A Statistical Approach to the Semantics of Verb-Particles

Colin Bannard
School of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW, UK
c.j.bannard@ed.ac.uk

Timothy Baldwin
CSLI
Stanford University
210 Panama Street
Stanford CA 94305, USA
tbaldwin@csli.stanford.edu

Alex Lascarides
School of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW, UK
alex@inf.ed.ac.uk

Abstract

This paper describes a distributional approach to the semantics of verb-particle constructions (e.g. *put up*, *make off*). We report first on a framework for implementing and evaluating such models. We then go on to report on the implementation of some techniques for using statistical models acquired from corpus data to infer the meaning of verb-particle constructions.

1 Introduction

The semantic representation of multiword expressions (MWEs) has recently become the target of renewed attention, notably in the area of hand-written grammar development (Sag et al., 2002; Villavicencio and Copestake, 2002). Such items cause considerable problems for any semantically-grounded NLP application (including applications where semantic information is implicit, such as information retrieval) because their meaning is often not simply a function of the meaning of the constituent parts. However, corpus-based or empirical NLP has shown limited interest in the problem. While there has been some work on statistical approaches to the semantics of compositional compound nominals (e.g. Lauer (1995), Barker and Szpakowicz (1998), Rosario and Hearst (2001)), the more idiosyncratic items have been largely ignored beyond attempts at identification (Melamed, 1997; Lin, 1999; Schone and Jurafsky, 2001). And yet the identification of non-compositional phrases, while valuable in itself, would by no means be the end of the matter. The unique challenge posed by MWEs for empirical NLP is precisely that they do not fall cleanly into the binary classes of compositional and non-compositional expressions, but populate a continuum between the two extremes.

Part of the reason for the lack of interest by computational linguists in the semantics of MWEs is that there is no established gold standard data from which to construct or evaluate models. Evaluation to date has tended to be fairly ad hoc. Another key problem is the lack of any firm empirical foundations for the notion of compositionality. Given this background, this paper has two aims. The first is to put the treatment of non-compositionality in corpus-based NLP on a firm empirical footing. As such it describes the development of a resource for implementing and evaluating statistical models of MWE meaning, based on non-expert human judgements. The second is to demonstrate the usefulness of such approaches by implementing and evaluating a handful of approaches.

The remainder of this paper is structured as follows. We outline the linguistic foundations of this research in Section 2 before describing the process of resource building in Section 3. Section 4 summarises previous work on the subject and Section 5 details our proposed models of compositionality. Section 6 lays out the evaluation of those models over the gold standard data, and we conclude the paper in Section 7.

2 Verb Particle Constructions

We selected the English verb-particle construction as our test case MWE in this paper. Verb-particle constructions (hereafter referred to as VPCs) consist of a head verb and one or more obligatory particles, in the form of intransitive prepositions (e.g. *hand in*), adjectives (e.g. *cut short*) or verbs (e.g. *let go*) . Here, we focus exclusively on prepositional particles due to their high productivity and variable compositionality. Examples of prepositional VPCs are *put up*, *finish up*, *gun down* and *make out* as used in the following sentences:

(1) Peter put the picture up

(2) Susan finished up her paper

(3) Philip gunned down the intruder

(4) Barbara and Simon made out

VPCs cause significant problems for NLP systems. Semantically, they often cannot be understood through the simple composition of their independent parts. Compare, for example, sentences (1) and (4). In (1), the meaning seems to be that Peter *put* the picture somewhere and that as a consequence the picture was *up*. That is, the verb and the particle make independent contributions to the sentence. A (partial) Parsons-style semantic analysis of this might be as follows:

$$\text{put}(e1, x, y) \wedge \text{peter}(x) \wedge \text{picture}(y) \wedge \text{up}(e1, y)$$

Sentence (4), on the other hand requires a rather different analysis. Neither Barbara nor Simon can be said to have *made* or to be *out*. The semantic analysis we would want then might be something like the following:

$$\text{make\_out}(e1, e2) \wedge \text{and}(e2, x, y) \wedge \text{barbara}(x) \wedge \text{simon}(y)$$

How are we to identify whether the first or the second kind of semantic representation is appropriate for any given item? If we look at the other two sentences we can see that the problem is even more complicated. In (2) it is the case that the paper is finished, but it would be hard to claim that anything or anyone is up. Only the verb then seems to be contributing its simplex meaning, and the semantic analysis is (roughly):

$$\text{finish}(e1, x, y) \wedge \text{susan}(x) \wedge \text{paper}(y)$$

In (3), by contrast, it is the particle that contributes its simplex meaning and not the verb. As a consequence of Philip's action the intruder is *down*, but since there is no simplex verb *to gun*, we would not say that anyone *gunned* or *was gunned* The semantic analysis is consequently as follows:

$$\text{gun\_down}(e1, x, y) \wedge \text{philip}(x) \wedge \text{intruder}(y) \wedge \text{down}(e1, y)$$

In the linguistic literature, the semantics of VPCs is frequently viewed in rather more complicated terms than we are suggesting here, with particles often seen as making significant construction-specific contributions in terms of aspect (e.g. Brinton (1985)). However no such existing linguistic account is completely robust, and for practical NLP purposes we are forced to adopt a rather straightforward definition of compositionality as meaning that the overall semantics of the MWE can be composed from the simplex semantics of its parts, as described (explicitly or implicitly) in a finite lexicon.

3 Building the Resource

Rather than attempting to model compositionality by anchoring word semantics to a given lexicon, our approach in this work is to defer to an empirical reference based on human judgements. We define MWE compositionality to be an entailment relationship between the whole and its various parts, and solicit entailment judgements based on a handful of example sentences.

Entailment is conventionally defined for logical propositions, where a proposition P entails a proposition Q iff there is no conceivable state of affairs that could make P true and Q false. This can be generalised to refer to the relationship between two verbs V1 and V2 that holds when the sentence *Someone V1s* entails the sentence *Someone V2s* (see, e.g., the treatment of verbs in the WordNet hierarchy (Miller et al., 1990)). According to this generalisation we would then say that the verb *run* entails the verb *move* because the sentence *He runs* entails the sentence *He moves*. The same idea can be generalised to the relationship between simplex verbs (e.g. *walk*) and VPCs (e.g. *walk off*). For example, sentence (1) can be said to entail that *Peter put the picture somewhere* and so we can say that *put up* entails *put*. The same might be said of *finish up* and *finish* in (2). However, (3) and (4) produce a rather different result. (4) does not entail that *Simon and Barbara made* something, and (3) cannot entail that *Philip gunned the intruder* because there is no simplex verb *to gun*. This is a very useful way of testing whether the simplex verb contributes to the meaning of the construction.

We can approach the relationship between VPCs and particles in this same way. For (1), while it is not true that *Peter was up*, it is true that *The picture was up*. We can therefore say that the VPC entails the particle here. For (2), it is not true that either Susan or the paper were up, and the VPC therefore does not entail the particle. In the case of (3), while it is not true that *Philip was down* it is true that *The intruder was down*, and the VPC therefore entails the particle. Finally, for (4), it is not true that Barbara and Simon were out, and the VPC therefore does not entail the particle.

We make the assumption that these relationships between the component words of the VPC and the whole are intuitive to non-experts, and aim to use their "entailment" judgements accordingly. This use of entailment in exploring the semantics of verb and preposition combinations was first proposed by Hawkins (2000), and applied to VPCs by Lohse et al. (in preparation).

3.1 Experimental Materials

In an attempt to normalise the annotators' entailment judgements, we decided upon an experimental setup where the subject is, for each VPC type, presented with a fixed selection of sentential contexts for that VPC. So as to avoid introducing any bias into the experiment through artificially-generated sentences, we chose to extract the sentences from naturally-occurring text, namely the written component of the British National Corpus (BNC, Burnard (2000)).

Extraction of the VPCs was based on the method of Baldwin and Villavicencio (2002). First, we used a POS tagger and chunker (both built using fnTBL 1.0 (Ngai and Florian, 2001)) to (re)tag the BNC. This allowed us to extract VPC tokens through use of: (a) the particle POS in the POS tagged output, for each instance of which we simply then look for the rightmost verb within a fixed window to the left of the particle, and (b) the particle chunk tag in the chunker output, where we similarly locate the rightmost verb associated with each particle chunk occurrence. Finally, we ran a stochastic chunk-based grammar over the chunker output to extend extraction coverage to include mistagged particles and also more reliably determine the valence of the VPC. The token output of these three methods was amalgamated by weighted voting.

The above method extracted 461 distinct VPC types occurring at least 50 times, attested in a total of 110,199 sentences. After partitioning the sentence data by type, we randomly selected 5 sentences for each VPC type. We then randomly selected 40 VPC types (with 5 sentences each) to use in the entailment experiment. That is, all results described in this paper are over 40 VPC types.

3.2 Participants

28 participants took part in our initial experiment. They were all native speakers of English, recruited by advertisements posted to newsgroups and mailing lists.

3.3 Experimental Method

Each participant was presented with 40 sets of 5 sentences, where each of the five sentences contained a particular VPC. The VPC in question was indicated at the top of the screen, and they were asked two questions: (1) whether the VPC implies the verb, and (2) whether the VPC implies the particle. If the VPC was *round up*, e.g., the subject would be asked "Does *round up* imply *round*?" and "Does *round up* imply *up*?", respectively. They were given the option of three responses: "Yes", "No" or "Don't Know". Once they had indicated their answer and pressed next, they advanced to the next VPC and set of 5 sentences. They were unable to move on until a choice had been indicated.

As with any corpus-based approach to lexical semantics, our study of VPCs is hampered by polysemy, e.g. *carry out*$_{\mathrm{TRANS}}$in the *execute* and *transport out (from a location)* senses.[1] Rather than intervene to customise example sentences to a prescribed sense, we accepted whatever composition of senses random sampling produced. Participants were advised that if they felt more that one meaning was present in the set of five sentences, they should base their decision on the sense that had the greatest number of occurrences in the set.

[1]The effects of polysemy were compounded by not having any reliable method for determining valence. We consider that simply partitioning VPC items into intransitive and transitive usages would reduce polysemy significantly.

VPC	*Component word*	*Yes*	*No*	*Don't Know*
get down	get	19	5	2
	down	14	10	2
move off	move	14	12	0
	off	19	7	0
throw out	throw	20	6	0
	out	15	10	1
pay off	pay	11	12	3
	off	16	8	2
lift out	lift	25	1	0
	out	26	0	0
roll back	roll	13	9	4
	back	14	12	0
dig up	dig	21	5	0
	up	18	7	1
lie down	lie	24	2	0
	down	25	1	0
wear on	wear	6	19	1
	on	3	22	1
fall off	fall	23	3	0
	off	25	1	0
move out	move	22	4	0
	out	26	0	0
hand out	hand	15	9	2
	out	19	7	0
seek out	seek	13	13	0
	out	15	11	0
sell off	sell	14	12	0
	off	16	9	1
trail off	trail	8	18	0
	off	10	16	0
stay up	stay	20	5	1
	up	21	5	0
go down	go	18	7	1
	down	22	3	1
hang out	hang	22	4	0
	out	25	1	0
get back	get	20	6	0
	back	19	6	1
throw in	throw	15	9	2
	in	13	12	1
put off	put	8	17	1
	off	5	19	2
shake off	shake	12	14	0
	off	15	11	0
step off	step	25	1	0
	off	26	0	0
give off	give	12	12	2
	off	21	5	0
carry away	carry	7	17	2
	away	6	18	2
throw back	throw	18	7	1
	back	21	4	1
pull off	pull	13	10	3
	off	13	6	7
carry out	carry	0	25	1
	out	0	25	1
brighten up	brighten	9	16	1
	up	16	10	0
map out	map	9	17	0
	out	10	16	0
slow down	slow	11	14	1
	down	19	7	0
sort out	sort	6	19	1
	out	11	15	0
bite off	bite	15	10	1
	off	16	8	2
add up	add	12	14	0
	up	19	6	1
mark out	mark	13	13	0
	out	14	12	0
lay out	lay	11	14	1
	out	10	14	2
catch up	catch	6	20	0
	up	7	18	1
run up	run	12	13	1
	up	13	10	3
stick out	stick	20	6	0
	out	15	11	0
play down	play	10	15	1
	down	6	20	0

Table 1: Participant entailment judgements

	Overall	*Verbs only*	*Particles only*
Agreement	.677	.703	.650
Kappa (κ)	.376	.372	.352
% Yes	.575	.655	.495
% No	.393	.319	.467
% Don't Know	.032	.026	.038

Table 2: Summary of judgements for all VPCs

The experiment was conducted remotely over the Web, using the experimental software package WebExp (Corley et al., 2000). Experimental sessions lasted approximately 20 minutes and were self-paced. The order in which the forty sets of sentences were presented was randomised by the software.

3.4 Annotator agreement

We performed a pairwise analysis of the agreement between our 28 participants. The overall mean agreement was .655, with a kappa (κ) score (Carletta, 1996) of .329. An initial analysis showed that two participants strongly disagreed with the other, achieving a mean pairwise κ score of less than .1. We decided therefore to remove these from the set before proceeding. The overall results for the remaining 26 participants can be seen in Table 2. The κ score over these 26 participants (.376) is classed as fair (0.2–0.4) and approaching moderate (0.4–0.6) according to Altman (1991).

As mentioned above, a major problem with lexical semantic studies is that items tend to occur with more than one meaning. In order to test the effects of polysemy in the example sentences on inter-annotator agreement, we analysed the agreement obtained over those VPCs which have only one meaning according to WordNet (Miller et al., 1990). There was a total of 14 such items, giving 28 entailment judgements (one for the verb and one for the particle in each item). For these items, mean agreement and the κ score were .700 and .387, respectively. These are only very slightly higher than the overall scores, suggesting, although by no means proving, that polysemy was not a significant confounding factor.

The results for each VPC type can be seen in Table 1, broken down into the verb and particle entailment judgements and based on the 26 participants. We took two approaches to deriving a single judgement for each test. First, we took the majority judgement to be the correct one (**majority**). Second, we identified the participant who achieved the highest overall κ score with the other participants, and took their judgements to be correct (**centroid annotator**). Both sets of results will be referred to in evaluating our models.

It is interesting to look at the way in which the results for component entailment are distributed across the VPCs. According to the majority view, there are 21 fully-compositional items, 10 items where neither the verb nor the particle is entailed, 9 items where only the particle is entailed, and 0 items where the verb alone is entailed. According to the judgements of the centroid annotator, there are 10 fully-compositional items, 12 items where neither the verb nor the particle is entailed, 15 where only the verb is entailed, and 3 where only the particle is entailed. It is surprising to notice that the majority view holds there to be no items in which the verb alone is contributing meaning. It could be the case that items where only the verb contributes meaning are rare, or that they are not represented in our dataset. Another possible, and to our minds more likely, conclusion is that the contribution of the head verb strongly affects the way in which participants view the whole item. Thus if a verb is considered to be contributing simplex semantics, the participant is likely to assume that the VPC is completely compositional, and conversely if a verb is considered to not be contributing simplex semantics, the participant is more likely to assume the VPC to be non-compositional.

4 Related Work

We devote this section to a description of statistical NLP work on the non-compositionality of MWEs.

Perhaps the singularly most influential work on MWE non-compositionality is that of Lin (1999). We describe Lin's method in some detail here as it forms the basis of one of the methods tested in this research. Lin's method is based on the premise that non-compositional items have markedly different distributional characteristics to expressions derived through synonym substitution over the original word composition. Lin took his multiword items from a collocation database (Lin, 1998b). For each collocation, he substituted each of the component words with a word with a similar meaning. The list of similar meanings was obtained by taking the 10 most similar words according to a corpus-derived thesaurus, the construction of which is described in Lin (1998a). The mutual information value was then found for each item produced by this substitution by taking a collocation to consist of three events: the type of dependency relationship, the head lexical item, and the modifier. A phrase α was then said to be non-compositional iff there exists no phrase β where: (a) β can be produced by substitution of the components of α as described above, and (b) there is an overlap between the 95% confidence interval of the mutual information values of α and β. These judgements were evaluated by comparison with a dictionary of idioms. If an item was in the dictionary then it was said to be non-compositional. Scores of 15.7% for precision and 13.7% for recall are reported.

There are, to our minds, significant problems with the underlying assumptions of Lin's method. The theoretical basis of the technique is that compositional items should have a similar distribution to items formed by replacing components words with semantically similar ones. The idea presumably is that if an item is the result of the free combination of words, or a fully productive lexical rule, then word-substituted variants should be distributed similarly. This seems a reasonable basis for modelling **productivity** but not **compositionality**, as Lin claims. There are many examples in natural language of phrases that are not at all productive but are still compositional (e.g. *frying pan*); we term the process by which these expressions arise **institutionalisation** . Similar work to Lin's has been done in the area of collocation extraction (e.g. Pearce (2002)), to pick up on this alternate concept of institutionalisation.

Schone and Jurafsky (2001) employed Latent Semantic Analysis (LSA, Deerwester et al. (1990)) in an effort to improve on existing techniques for extracting MWEs from corpora. One property they try and pick up on in doing so is non-compositionality. They measure the cosine between the vector representation for the candidate MWE and a weighted vector sum of its component words, suggesting that a small cosine would indicate compositionality. They evaluate this by comparing the extracted items with those listed in existing dictionaries, and report that it offers no improvement in extracting MWEs over existing techniques. The assumption that non-compositionality is requisite for the presence of a MWE in a dictionary, while interesting, is not well-founded, and hence it does not seem to us that the poor results reflect a failure of the LSA approach in measuring compositionality.

Bannard (2002) used a combination of hand-built thesauri and corpus statistics to explore the compositionality of VPCs. The task was to predict whether the verb and/or the particle were contributing meaning to a given item, using statistical analysis of a set of VPCs extracted from the Wall Street Journal section of the Penn Treebank (Marcus et al., 1993). Two techniques were used. The first of these loosely followed Lin in measuring the extent to which the component verb or particle of any VPC could be replaced with items of a similar semantic class to form a corpus-attested VPC; WordNet (Miller et al., 1990) was used as the source for verb substitution candidates, and a hand-build semantic taxonomy for particles. The second technique explored the semantic similarity of a VPC to its component verb by comparing their subcategorisation preferences, assuming that semantic similarity between a VPC and its component verb indicates compositionality. Poor results were put down to data-sparseness, and the lexical resources not being well suited to the task. We use a larger corpus and an automatically-derived thesaurus for the research described in this paper, with the hope of overcoming these problems.

McCarthy et al. (2003) carry out research close in spirit to that described here, in taking VPC tokens automatically extracted from the BNC and using an automatically acquired thesaurus to classify their relative compositionality. One significant divergence from our research is that they consider compositionality to be an indivisible property of the overall VPC, and not the individual parts. Gold-standard data was generated by asking human annotators to describe the compositionality of a given VPC according to a 11-point scale, based upon which the VPCs were ranked in order of compositionality. Similarly to this research, McCarthy et al. in part used the similarity measure of Lin (1998a) to model compositionality, e.g., in taking the top N similar words to each VPC and looking at overlap with the top N similar words to the head verb. They also examine the use of statistical tests such as mutual information in modelling compositionality, and find the similarity-based methods to correlate more highly with the human judgements.

Baldwin et al. (2003) use LSA as a technique for analysing the compositionality (or decomposability) of a given MWE. LSA is suggested to be a construction-inspecific test for compositionality, which is illustrated by testing its effectivity over both English noun-noun compounds and VPCs. Baldwin et al. used LSA to calculate the distributional similarity between an MWE and its head word, and demonstrate a correlation between similarity and compositionality (modelled in terms of endocentricity) by way of items with higher similarity being more compositional. They do not go as far as to classify MWEs as being compositional or non-compositional, however.

5 Building a classifier

Having created our gold-standard data, we implemented some statistical techniques for automatic analysis. In this, we use the VPC tokens with sentential contexts extracted from the BNC as reported in Section 3, i.e. a superset of the data used to annotate the VPCs. We mapped the gold-standard data onto four binary (yes/no) classification tasks over VPC items:

TASK 1: The item is completely compositional.

TASK 2: The item includes at least one item that is compositional.

TASK 3: The verb in the item contributes its simplex meaning.

TASK 4: The particle in the item contributes its simplex meaning.

Note the partial conditional chaining between these tests, e.g. an item for which the verb and particle contribute their simplex meaning (i.e. positive exemplars for TASKS 3 and 4) is completely compositional (i.e. a positive exemplar for TASK 1).

The following sections describe four methods for modelling VPC compositionality, each of which is tested over the 4 individual compositionality classification tasks. The results for each method are given in Table 4, in which the baseline for each task is the score obtained when we assign the most frequent label to all items. Each method is evaluated in terms of precision (***Prec***), Recall (***Rec***) and F-score ($\beta = 1$, ***FB1***), and all values which exceed the baseline are indicated in **boldface**.

5.1 Method 1

We decided to gain a sense of the start-of-the-art on the task by reimplementing the technique described in Lin (1999) over VPCs. In our implementation we replaced Lin's collocations with our VPCs, treating the relationship between a verb and a particle as a kind of grammatical relation. In addition to the binary compositional/non-compositional judgement that Lin offers (which seems to be equivalent to TASK 1), we tested the method over the other three tasks. Acknowledging, as we must, that items can be partially compositional (i.e. have one component item contributing a conventional meaning), it would seem to be the case, according to the assumptions made by the technique, that the substitutability of each item will give us some insight into its semantic contribution. The thesaurus used by Lin has been generously made available online. However this is not adequate for our purposes since it includes only verbs, nouns and adjectives/adverbs. We therefore replicated the approach described in Lin (1998a) to build the thesaurus, using BNC data and including prepositions.

5.2 Method 2

Method 2 is very similar to Method 1, except that instead of using a thesaurus based on Lin's method, we took a knowledge-free approach to obtaining synonyms. Our technique is very similar to the approach taken to building a "context space" by Schütze (1998). We measured the frequency of co-occurrence of our target words (the 20,000 most frequent words, including all of our VPCs[2] and all of their component verbs and prepositions) with a set of 1000 "content-bearing" words (we used the 51st to the 1050th most frequent words, the 50 most frequent being taken to have extremely low infomation content). A target word was said to co-occur with a content word if that content word occurred within a window of 5 words to either side of it. These co-occurrence figures were stored as feature vectors. In order to overcome data sparseness, we used techniques borrowed from Latent Semantic Indexing (LSI, Deerwester et al. (1990)). LSI is an information retrieval technique based on Singular Value Decomposition (SVD), and works by projecting a term-document matrix onto a lower-dimensional subspace, in which relationships might more easily be observed between terms which are related but do not co-occur. We used this technique to reduce the feature space for our target words from 1000 to 100, allowing relations to be discovered between target words even if there is not direct match between their context words. We used the various tools in the GTP software package, created at the University of Tennessee[3] to build these matrices from the co-occurrence data, and to perform SVD analysis.

We calculated the similarity between two terms by finding the cosine of the angle between their vectors. We performed a pairwise comparison between all verbs and all particles. For each term we then sorted all of the other items of the same part-of-speech in descending order of similarity, which gave us the thesaurus for use in substitution. As with the Lin method, we performed substitutions by taking the 10 most similar items for the head verb and particle of each VPC.

	Majority	***Centroid***	***60% Agreement***
All	1.29 (p =.255)	4.09 (p=.043)	0.48 (p=.488)
Monosemous	2.19 (p=.137)	0.01 (p=.924)	5.56 (p =.018)

Table 3: Logistic regression for Method 4

5.3 Method 3

We noted in Section 4 that a significant problem with the substitution approach is that it is sensitive to institutionalisation rather than non-compositionality. Method 3 attempts to adapt substitution to more accurately reflect non-compositionality by removing the assumption that an item formed by substitution should have the same distributional characteristics as the original item. Rather than basing the compositionality judgement on the relative mutual information scores of the original items and the items resulting from substitution, we instead base it on the corpus-based semantic similarity between the original expression and word-substituted derivative expressions. The same method of substitution is used, with each component being replaced by each of its 10 nearest neighbours according to the knowledge-free similarity measure described above. We judge a VPC item to be compositional if an expression formed by substitution occurs among the nearest 100 verb-particle items to the original, and failing this, we judge it to be non-compositional. We experimented with a number of cut-off points for identifying semantically similar items, and found that a value of 100 gave the best results.

5.4 Method 4

While Method 3 softens the reliance upon productivity as a test for compositionality, it still confuses insti-

[2] Concatenated into a single-word item

[3] `http://www.cs.utk.edu/~lsi/soft.html`

	Task 1 (mean agreement = .693)						Task 2 (mean agreement = .750)					
	Majority			*Centroid annotator*			*Majority*			*Centroid annotator*		
	Prec	*Rec*	*FB1*	*Prec*	*Rec*	*FB1*	*Prec*	*Rec*	*FB1*	*Prec*	*Rec*	*FB1*
Baseline	.525	1.000	.680	.250	1.000	.400	.750	1.000	.860	.700	1.000	.820
Method 1	**.577**	.714	.638	**.269**	.700	.389	.731	.633	.678	**.731**	.679	.704
Method 2	**.575**	.714	.638	**.308**	.800	**.447**	**.769**	.667	.717	**.769**	.714	.739
Method 3	**.558**	.905	**.690**	.235	.800	.360	**.765**	.866	.810	**.735**	.892	.810
Method 4	.514	.857	.642	.200	.700	.280	**.771**	.900	.830	**.714**	.893	.794

	Task 3 (mean agreement = .729)						Task 4 (mean agreement = .688)					
	Majority			*Centroid annotator*			*Majority*			*Centroid annotator*		
	Prec	*Rec*	*FB1*	*Prec*	*Rec*	*FB1*	*Prec*	*Rec*	*FB1*	*Prec*	*Rec*	*FB1*
Baseline	.525	1.000	.690	.625	1.000	.770	.750	1.000	.857	.670	1.000	.800
Method 1	.474	.429	.450	**.632**	.480	.546	**.818**	.300	.442	.454	.385	.417
Method 2	**.608**	.666	.639	**.782**	.720	.749	**.818**	.300	.442	.454	.385	.417
Method 3	**.531**	.810	.641	**.625**	.800	.717	**.769**	.333	.480	.308	.308	.308
Method 4	**.666**	.286	.400	**.666**	.240	.353	**.758**	.833	.793	.303	.769	.435

Table 4: Results for the four methods over the different compositionality classification tasks

tutionalisation with non-compositionality somewhat in its reliance upon substitution. We now suggest another technique which we claim is based on sounder principles. The underlying intuition is that identifying the degree of semantic similarity between a VPC and its component verb and/or particle will indicate whether that component part contributes independent semantics. This is similar to the assumption made in Schone and Jurafsky (2001), except that we make a distinction between the contribution of the different component parts. We again used the knowledge-free semantic similarity measure. We performed a pairwise comparison of all VPCs with all verbs and all particles, obtaining cosine similarity scores for each pair.

In order to measure the usefulness of this score, we performed a logistic regression of the similarity scores and the human judgements as to whether the given verb or particle is entailed by the VPC. We did this for the majority human judgements, and also the centroid annotator scores. We also did the same using the majority scores but rejecting those items on which there was less than 60% agreement. In addition to performing a regression for all items (***All***), we also performed a regression for only those items which have only one meaning according to WordNet (***Monosemous***). The results for all of these are shown in Table 3. The figures shown are chi-squared scores, with their associated significance values. We observed significant correlations for a number of the regressions (notably all items vs. the centroid annotator, and monosemous items vs. 60% agreement). While the results are far from stable, such variation is perhaps to be expected on a test like this since the nature of context space models means that rogue items sometimes get extremely high similarity scores, and we are performing the regression over only 40 VPCs (80 VPC-component pairs).

In order to build a classifier for making compositionality decisions, we again used a neighbour-based approach with a cut-off. We said that a verb was contributing meaning to a VPC if it occurred in the 20 most similar items to the VPC. For particles, we said that the item was contributing meaning if it was among the 10 nearest neighbours. We tried out a range of different cut-offs for each item and found that these gave the best results.

6 Results

The results in Table 4 show that on all tasks (for the majority-view based data and three out of four for the centroid data), at least one of the four statistical methods offers an improvement in precision over the baseline, and that there is an improvement in F-score for Task 1 on both sets of data. There are swings in the relative scores obtained over the majority as compared to centroid annotator data for a given task. In terms of relative performance, the semantic similarity based approach of Methods 3 and 4 outperform the distribution based approach of Methods 1 and 2 in terms of F-score, on 6 of the 8 sets of results reported.

In order to get a reliable sense for how good these scores are, we compare them with the level of agreement across human judges. We calculated pairwise agreement across all participants on the four classification tasks, resulting in the figures given in Table 4. These agreement scores give us an upper bound for classification accuracy on each task, from which it is possible to benchmark the classification accuracy of the classifiers on that same task. On Task 1, three of the four classifiers achieved a classification accuracy of .575. On Task 2, the highest-performing classifier (Method 4), achieved a classification accuracy of .725. On Task 3, Method 2 achieved the highest classification accuracy at .600, and on Task 4, Method 4 achieved a classification accuracy of .675. We can see then that the best classifiers perform only marginally below the upper bound on at least two of the tasks.

While these results may appear at first glance to be less than conclusive, we must bear in mind that we are working with limited amounts of data and relatively simplistic models of a cognitively intensive task. We interpret them as very positive indicators of the viability of using empirical methods to analyse VPC semantics.

7 Conclusion

This paper has described the implementation and evaluation of four corpus-based approaches to the semantics of verb-particle constructions. We created a set of gold-standard data, based on non-expert judgements acquired via a web-based experiment. We then implemented four different techniques and showed that they offer a significant improvement over a naive approach.

Acknowledgements

We would like to thank Ann Copestake, Maria Lapata, Diana McCarthy, Aline Villavicencio, Tom Wasow, Dominic Widdows and the three anonymous reviewers for their valuable input on this research. Timothy Baldwin is supported by the National Science Foundation under Grant No. BCS-0094638 and also the Research Collaboration between NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation and CSLI, Stanford University. Colin Bannard is supported by ESRC Grant PTA-030-2002-01740

References

Douglas G. Altman. 1991. *Practical Statistics for Medical Research*. Chapman and Hall.

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, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proc. of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*. (this volume).

Colin Bannard. 2002. Statistical techniques for automatically inferring the semantics of verb-particle constructions. *LinGO Working Paper No. 2002-06*.

Ken Barker and Stan Szpakowicz. 1998. Semi-automatic recognition of noun modifier relationships. In *Proc. of the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics (COLING/ACL-98)*, pages 96–102, Montreal, Canada.

Laurel Brinton. 1985. Verb particles in English: Aspect or aktionsart. *Studia Linguistica*, 39:157–68.

Lou Burnard. 2000. *User Reference Guide for the British National Corpus*. Technical report, Oxford University Computing Services.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

Martin Corley, Frank Keller, and Christoph Scheepers. 2000. *Conducting psychological experiments over the world wide web*. Unpublished manuscript, University of Edinburgh and Saarland University.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6).

John A. Hawkins. 2000. The relative order of preposition phrases in English: Going beyond manner – place – time. *Language Variation and Change*, 11:231–266.

Mark Lauer. 1995. *Designing Statistical Language Learners: Experiments on Noun Compounds*. Ph.D. thesis, Macquarie University.

Dekang Lin. 1998a. Automatic retrieval and clustering of similar words. In *Proc. of the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics (COLING/ACL-98)*, Montreal, Canada.

Dekang Lin. 1998b. Extracting collocations from text corpora. In *First Workshop on Computational Terminology*.

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proc. of the 37th Annual Meeting of the ACL*, pages 317–24, College Park, USA.

Barbara Lohse, John A. Hawkins, and Tom Wasow. in preparation. Domain minimization in English verb-particle constructions.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, 19(2):313–30.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proc. of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*. (this volume).

I. Dan Melamed. 1997. Automatic discovery of non-compositional compounds in parallel data. In *Proceedings of 2nd Conference on Empirical Methods in Natural Language Processing*.

G.A. Miller, R. Beckwith, C. Fellbaum, D Gross, and K.J. Miller. 1990. Introduction to WordNet: an on-line lexical database. *International Journal of Lexicography*, 3(4):235–44.

Grace Ngai and Radu Florian. 2001. Transformation-based learning in the fast lane. In *Proc. of the 2nd Annual Meeting of the North American Chapter of Association for Computational Linguistics (NAACL2001)*, pages 40–7, Pittsburgh, USA.

Darren Pearce. 2002. A comparative evaluation of collocation extraction techniques. In *Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Canary Islands.

Barbara Rosario and Marti Hearst. 2001. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In *Proc. of the 6th Conference on Empirical Methods in Natural Language Processing (EMNLP 2001)*, Pittsburgh, USA.

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

Patrick Schone and Dan Jurafsky. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proc. of the 6th Conference on Empirical Methods in Natural Language Processing (EMNLP 2001)*, pages 100–108.

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

Aline Villavicencio and Ann Copestake. 2002. Phrasal verbs and the LinGO-ERG. *LinGO Working Paper No. 2002-01*.