Detecting a Continuum of Compositionality in Phrasal Verbs.

Diana McCarthy & Bill Keller & John Carroll
Cognitive & Computing Sciences,
University of Sussex
Brighton BN1 9QH, UK
dianam,billk,johnca@cogs.susx.ac.uk

Abstract

We investigate the use of an automatically acquired thesaurus for measures designed to indicate the compositionality of candidate multiword verbs, specifically English phrasal verbs identified automatically using a robust parser. We examine various measures using the nearest neighbours of the phrasal verb, and in some cases the neighbours of the simplex counterpart and show that some of these correlate significantly with human rankings of compositionality on the test set. We also show that whilst the compositionality judgements correlate with some statistics commonly used for extracting multiwords, the relationship is not as strong as that using the automatically constructed thesaurus.

1 Introduction

Many people are working on acquisition of multiword expressions, although terminology varies. In this paper, we are interested in lexicalised expressions (Sag et al., 2002) where special interpretation is required because of some degree of non-compositionality or semantic opacity. We are specifically concerned with what are commonly referred to as phrasal verbs, or verb and particle constructions (Baldwin and Villavicencio, 2002). As well as having idiosyncratic semantics, phrasals also display specific syntactic behaviour such as permitting particle movement when used in the transitive; for example:
Jo ate up her food ↔ *Jo ate her food up*

We are interested in phrasal verbs because we want to acquire predicate selectional preferences for word sense disambiguation (McCarthy et al., 2001). When acquiring such lexical information for a verb it is important to know when there is a special interpretation required for the verb and particle combination so that these combinations are handled separately from the simplex case. Whilst it is possible to put every single occurrence of a verb and particle combination into a lexicon this is not desirable. One wants to achieve generalisation and avoid redundancy, only storing details which cannot be created from what is already there. Not every verb modified by a particle may be a genuine multiword unit, but may instead be a fully compositional verb modified by an adverbial e.g. *fly up*. Also very productive verb particle combinations such as those involving verbs of motion, which often occur with a particle e.g. *up*, such as *wander*, *stroll*, *go* etc... might be better handled in the grammar (Villavicencio and Copestake, 2002).

Additionally, in lexical acquisition, and for word sense disambiguation, it is important that related senses of words are identified. For example, if the verb *eat* is closer in meaning to a phrasal construction *eat up*, compared to other simplex verbs with their phrasal constructions such as *blow/blow up*, then the lexicon should reflect that. Having a measure of compositionality should help in this.

In this paper we are not concerned with evaluation of precision and recall of the extraction of phrasal verbs from a parser, although we have done some preliminary experiments in this direction on the Wall Street Journal (WSJ), see section 3. Instead, our focus is on methods of using an automatically acquired thesaurus for detecting compositionality of candidate phrasals output from our parser. We contrast this with some statistics commonly used for multiword extraction. The thesaurus is acquired from the grammatical relations occurring with verbs, both the target phrasals and their simplex counterparts. The intuition is that the neighbours of the simplex verb should be similar to those of the phrasal where the phrasal has a compositional meaning, and that the phrasal neighbours should include phrasal candidates with the same particle.

For evaluation, we obtain a sample of multiword candidates from our parser and then obtain human judgements of compositionality using an ordinal scale for compositionality. We demonstrate that there is highly significant agreement on the rank order of these judgements and use the average ranks for each item as a gold-standard to compare various measures aimed at detecting non-compositionality.

In the following section we look at related work. In section 3 we show how phrasals are identified by our parser. We talk about the generation of the gold-standard set of compositionality judgements in section 4. In section 5 we describe the construction of the automatic thesaurus and the measurements we explored for detecting compositionality. In section 6 we show the correlations of our measures with the gold-standard, and compare these to some statistics commonly used for identifying compositional multiwords. In section 7 we analyse our findings, and conclude (section 8) with directions for future work.

2 Related Research

There has been a lot of recent work on extraction of multiwords from corpora we focus specifically on work involving multiword verbs, and detecting compositionality of multiwords.

2.1 Multiword Verb Extraction

There have been a number of methods proposed in the literature for extracting multiword verb constructions from corpora. Baldwin and Villavicencio (2002) demonstrated that combining syntactic evidence using automatic PoS taggers and statistical chunkers, and feeding evidence from a number of tokens into a memory based-learner gave high precision and recall, using marked up WSJ text to gauge precision, and phrasals listed in the Alvey Natural Language Tools (ANLT) (Grover et al., 1993) attested in the same corpus for recall. No distinction on opaqueness of the verb and particle constructions was made.

Blaheta and Johnson (2001) used log-linear models to extract English multiword verbs involving verb and particle constructions from parsed data; these include phrasal and prepositional verbs. [1]

[1] Prepositional verbs also have some degree of idiosyncratic semantic interpretation, but the particle functions as a preposition and selects for the following noun phrase. There is therefore no particle movement e.g. **she referred the problem to.*

Krenn and Evert (2001) investigated German support verb constructions (identifiable on grammatical grounds) and figurative expressions (having idiomatic interpretations). In their experiments, true positives were typically defined as such according to the annotator scanning the list. Krenn and Evert found that different statistics are suited to different types of collocation - there is no easy route for collocation extraction. Moreover, they found that a simple co-occurrence frequency fares comparably, if not better, than most statistical tests of significance.

2.2 Compositionality of Multiwords

Most people researching into multiwords assume some degree of non-compositionality. Blaheta and Johnson took human judgements on phrasality, opaqueness (a dichotomous scale) and a subjective judgement of relatedness (on a scale between 1 and 5). They showed that the opaqueness judgements correlated with the relatedness (good collocation) judgement. Also, those constructions judged to be phrasals tended to have higher ranks (higher opaqueness and relatedness) than prepositional verb particle constructions.

Both Lin (1999) and Schone and Jurafsky (2001) have used distributional similarity to detect compositionality in multiwords. Schone and Jurafsky used measures on the vectors representing the multiword candidates compared to measures for the words that the multiword contains but this failed to improve performance, using WordNet and other machine readable resources as gold-standards for evaluation. There was some success though in using latent semantic analysis (LSA) models to identify multiwords by the fact that the component words are typically non-substitutable, but they felt that much of what is captured by this is already handled by the statistics that they employ.

Lin (1999) had already done something similar to the substitutability experiments using the method he had proposed earlier (Lin, 1998a) for automatic thesaurus construction. He identified general multiwords involving several open-class words output from his parser and filtered by the log-likelihood statistic. Using the parser yielded much better results than just a simple window for co-occurrence relationships. Lin proposed that if there is a multiword obtained by substitution of either the head or modifier in the multiword with a near neighbour, then the mutual information of this and the original multiword must be significantly different for the original multiword to be considered non-compositional. He evaluated this manually on a sample. As well as finding non-compositional multiwords, there were also a higher proportion of parser errors that met these criteria.

Bannard et al. (2003) are investigating compositionality by looking at the contribution of the verb, and the particle to the semantics of the verb and particle combination; this follows on from Bannard's earlier work (2002) where he showed that compositionality judgements correlate with human judgements of similarity between the head verb and the verb and particle combination. Bannard et al. (2003) point out that Lin's method of using substitution of component words in a multiword with semantic neighbours is a good indication of productivity, but not necessarily of compositionality, since an institutionalised non-productive combination, such as *frying pan* would not have near neighbour substitutes, but would nevertheless be compositional. They explore four methods for detecting compositionality using resources acquired from distributional data. They use these on 40 candidates on 4 separate tasks which aim to determine whether i) the item is compositional, ii) one component word contributes its meaning iii) the verb contributes its meaning iv) the particle contributes its meaning. The classifications on each of these tasks according to these methods are contrasted with a gold standard classification from 26 judges on the same data. The methods exceed the mean agreement of the annotators in some cases, particularly as regards the contribution from the particle.

Baldwin et al. (2003) are also exploring empirical models of compositionality using LSA with noun-noun compounds and verb-particle constructions. In their study, they compare the similarities of the component words with WordNet based similarity scores and demonstrate a moderate correlation, lower for noun-noun compounds.

We are also exploring the relation between a verb and verb and particle combination (we use the term *phrasal verb*) using distributional techniques, but our evaluation is somewhat different.

2.3 Evaluation

Evaluation of collocation extraction is a notoriously thorny problem (Krenn and Evert, 2001; Pearce, 2002). People do use MRDs such as WordNet (Schone and Jurafsky, 2001) even though they acknowledge that there will be omissions in these resources, and the phenomena in the resource may be rare or simply not attested in the particular corpus used for acquisition. Many researchers use manually annotated samples, where the judges make a binary decision on whether each candidate multiword is "genuine" or not (Lin, 1999; Blaheta and Johnson, 2001; Krenn and Evert, 2001; Baldwin and Villavicencio, 2002). As Krenn and Evert point out, there is low agreement between annotators who are asked to mark "typical" multiwords, or collocations. The intuitions behind what is typical vary, and likewise association scores vary in their ability to partition the set depending on the notion of "typicality" employed by the annotators. Researchers also sometimes show how well the results accord with the contents of MRDs, even though these cannot be taken as definitive.

In this study we are less interested in the dichotomy of whether a putative phrasal candidate is indeed a genuine multiword or not (although it is more likely that those with low compositionality are likely to be) but we use empirical methods to gauge the position of a candidate on a continuum between the fully opaque idiom and transparent compositional phrases. Variability of idioms on a scale of compositionality has been discussed by Nunberg et al. (1994) and in the psycholinguistics literature, see (Gibbs and Nayak, 1989). Tseng (2000) also advocates use of a spectrum when considering the semantics of prepositions. We will consider compositionality as a continuous scale and ask human judges to rank multiword candidates along this. We investigate the use of these ranked judgements for evaluating compositionality measures. We also look at the relation between these judgements and appearance of the candidates in gold-standard resources such as WordNet (Miller et al., 1993) or the ANLT lexicon (Grover et al., 1993), on the premise that non-compositional phrases are more likely to be listed as multiwords in man-made resources.

3 Parser Output

For these experiments we use data from the ninety million words of the written portion of the British National Corpus parsed with the RASP parser (Briscoe and Carroll, 2002). The output of the parser is a set of *grammatical relations* (Carroll et al., 1998) specifying the syntactic dependency between each head and its dependent(s), read off from the phrase structure tree that is returned from the disambiguation phase. The parser uses information from ANLT such as phrasals in its dictionary. This makes it more likely to spot phrasal constructions from this list. We have already looked at recognition of verb and particle constructions in the WSJ identified purely on syntactic grounds using the parses provided with the WSJ Penn Treebank 2 (Marcus et al., 1995) as a gold standard. The results for identifying verb and particle tokens are reported in table 1, both with and without the ANLT phrasal list (ANLT

	Precision	Recall
MINIPAR	78.9	44.1
RASP (without ANLT phr)	87.6	49.4
RASP (with ANLT phr)	92.6	64.2

Table 1: Identification of verb and particle attachments in WSJ data

phr). We also give results for comparison obtained on the same data for another wide coverage parser, (MINIPAR (Lin, 1998b)). [2]

In the RASP parser grammatical relation output we identify phrasal verbs as being a verb modified by a particle (tagged RP) under the ncmod (non-clausal modifier) relation. It is quite possible that some particle tags have been given erroneously and that some genuine particles are not recognised as such by the parser, or are not attached to the verb by the parser. We only look at tokens in isolation and therefore do not collate evidence to look for syntactic evidence of particle movement as Baldwin and Villavicencio do. This would be a good way to improve phrasal extraction accuracy, particularly where a particle follows a pronoun.

4 Human Compositionality Judgements

In our experiments we asked human judges to rank phrasal verb candidates as to how compositional they are.

4.1 Test set

From the full set of 4272 phrasal verb candidate types output from the RASP parser we obtained 100 candidates randomly subject to the constraint that 33 [3] each came from one of 3 frequency ranges (each range covering an even number of phrasal types) from 20 to the maximum frequency. A further 16 manually selected phrasals were added to this test set.

Three native English speakers ranked the 116 candidates on a numerical score 0 to 10 (10 for fully-compositional, 0 for totally opaque), or gave a "don't know" response. We discounted any item where any of the judges had put such a "don't know". This only removed a total of 5 items, leaving a ranking from all 3 judges on 111 candidates.

4.2 Human Agreement

To investigate if the rankings from the 3 judges agreed we employed the Kendall Coefficient of Concordance (W) (Siegel and Castellan, 1988). This statistic is useful for determining inter-rater agreement where there are 3 or more judges and the judgements are ordinal, and one is interested in the ranks rather than the actual numerical values. W ranges between 0 (little agreement) and 1 (full agreement) and bears a linear relationship to the average Spearman Rank-order Correlation Coefficient taken over all possible pairs of the rankings.

W is calculated as shown in equation 1 below, where n is the number of items (111 in this case), $\bar{R}_i$ is the average rank for the i^{th} item and k is the number of raters.

$$W = \frac{12 \sum_{i=0}^{n} \bar{R}_i^2 - 3n(n+1)^2}{n(n^2-1) - \frac{\sum_{j=1}^{k} T_j}{k}} \quad (1)$$

The second term in the denominator includes a correction for ties where:

$$T_j = \sum_{i=1}^{g_j} (t_i^3 - t_i) \quad (2)$$

where t_i is the number of tied ranks in the i^{th} grouping of ranks.

The value $k(n-1)W$ is approximately distributed as χ^2 with $n-1$ degrees of freedom. We obtained a W score of 0.594 which gives a χ^2 score of 196.30 for 110 degrees of freedom which is highly significant (probability of this value $<= 0.000001$).

5 Detecting Compositionality

5.1 Using an Automatically Constructed Thesaurus:

Using the method proposed by Lin (1998a) we produced a thesaurus with 500 nearest neighbours for the set of phrasal verbs as described above. Tuples of the form <*verb, argument head, grammatical relation*> from the parsed BNC data were used for this purpose where the verb was the multiword phrasal and the grammatical relations used were subjects and direct objects. We did likewise for the simplex verbs contained within the phrasals (e.g. *blow* from *blow up*).

We investigated various measures which compare the nearest neighbours of the phrasal verb to the neighbours of the corresponding simplex verb. We also tried various measures on the neighbours of the phrasal verb. We supply short labels for these for ease of reference.

overlap The size of the overlap of the top X phrasal neighbours with the same number of the corresponding simplex verb's neighbours, not including the simplex verb itself. We tried this for

[2] We are indebted to Mirella Lapata for the results using MINIPAR.

[3] This was 34 from the lowest frequency range.

$X = 30$, 50, 100, and 500 [4]. Thus for example, the overlap of 50 nearest neighbours for *climb down* with those of *climb* is shown in figure 1. The intuition is that the more compositional the phrasal, the closer will be the neighbours of the phrasal and the corresponding simplex verb.

sameparticle The number of neighbours of the phrasal with the same particle as the phrasal. The intuition behind this is that the particle contributes to the semantics in the compositional case.

sameparticle-simplex The number of neighbours with the same particle as the phrasal (in the top 500), minus the equivalent number of the simplex neighbours (i.e. having the same particle as the target phrasal). Thus we control for the case that this particle appears to the same extent in the simplex neighbours.

simplexasneighbour Whether the simplex verb occurs in the top 50 nearest neighbours of the phrasal.

rankofsimplex The rank of the simplex in the top 500 nearest neighbours of the phrasal.

scoreofsimplex The similarity score of the simplex in the top 500 nearest neighbours of the phrasal.

overlapS The overlap of neighbours (in the top 30, top 50 and top 500 neighbours) where we used the simplex form of phrasals in the phrasal neighbours; so for example this variation for the overlap for the top 50 neighbours of *climb down* is shown in figure 2. The intuition here is that the particle includes some semantics in the compositional case, and removing particles from the neighbours goes some way to reducing the sparse data problem and removing the semantics of the particles.

5.2 Statistics Used for Comparison

In our experiments we compared the results using the nearest neighbours to various statistics commonly used for multiword extraction. We used the χ^2 statistic, the log-likelihood ratio statistic (LLR) (Dunning, 1993) and point-wise mutual information (MI) (Church and Hanks, 1990) and looked at the correlation of these statistics with the compositionality judgements. We also looked at the frequency of the phrasal (i.e. the co-occurrence frequency of the verb when modified by the particle), and the frequency of the simplex verb.

[4] We do not do this for overlaps of 10 or 20 neighbours because there was no overlap for these sizes

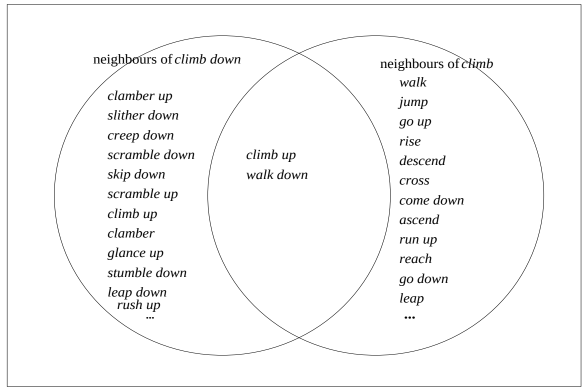

Figure 1: Overlap of top 50 neighbours of *climb down* with those of *climb*.

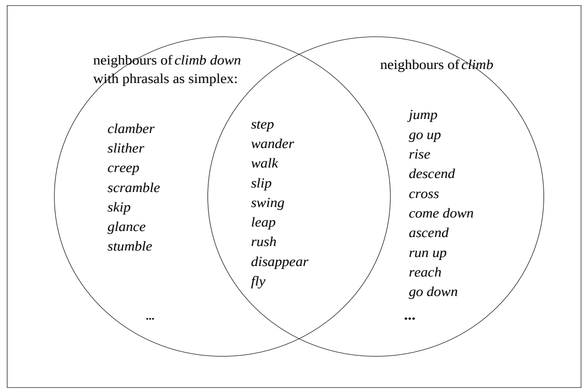

Figure 2: Overlap of top 50 neighbours, with phrasals reduced to simplex form.

5.3 Resources Used for Comparison

We compared the average rank of the human judgements to whether the phrasal was in WordNet [5] and also whether the phrasal was in ANLT. [6] We are not evaluating precision of our parser, we simply want to see how well our average ranks as gold-standard compositionality judgements correlated with these resources.

6 Results

Table 2 provides the values of correlation statistics on the various measures of compositionality against our gold-standard from the human judges. For the mea-

[5] Though in WordNet of course the type of multiword verb is not marked.

[6] Although our parser has been seeded with ANLT phrasal data, not all phrasals output from the parser are in ANLT. Of the 4272 phrasal types extracted, only 1531 (36%) are in the ANLT lexicon.

Correlation with Measures Using the Thesaurus			
measure	correlation statistic	Z score	probability under H_o
overlap PN SN 500	r_s = - 0.032	-0.38	0.35
overlap PN SN 100	$r_s = 0.037$	0.39	0.35
overlap PN SN 50	$r_s = 0.136$	1.43	0.08
overlap PN SN 30	$r_s = 0.166$	1.74	0.04
sameparticle PN 500	$r_s = 0.414$	4.34	<0.00003
sameparticle-simplex PN SN 500	$r_s = 0.49$	5.17	<0.00003
simplexasneighbour PN 500	Mann W	0.950	0.171
simplexrank PN 500	r_s = -0.115	-1.21	0.113
simplexscore PN 500	$r_s = 0.052$	0.54	0.295
overlapS PN SN 30	$r_s = 0.306$	3.21	<0.0007
overlapS PN SN 50	$r_s = 0.303$	3.18	<0.0007
overlapS PN SN 500	$r_s = 0.167$	1.75	0.040
Correlation with Man-made Resources			
WordNet	Mann W	2.39	0.008
ANLT Phrasals	Mann W	3.03	0.012
ANLT Prepositionals	Mann W	0.430	0.334
Correlation with Statistics (used for multiword extraction)			
χ^2	r_s = -0.213	-2.22	0.0139
LLR	r_s = -0.168	-1.76	0.0392
MI	r_s = -0.248	-2.60	0.0047
phrasal Freq	r_s = -0.096	-1.01	0.156
simplex Freq	$r_s = 0.092$	0.96	0.169

Table 2: Correlation with human compositionality judgements

sures which use the automatic thesaurus we indicate whether the measure relies only on the phrasal neighbours (PN), or the simplex neighbours (SN) or some combination of both (PN SN). In this first column, we also indicate how many of the top ranked neighbours were used. Where we are evaluating scores on a numerical scale, such as the size of the overlap, we use the ranks of the numerical values and compare these to the average ranks of our gold-standard using the Spearman Rank-Order Correlation Coefficient (r_s). Since we have a large enough sample, these can be used to obtain a normally distributed Z score and we can thus obtain the probability of obtaining a score such as this by chance under the null hypothesis (that there is no relationship). For the scores which involve a binary decision, such as whether a score is in WordNet or not, we use the Mann Whitney U test, which compares the gold-standard ranks for the partitioned set and gives a Z score. We use one-tailed tests because we predict the direction of the relationship. For all the scores using the automatic thesaurus, we assume that the larger the value, the more compositional the item.

For the statistics (commonly used for multiword extraction) the relationship is in the other direction: high values are indicative of a non-compositional reading. We change the log-likelihood statistic to add a sign where the joint frequency of particle and verb is smaller than anticipated from that expected.

From these results we can see that some of the measures from the automatic thesaurus correlate significantly with the human compositionality judgements and that these correlations are slightly stronger than those of any of the statistics used. The statistics used all correlate (in the other direction) with the human compositionality judgements, although this is slightly less so for the log-likelihood ratio. The frequency of the verb and particle seems to bear no significant relation to compositionality judgements. This is interesting because Krenn and Evert found that co-occurrence frequency was a good indication of the German multiwords, although the task there was identification of the multiwords, as opposed to measuring compositionality.

7 Analysis

MI is the statistic with the strongest value of r_s and the thesaurus measure with the strongest relationship was **sameparticle-simplex**. These two measures correlated well together (r_s = -0.51, z = -5.37) and both are significantly correlated (using the Mann Whitney U test) with whether the candidate is found in either WordNet or ANLT, see table 3, although the relationship using the automatic thesaurus is slightly higher.

Lin uses a log-likelihood ratio to filter multiword candidates before using his automatic thesaurus to detect compositionality in multiwords containing 2 or more open class words. For phrasal candidates at least, it might be worth using evidence from the thesaurus on the unfiltered list.

We were surprised, and a little disappointed that the straight overlap of neighbours did not give a significant relationship, other than for the overlap of 30 neighbours. We believe this is due to the large scope for open class words as neighbours, and that there is often some element of meaning added by the particle. Thus the overlap where we reduce neighbours of the phrasal to simplex form compensated for this.

We have not yet explored varying the number of neighbours for methods other than the **overlap** and **overlapS**. We feel that it would be worth exploring the effect of the number of neighbours further, and also to use the similarity scores of the neighbours, rather than simple measures operating on the types occurring as neighbours. This would help control for the fact that for some verbs there are not many close neighbours and neighbours further down the ranked list may in fact be quite distant.

Whilst statistics are useful indicators of non-compositionality, there are compositional multiwords which have low values for these statistics, yet are highly non-compositional. A good example is *cock up*; it is the lowest ranked for compostionality by the human judges, but its MI value is only 5.02, and according to MI it is ranked between the somewhat more compositional candidates *tie down* and *come down*. The automatic thesaurus measures such as **sameparticle-simplex** give a low compositionality score and place it at the end between *carry out* and *latch on*.

There are also candidates with high values of the statistics, yet they are in the middle range of the compositionality judgements, for example, *plod on*. This is simply because of a high co-occurrence frequency. Whether such an unexpectedly high co-occurrence frequency warrants an entry in the lexicon depends on the type of lexicon being built.

8 Conclusions

We can see that there is a significant relationship between the human compositionality judgements and some of the measures from the automatic thesaurus, particularly those that endeavour to take into account the semantics of the particle. This relationship is stronger than statistics which have previously been used for filtering candidate multiwords which suggests that it might be better not to filter with statistics before looking at compositionality using an automatic thesaurus.

Measure	in WordNet	in ANLT
MI	-2.61	-4.53
sameparticle-simplex	3.71	4.59

Table 3: Mann Whitney Z scores showing correlation of measures with man-made resources

We have not yet exploited these measures in the construction of a lexicon for phrasal verbs. Identifying non-compositional phrasals by employing thresholds to force a binary decision is one option. This would help in determining which candidate phrasals should be treated separately from the simplex for purposes such as selectional preference acquisition and word sense disambiguation. The thresholds might be acquired empirically from some training data, such as the compositionality judgements we have used. However, we believe that permitting measurements and evaluation on a continuum of compositionality allows for a more natural exploration of relationships, without imposing an arbitrary cut-off point required only when finally categorising items for a lexicon. It also could be useful to use the measures to tell whether the meaning comes from the verb or the particle or both, as Bannard et al. (2003) do, because if the verb contributes its meaning then data for selectional preference acquisition might be amalgamated with those of the simplex counterpart.

9 Acknowledgements

This work was supported by the EPSRC-funded RASP project (grant GR/N36493), and the EU 5th Framework project MEANING – Developing Multilingual Web-scale Language Technologies (IST-2001-34460). We are grateful to Timothy Baldwin and Colin Bannard for their helpful comments and useful references.

References

T. Baldwin and A. Villavicencio. 2002. Extracting the unextractable: A case study on verb-particles. In *Proceedings of the Sixth Conference on Computational Natural Language Learning (CoNLL 2002)*, pages 98–104, Taipei, Taiwan.

T. Baldwin, C. Bannard, T. Tanaka, and D. Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL Workshop on multiword expressions: analysis, acquisition and treatment.*

C. Bannard, T. Baldwin, and A. Lascarides. 2003. A statistical approach to the semantics of verb-particles. In *Proceedings of the ACL Workshop on multiword expressions: analysis, acquisition and treatment.*

C. Bannard. 2002. Statistical techniques for automatically inferring the semantics of verb-particle constructions. Technical Report WP-2002-06, University of Edinburgh, School of Informatics. http://lingo.stanford.edu/pubs/WP-2002-06.pdf.

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

E. Briscoe and J. Carroll. 2002. Robust accurate statistical annotation of general text. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*, pages 1499–1504, Las Palmas, Canary Islands, Spain.

J. Carroll, E. Briscoe, and A. Sanfilippo. 1998. Parser evaluation: a survey and a new proposal. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 447–454.

K.W. Church and P. Hanks. 1990. Word association norms, mutual information and lexicography. *Computational Linguistics*, 19(2):263–312.

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

R.W. Gibbs and N. P. Nayak. 1989. Psycholinguistic studies on the syntactic behaviour of idioms. *Cognitive Psychology*, 21:100–38.

C. Grover, J. Carroll, and T. Briscoe. 1993. The Alvey Natural Language Tools grammar. Technical Report 284, Computer Laboratory, University of Cambridge.

B. Krenn and S. Evert. 2001. Can we do better than frequency? A case study on extracting PP-verb collocations. In *Proceedings of the ACL Workshop on Collocations*, pages 39–46, Toulouse, France.

D. Lin. 1998a. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL 98*, Montreal, Canada.

D. Lin. 1998b. Dependency-based evaluation of MINIPAR at LREC. In *Proceedings of the Workshop on the Evaluation of Parsing Systems*, Granada, Spain. http://www.cs.ualberta.ca/l̃indek/minipar.htm.

D. Lin. 1999. Automatic identification of non-compositional phrases. In *Proceedings of ACL-99*, pages 317–324, Univeristy of Maryland, College Park, Maryland.

M. Marcus, G. Kim, M. A. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger. 1995. The Penn Treebank: annotating predicate argument structure. Technical report, University of Pennsylvania. Distributed on The Penn Treebank 2 CD-ROM by the Linguistic Data Consortium.

D. McCarthy, J. Carroll, and J. Preiss. 2001. Disambiguating noun and verb senses using automatically acquired selectional preferences. In *Proceedings of the* SENSEVAL-2 *workshop*, pages 119–122.

G. Miller, R. Beckwith, C. Fellbaum, D Gross, and K. Miller, 1993. *Introduction to WordNet: an On-Line Lexical Database.* ftp://clarity.princeton.edu/pub/WordNet/5papers.ps.

G. Nunberg, I. A. Sag, and T. Wasow. 1994. Idioms. *Language*, 70:491–538.

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

I. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, pages 1–15, Mexico City, Mexico.

P. Schone and D. Jurafsky. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 100–108, Hong Kong.

S. Siegel and N. John Castellan, editors. 1988. *Non-Parametric Statistics for the Behavioral Sciences.* McGraw-Hill, New York.

J. L. Tseng. 2000. *The Representation and Selection of Prepositions.* Ph.D. thesis, University of Edinburgh.

A. Villavicencio and A. Copestake. 2002. Verb-particle constructions in a computational grammar. In *Proceedings of the 9th International Conference on Head-Driven Phrase Structure Grammar (HPSG-2002)*, Seoul, South Korea.