at these issues, for the particular case of trigrams, by testing different corpora and different statistical measures. For that we use 1039 trigrams that are the output of Zhang et al. (2006) error mining system, and frequencies collected from the BNC and from the World Wide Web. The former were collected from two different portions of the BNC, namely the fragment of the BNC (BNC_f) used in the error-mining experiments, and the complete BNC (from the site http://pie.usna.edu/), to test whether a larger sample of a more homogeneous and well balanced corpus improves results significantly. For the latter we used two different search engines: Google and Yahoo, and the frequencies collected reflect the number of pages that had exact matches of the n-grams searched, using the API tools for each engine.

5.1 Comparing Corpora

A corpus for NLP related work should be a reliable sample of the linguistic output of a given language. For this work in particular, we expect that the relative ordering in frequency for different n-grams is preserved across corpora, in the same domain (e.g. a corpus of chemistry articles). For, if this is not the case, different conclusions are certain to be drawn from different corpora.

The first test we performed was a direct comparison of the rank plots of the relative frequency of trigrams for the four corpora. We ranked 1039 MWE-candidate trigrams according to their occurrence in each corpus and we normalised this value by the total number of times any one of the 1039 trigrams appeared for each corpus. These normalisation values were: 66,101 times in BNC_f, 322,325 in BNC, 224,479,065 in Google and 6,081,786,313 in Yahoo. It is possible to have an estimate of the size of each corpus from these numbers: the trigrams account for something like 0.3% of the BNC corpora, while for Google and Yahoo nothing can be said since their sizes are not reliable numbers. Figure 1 displays the results. The overall ranking distribution is very similar for these corpora showing the expected Zipf like behaviour in spite of their different sizes.

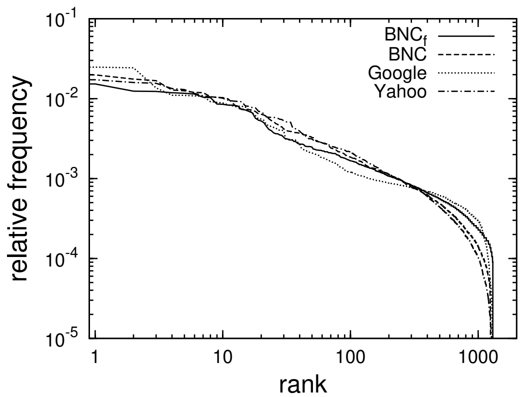

Figure 1: Relative frequency rank for the 1039 trigrams analysed.

Of course, the information coming from Figure 1 is not sufficient for our purposes. The order of the trigrams could be very different inside each corpus. Therefore a second test is needed to compare the rankings of the n-grams in each corpus. In order to do that we measure the Kendall's τ scores between corpora. Kendall's τ is a non-parametric method for estimating correlation between datasets (Press et al., 1992). For the number of trigrams studied here the Kendall's scores obtained imply a significant correlation between the corpora with p<0.000001. The significance indicates that the data are correlated and the *null* hypothesis of statistical independence is certainly disproved. Unfortunately disproving the *null* hypothesis does not give much information about the degree of correlation; it only asserts that it exists. Thus, it could be a very insignificant correlation. In table 1, we display a more intuitive measure to estimate the correlation, the probability Q that any 2 trigrams chosen from two corpora have the same relative ordering in frequency. This probability is related to Kendall's τ through the expression $Q = (1 + \tau)/2$.

	BNC	Google	Yahoo
BNC$_f$	0.81	0.73	0.78
BNC		0.73	0.77
Google			0.86

Table 1: The probability Q of 2 trigrams having the same frequency rank order for different corpora.

The results show that the four corpora are certainly correlated, and can probably be used interchangeably to access most of the statistical properties of the trigrams. Interestingly, a higher correlation was observed between Yahoo and Google than between BNC$_f$ and BNC, even though BNC$_f$ is a fragment of BNC, and therefore would be expected to have a very high correlation. This suggests that as corpora sizes increase, so do the correlations between them, meaning that they are more likely to agree on the ranking of a given MWE.

5.2 Comparing statistical measures - are they equivalent?

Here we concentrate on a single corpus, BNC$_f$, and compare the three statistical measures for MWE identification: Mutual Information (MI), χ^2 and Permutation Entropy (PE)(Zhang et al., 2006), to investigate if they order the trigrams in the same fashion.

MI and χ^2 are typical measures of association that compare the joint probability of occurrence of a certain group of events $p(abc)$ with a prediction derived from the *null* hypothesis of statistical independence between these events $p_\emptyset(abc) = p(a)p(b)p(c)$ (Press et al., 1992). In our case the events are the occurrences of words in a given position in an n-gram. For a trigram with words $w_1 w_2 w_3$, χ^2 is calculated as:

$$\chi^2 = \sum_{a,b,c} \frac{[\, n(abc) - n_\emptyset(abc)\,]^2}{n_\emptyset(abc)}$$

where a corresponds either to the word w_1 or to $\neg w_1$ (all but the word w_1) and so on. $n(abc)$ is the number of trigrams abc in the corpus, $n_\emptyset(abc) = n(a)n(b)n(c)/N^2$ is the predicted number from the *null* hypothesis, $n(a)$ is the number of unigrams a, and N the number of words in the corpus. Mutual Information, in terms of these numbers, is:

$$\mathrm{MI} = \sum_{a,b,c} \frac{n(abc)}{N} \log_2 \left[\frac{n(abc)}{n_\emptyset(abc)} \right]$$

The third measure, permutation entropy, is a measure of order association. Given the words w_1, w_2, and w_3, PE is calculated in this work as:

$$\mathrm{PE} = - \sum_{(i,j,k)} p(w_i w_j w_k) \ln [\, p(w_i w_j w_k) \,]$$

where the sum runs over all the permutations of the indexes and, therefore, over all possible positions of the selected words in the trigram. The probabilities are estimated from the number of occurrences of each permutation of a trigram (e.g. *by and large, large by and, and large by, and by large, large and by*, and *by large and*) as:

$$p(w_1 w_2 w_3) = \frac{n(w_1 w_2 w_3)}{\sum_{(i,j,k)} n(w_i w_j w_k)}$$

PE was proposed by Zhang et al. (2006) as a possible measure to detect MWEs, under the hypothesis that MWEs are more rigid to permutations and therefore present smaller PEs. Even though it is quite different from MI and χ^2, PE can also be thought as an indirect measure of statistical independence, since the more independent the words are the closer PE is from its maximal value ($\ln 6$, for trigrams). One possible advantage of this measure over the others is that it does not rely on single word counts, which are less accurate in Web based corpora.

Given the rankings produced for each one of these three measures we again use Kendall's τ test to assess correlation and its significance. Table 2 displays the Q probability of finding the same ordering in these three measures. The general conclusion from the table is that even though there is statistical significance in the correlations found (the p values are not displayed, but they are very low as before) the different measures order the trigrams very differently. There is a 70% chance of getting the same order from MI and χ^2, but it is safe to say that these measures are very different from the PE, since their Q values are very close to pure chance.

	MI$\times\chi^2$	MI$\times$PE	$\chi^2\times$PE
Q	0.71	0.55	0.45

Table 2: The probability Q of having 2 trigrams with the same rank order for different statistical measures.

5.3 Comparing Statistical Measures - are they useful?

The use of statistical measures is widespread in NLP but there is no consensus about how good these measures are for describing natural language phenomena. It is not clear what exactly they capture when analysing the data.

In order to evaluate if they would make good predictors for MWEs, we compare the measures distributions for MWEs and non-MWEs. For that we selected as gold standard a set of around 400 MWE candidates annotated by a native speaker[1] as MWEs or not. We then calculated the histograms for the values of MI, χ^2 and PE for the two groups. MI and χ^2 were calculated only for BNC_f. Table 3 displays the results of the Kolmogorov-Smirnof test (Press et al., 1992) for these histograms, where the first value is Kolmogorov-Smirnov D value ($D \in [0,1]$ and large D values indicate large differences between distributions) and the second is the significance probability (p) associated to D given the sizes of the data sets, in this case 90 for MWEs and 292 for non-MWEs.

	MI_{BNC_f}	$\chi^2_{BNC_f}$	PE_{Yahoo}	PE_{Google}
D	0.27	0.13	0.27	0.24
p<	0.0001	0.154	0.0001	0.0005

Table 3: Comparison of MI, χ^2 and PE

The surprising result is that there is no statistical significance, at least using the Kolmogorov-Smirnov test, that indicates that being or not an MWE has some effect in the value of the trigram's χ^2. The same does not happen for MI or PE. They do seem to differentiate between MWEs and non-MWEs. As discussed before the statistical significance implies the existence of an

[1]The native speaker is a linguist expert in MWEs.

effect but has very little to say about the intensity of the effect. As in the case of this work our interest is to use the effect to predict MWEs, the intensity is very important. In the figures that follow we show the normalised histograms for MI, χ^2(for the BNC_f) and PE (for the case of Yahoo) for MWEs and non-MWEs. The ideal scenario would be to have non overlapping distributions for the two cases, so a simple threshold operation would be enough to distinguish MWEs. This is not the case in any of the plots. Starting from Figure 3 it clearly illustrates the negative result for χ^2 in table 3. The other two distributions show a visible effect in the form of a slight displacement of the distributions to the left for MWEs. In particular for the distribution of PE, the large peak on the right, representing the n-grams whose word order is irrelevant with respect to its occurrence, has an important reduction for MWEs.

The statistical measures discussed here are all different forms of measuring correlations between the component words of MWEs. Therefore, as some types of MWEs may have stronger constraints on word order, we believe that more visible effects can be seen in these measures if we look at their application for individual types of MWEs, which is planned for future work. This will bring an improvement to the power of MWE prediction of these measures.

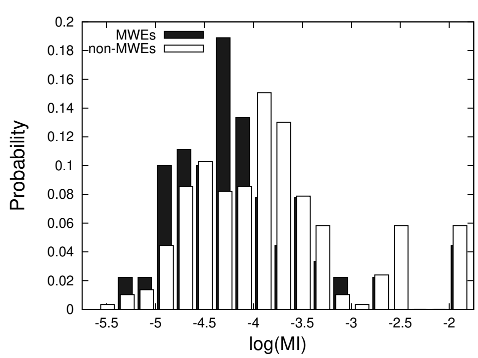

Figure 2: Normalised histograms of MI values for MWEs and non-MWEs in BNC_f.

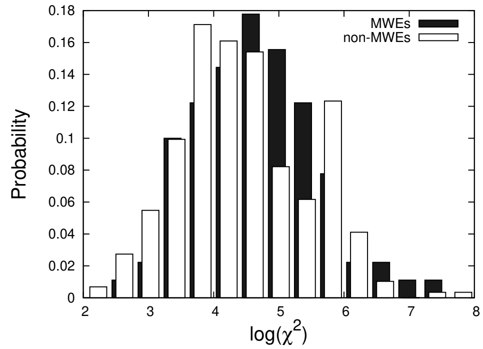

Figure 3: Normalised histograms of χ^2 values for MWEs and non-MWEs in BNC_f.

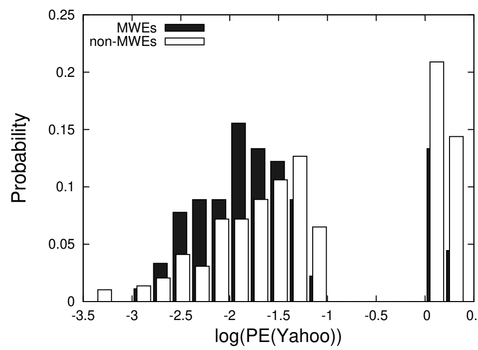

Figure 4: Normalised histograms of PE values for MWEs and non-MWEs in Yahoo.

6 Evaluation of the Extensions to the Grammar

Our ultimate goal is to maximally automate the process of discovering and handling MWEs. With good statistical measures, we are able to distinguish genuine MWE from non-MWEs among the n-gram candidates. However, from the perspective of grammar engineering, even with a good candidate list of MWEs, great effort is still required in order to incorporate such word units into a given grammar automatically and in a precise way.

Zhang et al. (2006) tried a simple "word with spaces" approach. By acquiring new lexical entries for the MWEs candidates validated by the statistical measures, the grammar coverage was shown to improve significantly. However, no further investigation on the parser accuracy was reported there.

Taking a closer look at the MWE candidates proposed, we find that only a small proportion of them can be handled appropriately by the "word with spaces" approach of Zhang et al. (2006). Simply adding new lexical entries for all MWEs can be a workaround for enhancing the parser coverage, but the quality of the parser output is clearly linguistically less interesting.

On the other hand, we also find that a large proportion of MWEs that cannot be correctly handled by the grammar can be covered properly in a constructional way by adding one lexical entry for the head (governing) word of the MWE. For example, the expression *foot the bill* will be correctly handled with a standard head-complement rule, if there is a transitive verb reading for the word *foot* in the lexicon. Some other examples are: *to* ***put*** *forward*, *the* ***good*** *of*, *in* ***combination*** *with*, ..., where lexical extension to the words in ***bold*** will allow the grammar to cover these MWEs. In this paper, we employ a constructional approach for the acquisition of new lexical entries for the head words of the MWEs.[2]

It is arguable that such an approach may lead to some potential grammar overgeneration, as there is no selectional restriction expressed in the new lexical entry. However, as far as the parsing task is concerned, such overgeneration is not likely to reduce the accuracy of the grammar significantly as we show later in this paper through a thorough evaluation.

6.1 Experimental Setup

With the complete list of 1039 MWE candidates discussed in section 5, we rank each n-gram according to each of the three statistical measures. The average of all the rankings is used as the combined measure of the MWE candidates. Since we are only interested in acquiring new lexical entries for MWEs which are not covered by the grammar, we used the error mining results (Zhang et al., 2006; van Noord, 2004) to only keep those candidates with parsability ≤ 0.1. The top 30 MWE candidates are used in this experiment.

We used simple heuristics in order to extract the head words from these MWEs:

- the n-grams are POS-tagged with an automatic tagger;
- finite verbs in the n-grams are extracted as head words;
- nouns are also extracted if there is no verb in the n-gram.

Occasionally, the tagger errors might introduce wrong head words. However, the lexical type predictor of Zhang and Kordoni (2006) that we used in our experiments did not generate interesting new entries for them in the subsequent steps, and they were thus discarded, as discussed below.

With the 30 MWE candidates, we extracted a sub-corpus from the BNC with 674 sentences which included at least one of these MWEs. The lexical acquisition technique described in Zhang and Kordoni (2006) was used with this sub-corpus in order to acquire new lexical entries for the head words. The lexical acquisition model was trained with the Redwoods treebank (Oepen et al., 2002), following Zhang et al. (2006).

The lexical prediction model predicted for each occurrence of the head words a most plausible lexical type in that context. Only those predictions that occurred 5 times or more were taken into consideration for the generation of the new lexical entries. As a result, we obtained 21 new lexical entries.

These new lexical entries were later merged into the ERG lexicon. To evaluate the grammar performance with and without these new lexical entries, we

1. parsed the sub-corpus with/without new lexical entries and compared the grammar coverage;
2. inspected the parser output manually and evaluated the grammar accuracy.

In parsing the sub-corpus, we used the PET parser (Callmeier, 2001). For the manual eval-

[2]The combination of the "word with space" approach of Zhang et al. (2006) with the constructional approach we propose here is an interesting topic that we want to investigate in future research.

uation of the parser output, we used the treebanking tools of the [incr tsdb()] system (Oepen, 2001).

6.2 Grammar Performance

Table 4 shows that the grammar coverage improved significantly (from 7.1% to 22.7%) with the acquired lexical entries for the head words of the MWEs. This improvement in coverage is largely comparable to the result reported in (Zhang et al., 2006), where the coverage was reported to raise from 5% to 18% with the "word with spaces" approach (see also section 4).

It is also worth mentioning that Zhang et al. (2006) added 373 new lexical entries for a total of 311 MWE candidates, with an average of 1.2 entries per MWE. In our experiment, we achieved a similar coverage improvement with only 21 new entries for 30 different MWE candidates, with an average of 0.7 entries per MWE. This suggests that the lexical entries acquired in our experiment are of much higher linguistic generality.

To evaluate the grammar accuracy, we manually checked the parser outputs for the sentences in the sub-corpus which received at least one analysis from the grammar before and after the lexical extension. Before the lexical extension, 48 sentences are parsed, among which 32 (66.7%) sentences contain at least one correct reading (table 4). After adding the 21 new lexical entries, 153 sentences are parsed, out of which 124 (81.0%) sentences contain at least one correct reading.

Baldwin et al. (2004) reported in an earlier study that for BNC data, about 83% of the sentences covered by the ERG have a correct parse. In our experiment, we observed a much lower accuracy on the sub-corpus of BNC which contains a lot of MWEs. However, after the lexical extension, the accuracy of the grammar recovers to the normal level.

It is also worth noticing that we did not receive a larger average number of analyses per sentence (table 4), as it was largely balanced by the significant increase of sentences covered by the new lexical entries. We also found that the disambiguation model as described by Toutanova et al. (2002) performed reasonably well, and the best analysis is ranked among top-5 for 66% of the cases, and top-10 for 75%.

All of these indicate that our approach of lexical acquisition for head words of MWEs achieves a significant improvement in grammar coverage without damaging the grammar accuracy. Optionally, the grammar developers can check the validity of the lexical entries before they are added into the lexicon. Nonetheless, even a semi-automatic procedure like this can largely reduce the manual work of grammar writers.

7 Conclusions

In this paper we looked at some of the issues involved in the evaluation of the identification of MWEs. In particular we evaluated the use of three statistical measures for automatically identifying MWEs. The results suggest that at least two of them (MI and PE) can distinguish MWEs. In terms of the corpora used, a surprisingly higher level of agreement was found between different corpora (Google and Yahoo) than between two fragments of the same one. This tells us two lessons. First that even though Google and Yahoo were not carefully built to be language corpora their sizes compensate for that making them fairly good samples of language usage. Second, a fraction of a smaller well balanced corpus may not necessarily be as balanced as the whole.

Furthermore, we argued that for precise grammar engineering it is important to perform a careful evaluation of the effects of including automatically acquired MWEs to a grammar. We looked at the evaluation of the effects in coverage, size of the grammar and accuracy of the parses after adding the MWE-candidates. We adopted a compositional approach to the encoding of MWEs, using some heuristics to detect the head of an MWE, and this resulted in a smaller grammar than that by Zhang et al. (2006), still achieving a similar increase in coverage and maintaining a high level of accuracy of parses, comparable to that reported by Baldwin et al. (2004).

The statistical measures are currently only

	item #	parsed #	avg. analysis #	coverage %
ERG	674	48	335.08	7.1%
ERG + MWE	674	153	285.01	22.7%

Table 4: ERG coverage with/without lexical acquisition for the head words of MWEs

used in a preprocessing step to filter the non-MWEs for the lexical type predictor. Alternatively, the statistical outcomes can be incorporated more tightly, i.e. to combine with the lexical type predictor and give confidence scores on the resulting lexical entries. These possibilities will be explored in future work.

References

Timothy Baldwin and Aline Villavicencio. 2002. Extracting the unextractable: A case study on verb-particles. In *Proc. of the 6th Conference on Natural Language Learning (CoNLL-2002)*, Taipei, Taiwan.

Timothy Baldwin, Emily M. Bender, Dan Flickinger, Ara Kim, and Stephan Oepen. 2004. Road-testing the English Resource Grammar over the British National Corpus. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal.

Gosse Bouma, Gertjan van Noord, and Robert Malouf. 2001. Alpino: Wide-coverage computational analysis of dutch. In *Computational Linguistics in The Netherlands 2000*.

Ted Briscoe and John Carroll. 1997. Automatic extraction of subcategorization from corpora. In *Fifth Conference on Applied Natural Language Processing*, Washington, USA.

Ulrich Callmeier. 2001. Efficient parsing with large-scale unification grammars. Master's thesis, Universität des Saarlandes, Saarbrücken, Germany.

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

Dan Flickinger. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering*, 6(1):15–28.

Ray Jackendoff. 1997. Twistin' the night away. *Language*, 73:534–59.

Jeremy Nicholson and Timothy Baldwin. 2006. Interpretation of compound nominalisations using corpus and web statistics. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 54–61, Sydney, Australia. Association for Computational Linguistics.

Stephan Oepen, Kristina Toutanova, Stuart Shieber, Christopher Manning, Dan Flickinger, and Thorsten Brants. 2002. The LinGO Redwoods treebank: Motivation and preliminary applications. In *Proceedings of COLING 2002: The 17th International Conference on Computational Linguistics: Project Notes*, Taipei.

Stephan Oepen. 2001. [incr tsdb()] — competence and performance laboratory. User manual. Technical report, Computational Linguistics, Saarland University, Saarbrücken, Germany.

Darren Pearce. 2002. A comparative evaluation of collocation extraction techniques. In *Third International Conference on Language Resources and Evaluation*, Las Palmas, Canary Islands, Spain.

William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. 1992. *Numerical Recipes in C: The Art of Scientific Computing. Second edition.* Cambridge University Press.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15, Mexico City, Mexico.

Kristina Toutanova, Christoper D. Manning, Stuart M. Shieber, Dan Flickinger, and Stephan Oepen. 2002. Parse ranking for a rich HPSG grammar. In *Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT2002)*, pages 253–263, Sozopol, Bulgaria.

Leonoor van der Beek. 2005. The extraction of determinerless pps. In *Proceedings of the ACL-SIGSEM Workshop on The Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, Colchester, UK.

Gertjan van Noord. 2004. Error mining for wide-coverage grammar engineering. In *Proceedings of* *the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 446–453, Barcelona, Spain, July.

Aline Villavicencio, Francis Bond, Anna Korhonen, and Diana McCarthy. 2005. Introduction to the special issue on multiword expressions: having a crack at a hard nut. *Journal of Computer Speech and Language Processing*, 19(4):365–377.

Aline Villavicencio. 2005. The availability of verb-particle constructions in lexical resources: How much is enough? *Journal of Computer Speech and Language Processing*, 19.

Yi Zhang and Valia Kordoni. 2006. Automated deep lexical acquisition for robust open texts processing. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy.

Yi Zhang, Valia Kordoni, Aline Villavicencio, and Marco Idiart. 2006. Automated multiword expression prediction for grammar engineering. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 36–44, Sydney, Australia. Association for Computational Linguistics.