Differentiating Types of Verb Particle Constructions

Jon Patrick and Jeremy Fletcher
Sydney Language Technology Research Group
School of Information Technologies
University of Sydney
Sydney, Australia, 2006
jonpat@it.usyd.edu.au, jfletch1@it.usyd.edu.au

Abstract

A verb particle construction (VPC) classification scheme gleaned from linguistic sources has been used to assess its usefulness for identifying issues in decomposability. Linguistic sources have also been used to inform the features suitable for use in building an automatic classifier for the scheme with a series of good performance results. The notions of how to define the task of computing phrasal verbs are discussed and new proposals are presented.

1 Introduction

Our area of research focuses on verb-particle constructions (VPCs), a sub-section of multi-word expressions (MWEs). MWEs are a generic term for the group of expressions that include idioms (e.g. *over the moon*), lexically fixed expressions (e.g. *ad hoc*), light verb constructions (e.g. *make a mistake*), institutionalised phrases (e.g. *kindle excitement*), and verb-particle constructions (e.g. *run away*). All these expressions have in common the occurrence of words adjacent to each other that would be more frequent than if they were simply random words put together. Hence the words which constitute them have some particular meaning together that they would not have apart.

VPCs consist of a simplex (single-word) verb, and a particle, whether preposition or adverb. A particular subset of interest are "phrasal" verbs which are considered to be non-decomposable structures where the meaning is in the whole and not the parts of the phrase (Dixon, 1978).

Previous research into the area of MWEs, shows a number of places in which there has been a lack of research. The area which is of interest to us is the approach of bringing reliable resources and specific encodeable linguistic knowledge for use in feature selection in supervised learning tasks of categorisation and WSD.

2 Previous Research

There has been only limited research done in the field of MWEs in computational linguistics, mostly focusing on VPCs; although much research (Abeillé, 1988, 1995; Barkema, 1994; Wehrli, 1998) has been done on such MWEs as idioms (Abeillé, 1995; Barkema, 1994; Wehrli, 1998) and light verbs (Abeillé, 1988) within the traditional linguistics field. MWEs are very idiosyncratic constructs making progress in their computation difficult. They have been called "a serious problem for many areas of language technology" (Copestake, Lambeau, Villavicencio, Bond, Baldwin, Sag, Flickinger, 2002) "unpredictable" (Baldwin & Villavicencio, 2002) and "a pain in the neck" (Sag, Baldwin, Bond, Copestake and Flickinger. 2002), and thus many applications of computational linguistics such as machine translation put multiword expressions in the "too-hard" basket. However, the use of MWEs in natural language is widespread, and thus comes about the need for what has been called "a robust, structured handling [of MWEs]" (Baldwin & Villavicencio, 2002, Calzolari, Fillmore, Grishman, Ide, Lencu, MacLeod, and Zampolli. 2002).

Much of the previous work on VPCs has revolved around tasks besides WSD, focused on the extraction of multiple word constructs from corpora (Abeillé,1988, Baldwin & Villavicencio, 2002, Maynard & Ananiadou, 1999.) which we will discuss in some detail below, or what is termed the "decomposability" or "compositionality" of VPCs.

2.1 Extraction of MWEs from Corpora

One of the important components of research in MWEs is their automatic extraction from some corpora. The best results for precision and recall on this task is clearly the work of Baldwin and Villavicencio (2002), who used a combination of part-of-speech (POS)-based extraction (using Brill's POS tagger (Brill, 1995)), chunk-based

Proceedings of the Australasian Language Technology Workshop 2004
Macquarie University, Sydney, December 8th, 2004.

Accepted after full paper review

extraction, and chunk-grammar-based extraction, to extract VPCs from the Wall Street Journal section of the Penn Treebank.

They report precision of 0.889 and recall of 0.903 (Fβ1 = 0.896) for this task. Although they cite some other research into a similar area, there had been very little research in this specific area of extracting VPCs automatically from corpora, although some studies without quantitative analysis had been done previously (Kaalep & Muischnek, 2002, Krenn & Evert, 2001), along with work on the extraction of other collocations other than MWEs (Abeillé, 1988, Basili, Pazienza and Velardi, 1993, Maynard and Ananiadou, 1999).

VPC Example	Verb Contributes to Meaning	Particle Contributes to Meaning
1. Peter put the picture up	Yes	Yes
2. Susan finished up her paper	Yes	No
3. The thief made away with the cash	No	Yes
4. Barbara and Simon made out	No	No

Table 1. Summativity of VPCs; whether individual elements determine the meaning of the construction.

2.2 Determining the Decomposability of MWEs

As the semantics of most MWEs are more difficult to ascertain than the semantics of simplex words (even taking into account the problem of disambiguating between different senses of a simplex word), there has been some research done into what is termed the "compositionality" or "decomposability" of VPCs. In Bannard, Baldwin & Lascarides (2003) some examples are given of VPCs which illustrate one way of describing "summativity" (See table 1). In these examples, 1) "put up", is entirely composed of its constituent parts, as at the end of the action, the painting is both "put" somewhere, and is now "up". In 2) "finished up", the paper is "finished", but nothing is "up", in 3) "made away", the thief is "away", but nothing is "made", and in 4) nothing is either "made" or "out".

Lin (1999) discusses the principle of decomposability with regards to constructs other than VPCs, (e.g. "red tape"), and conducted an experiment whereby individual words in the collocation are replaced by words with similar meanings taken from a thesaurus.

This word-substitution technique is transferred to VPCs in Bannard (2002) where he suggests that a similar approach could be used to determine the "decomposability" of VPCs. He obtains disappointing results ranging from precision of 0.516 and recall of 0.739 (for the largest class) to precision of 0.286 and recall of 0.083 (for the smallest class).

Bannard et al (2003) describe a statistical distribution modelling framework for determining whether specific VPCs are "decomposable", and hence automatically infer their meaning.

They conducted an internet-based experiment whereby non-expert native English speakers were asked whether certain VPCs, in sets of 5 sentences for each exemplar VPC, entailed the meaning of the simplex verb and/or the meaning of the preposition. They used this as their gold standard test data to evaluate their results. This contrasts with the approach taken by Lin (1999), in that their evaluation is based on a more intuitive level, although this could in fact lead to their results becoming affected by subjective judgments. They then encapsulated the problem as a classification task, to classify VPCs into classes depending on whether or not they were composed of their individual elements. They achieved results that improved on a baseline of classifying everything as the most common class, with results ranging from 0.735 for precision and 0.892 for recall (Fβ1= 0.810) to 0.303 for precision and 0.769 for recall (Fβ1=0.435) by using their distribution modelling approach.

2.3 Word Sense Disambiguation (WSD)

There has been much general research done in the field of WSD (Krovetz & Croft, 1989, Maynard & Ananiadou, 1998, Yarowsky, 1992, 1995, for example), although very little relating specifically to disambiguating MWEs. O'Hara and Wiebe, 2003 do perhaps what is the most relevant research, on the task of disambiguating prepositions as having a locative, temporal or other meaning. They report average accuracy of 0.703, although they don't provide precision and recall scores, so it is difficult to ascertain the particular shortcomings of the system, and where improvements could be made. However, we can make the conjecture that had they determined the compositionality of the VPC to which the prepositional particle belongs, their reported accuracy would be higher.

Proceedings of the Australasian Language Technology Workshop 2004
Macquarie University, Sydney, December 8th, 2004.

Accepted after full paper review

3 Method

Our research goal is to create a system that can accurately disambiguate between the different semantics of different instances of VPCs. It is interesting to note that the task of disambiguating VPCs has not been undertaken by other researchers working on MWEs possibly for a number of reasons.

Firstly, there is no readily available corpus that has VPCs tagged for semantic differences. Hence, a major part of the work described is to collect a suitable subset of a given corpus – in this case, we use the British National Corpus (BNC) – and manually tag target VPCs as having certain semantic features.

Another reason why semantic disambiguation has not been undertaken to such a fine degree for VPCs is that there is, in fact, no comprehensive electronic resource which contains different senses of a given verb-particle construction, although there are large, readily available resources for different simplex verbs (the most obvious being WordNet (G. Miller, Beckwith, Fellcaum, Gross, and K. Miller. 1990).

Hence, the main difference between the research being undertaken in this project and that which has been done previously is the exploitation of a lexical database of phrasal verbs (constructed from Meyer, 1975). This will be used to increase the accuracy of our task on VPCs, and provide us with a compendium of "valid" verb-particle constructions. Also, whilst most semantic identification tasks relate to simplex words, there has been little research on disambiguation of MWEs. Disambiguation in this sense is most applicable to VPCs, as most other idiomatic MWEs have a single, fixed (if still metaphorical) meaning. For example, the idiom "kick the bucket", once it has been identified as a MWE has only the sense of "dying". Although the phrase itself could also have a literal sense of kicking, once it has been extracted and identified as an idiom, only this idiomatic, metaphoric sense applies. It is a similar situation with other idioms.

However, with VPCs, there are many examples that have different senses. For example the phrasal verb "to check out" has the sense of leaving a hotel, and the sense of checking something to make sure it is correct. Other examples of VPCs with multiple senses are "pick up" (understand/comprehend, retrieve from the ground, hook up with someone of the opposite sex), "look out" (look out of a window, watch out for) and "set up" (put into position, furnish with money/resources, establish, etc…).

4 Motivation

Unlike the current trend in other works that focus on recognition of VPCs from thesaural expansion of context whether automatically (Lin, 1999, Bannard et al, 2003, McCarthy et al, 2003) or using WordNet (Bannard et al 2003) and the establishment of gold standards by survey (Bannard et al, 2003, McCarthy et al, 2003) we prefer a principled method using the understandings developed by linguistic studies. The identification of (virtually) all candidate VPCs have been captured in 3 Phrasal Verb dictionaries (Collins, Oxford, and Meyer) so as a resource to determine the definitions of VPCs we use the dictionary of Meyer (1975). As well as the dictionaries, a variety of linguistic studies provide an extensive analysis of the features of VPCs. Our procedure is to use the dictionaries to guide the compilation of our gold standard and the studies to guide feature selection for automated classification to both establish the parameterisation of a generative model for recognising VPCs and to identify the strengths of the various features established by linguists.

The most comprehensive analysis of this problem has been completed by Dixon, (1982) and we use his classification scheme to guide the development of our own classification scheme and much feature selection has been gleaned from his work. For the definition of verbs we use the new Shorter Oxford English Dictionary (1993) and for particle description Lindstromberg, (1998). Dixon produced a 5 class classification scheme:

A. Literal usage of all VPC components, e.g. *John walked on the grass,*
B. Like A but with missing arguments that are reasonably inferable e.g. *He ran down (the bank) to the railway line,*
C. Obvious metaphorical extensions form literal phrases e.g. *the firm went under,*
D. non-literal constructions that cannot obviously be related to the literal form e.g. *They are going to have it out;*
E. Full idioms, e.g. *turn over a new leaf.*

This scheme is more extensive than that used in any computational study we have found, most of which attempt to resolve the compositional/non-compositional dichotomy of a VPC or at best provide a graded scale. Whilst the Dixon scheme is more detailed than other schemes we are not entirely satisfied with it. In our own studies we have come to recognise that there is a dimension of diversity in VPCs not captured by it. Some VPCs have become so established that they

Proceedings of the Australasian Language Technology Workshop 2004
Macquarie University, Sydney, December 8th, 2004.

Accepted after full paper review

Class	Description	Example(s)
N – Non-decomposable VPCs (Phrasal Verbs)	Verb-preposition pairs which are semantically related, and whose meaning is somewhat or wholly idiomatic.	*"Leeds United carried off a massive victory."* *"John and Julie made out."*
D – Decomposable VPCs	Verb-preposition pairs which are semantically related, but whose meaning is literal, or where the preposition is redundant.	*"The bee carried the pollen off to another flower"*
I – Independent verb-preposition pairs	Verb-preposition pairs which have no semantic relationship.	*"The cables carry around 1,000 volts"*

Table 2. Description of NDI classification scheme.

have a metaphorical sense derived from the original components and are found in the dictionaries catalogues as such, *have it off, kick off*. However they, like fully compositional VPCs, are used in both literal and metaphorical contexts based on the literality of the phrasal arguments, leading to the perception that compositionality is a continuum. In the case of *kick off*, our own experience is that even with literal arguments it feels metaphorical, e.g. *The game kicks off at 7pm*.

We believe that individual assessment of the literality of the arguments will establish a more reliable predictor of metaphoricity and thereby compositionality. For example, *the Government is driving down the road of disaster* has one more metaphorical argument than, *the man is driving down the road of disaster*. Furthermore, the relationship between the VPC and the head of the object is literal, and the metaphorical meaning is only created by the head modifier. Due to the manual effort in calibrating a corpus with this level of detail we have not incorporated such an analysis here, but we foreshadow it as future work.

A further deficit in this scheme is that it doesn't permit the distinction between full literal usage (class A) and literal usage of the compound where the particle is non-contributing (*do you think he will end up getting married*) or may even be removed (*are the computers linked/hooked (up) to the network*).

Further complications arise from features that can exist as one value at the sense level in the way you might define it in a dictionary but the context of usage changes those values, e.g. blow off.

Thus, we use the principles behind Dixon's classifications, along with our analysis of the problem to create a new 3-class classification scheme, which reflects more closely the real-life problem of identifying phrasal verbs.

The description of our NDI classification scheme appears in Table 2.

5 Determining Gold-standards

The work of Bannard et al (2003) determined a gold standard by randomly collecting 5 examples of a given VPC and as a collection asking non-experts to classify the components as compositional/non-compositional. We prefer an alternative approach where each individual sample sentence is assessed for all classifications of interest. Our work in this study has shown a significant diversity in random sampling so that we don't believe it can lead to a consistent result. This random sampling leads to imprecision in the classification task, due to a failure to create a homogenous data set. Indeed the poor inter-rater reliability in their study is testimony to this problem. Rather we have categorised each sample sentence extracted from the BNC.

We perceive that the notion of sense for VPCs appears at three major levels without restricting granularity within those levels. The first level is the compositional sense brought together by the union of the components, the second is the intrinsic sense that is more (or not) than the sum of the parts and conventionally recorded in a phrasal dictionary, and the third is the contextual sense that varies either of the other two senses in language usage. Hence automatic WSD will only be achieved by identifying each of these types of meaning making through fixed resources like dictionaries, manual analysis of the idiosyncratic usage in real language examples, that is corpus tagging, and the machine learning methods

Proceedings of the Australasian Language Technology Workshop 2004
Macquarie University, Sydney, December 8th, 2004.

Accepted after full paper review

appropriately parameterised for all variables of a linguistically motivated model. The current work is a beginning on the larger task of WSD for VPCs.

6 Data Selection

The data selection process consisted of firstly constructing a lexical database of Meyer's phrasal verb dictionary (1975) (denoted PV-Lex-Meyer) (Only entries A-O have been completed). The text was scanned and OCRed and then converted from a Word file into an XML database using the Ferret software (Patrick, Palko, Munro, Zappavigna, 2003). All VPCs in the database were extracted and all matching examples in the BNC retrieved on the criteria the verb and the particle had no intervening verb. This yielded over 600,000 sample sentences. To cut down the examples used, we sampled those VPCs which had medium density in the corpus; those verb-particle pairs which occurred in more than 10, but less than 40 sentences. There were approximately 70 such verb-particle pairs, giving us a reduced corpus of approximately 6000 sentences.

We used this reduced corpus as the basis for our initial classification task over the 3 possible tags for particles which occur in the BNC's CLAWS tag set, AV0 (adverb), AVP (adverbial preposition), PRP(general preposition). Our results for this experiment are shown in Table 3.

Class	Precision	Recall	F-Score
AV0	0.975	0.963	0.969
AVP	0.984	0.996	0.990
PRP	0.911	0.767	0.833

Table 3. Precison, Recall and F scores for estimating POS tags of particles.

The features used in this preliminary experiment were the POS tags of the words on either side of the verb and particle, the distance between the verb and particle, the particle's value. This experiment provided the separation of the data into the Clean and Noisy data sets, where Clean is correctly classified and Noisy is incorrectly classified.

Once we had achieved our best accuracy, we had a list of approximately 400 sentences which were not correctly identified by our classifier. We then manually tagged these 400 sentences using our VPC classification schemes.

This set provided examples of VPCs used in an atypical fashion. We use this set in our classification task, as the value of linguistic analysis is not tested by an entirely compliant data set. So although this sample is not representative of general usage, if gives a better measure of the performance of the linguistic features.

To complement these 400 sentences, and give us a rich problem to solve we sampled across the 5600 correctly classified sentences, to extract another 400. Our goal was to extract a sample of sentences which were representative of the trends in preposition usage (i.e. the POS tag of the particle), while still maintaining as wide a spectrum of verb-particle pairs as possible.

Of the extracted sentences, around 80% had particles labelled as adverbial particles, 10% as general adverbs and 10% as general prepositions. Thus, to maintain this distribution, we extracted 320 sentences which contained adverbial particles, 40 with general adverbs, and 40 sentences with general prepositions.

The representation of different VPCs according to CLAWS tags are: AV0-28, AVP-64, PRP-65. We thus targeted extraction numbers of 5 sentences for each VPC with an AVP tag, 2 sentences for each with an AV0 tag, and 1 sentence with a PRP tag. Given that some verbs occurred in less than the targeted number of sentences, we actually extracted a total of 423 example sentences.

We then manually tagged these 423 examples using our VPC schemes which was reduced to 376 when unusable sentences were deleted (principally for incorrect pre-processing, and in a few cases for being unintelligible). This set of approximately 800 examples was then used for our classification tasks. Our aim was to have a representative sample of the corpus, while maximising those examples where the classification was less obvious, and the number of different VPCs. Thus we included those examples where the use of the particle was more difficult to classify.

7 Results

The compositionality experiments are based on the NDI classification scheme, decomposable (D) (~44%), non-decomposable (N) (~36%) and Independent (I) (~19%), as this recognises the processing situation of identifying phrasal verbs in real text, rather than the somewhat artificial task of just discriminating between summative and non-summative constructs. The D class is the largest class and so sets the baseline as P=0.472, R=1, F=0.641 on the Noisy data set, P=0.412, R=1, F=0.584 on the Clean data set and P=0.442, R=1, F=0.613 on the Combined. Based on our initial linguistic analysis of the problem, we looked for a number of features within the target sentence. In

Proceedings of the Australasian Language Technology Workshop 2004
Macquarie University, Sydney, December 8th, 2004.

Accepted after full paper review

the first experiment (Table 4), the features included the particle being used, the distance between the verb and the particle in the VPC, the number of words this sentence has in common with example sentences extracted from our resource for this VP pair, whether this particle was part of one of the compound particles extracted during the manual annotation of the corpus, whether this VP pair has more senses in which it is transitive or intransitive (or neither, if there are an equal number of transitive and intransitive senses) also extracted from PV-Lex-Meyer, and the Dixon sub-categorisation frame.

In experiment 2 (Table 4) we constructed a more fine-grained scale for the transitivity measure, distinguishing between those PVs which only have transitive or intransitive senses, and those which have more transitive than intransitive senses.

	Exp 1 - Noisy			**Exp 2 - Noisy**		
	P	**R**	**F**	**P**	**R**	**F**
N	0.500	0.534	0.516	0.523	0.585	0.552
D	0.579	0.565	0.572	0.612	0.554	0.581
I	0.420	0.420	0.433	0.507	0.522	0.514
	Exp 1 - Clean			**Exp 2 - Clean**		
	P	**P**	**P**	**P**	**R**	**F**
N	0.502	0.502	0.502	0.513	0.649	0.573
D	0.461	0.461	0.461	0.485	0.406	0.442
I	0.550	0.550	0.550	0.508	0.411	0.455
	Exp 1 - Combined			**Exp 2 - Combined**		
	P	**R**	**F**	**P**	**R**	**F**
N	0.459	0.609	0.523	0.502	0.605	0.549
D	0.506	0.390	0.441	0.545	0.452	0.494
I	0.550	0.452	0.561	0.556	0.556	0.556

Table 4. Performance statistics for experiments 1 and 2.

In Experiment 3a the length of the verb was used as a feature, given that, as Dixon says, "phrasal verbs are almost exclusively based on monosyllabic verbs of Germanic origin".

We considered that this length would give us a reasonable estimate of the number of syllables. This result looked encouraging, so we manually annotated each of the verbs in our resource for the number of syllables in the word.

In experiment 3b we used this number instead of the verb length. This showed an increase in two of the classes, but a decrease in the largest class. We hence included both these features in experiment 3c (see Table 5).

Some further analysis of the linguistics of these constructs led to the creation of a measure of the differences of the arguments of the verb. To do this, in experiment 4a we identified the POS tag of the head of each of the noun phrases surrounding the preposition. In experiment 4b we then generalised this distinction, to being either a named entity, or a general noun, given that this distinction occurs within the CLAWS tag set. These features were added to the features from experiment 3c. The results are shown in Table 6.

Following on from the potential phonological interpretation of Dixon's "Germanic origin" thesis, we also used the last three letters of the verb, each as an individual feature in experiment 5. However, we appreciate this is a primitive representation of the underlying linguistic model, and needs further maturity. See Table 7.

8 Conclusions

We have shown in the research a classification system built on encodeable linguistic knowledge can predict certain semantic information about a given instance of a VPC, to a level of accuracy comparable to that which has been achieved on the more coarse-grained approach of dealing with each verb-particle pair as a unit whose semantics remain the same in different contexts.

While VPCs in different contexts have vastly different semantic properties, we have also shown that it is possible to compute these semantics from features such as the syntactic structure of the VPC, and features of the arguments of the VPC.

Although the results are not presented here, there is evidence to suggest that if the distinction between these classes could be predicted reliably we could also get good results on predicting Dixon's class assignment of a given instance of a phrasal verb.

Our approach has come from grounding in the linguistic features of VPCs, to determine what the key distinctions between VPCs with different semantics are. Whilst no direct comparison can be drawn between the work presented here and the previous studies done on the "decomposability" of VPCs, our results show more stable solutions on a more complex task, which is also closer to realistic language processing.

Proceedings of the Australasian Language Technology Workshop 2004
Macquarie University, Sydney, December 8th, 2004.

Accepted after full paper review

	Experiment 3a Noisy			Experiment 3a Clean			Experiment 3a Combined		
Class	**Prec**	**Recall**	**F-Score**	**Prec**	**Recall**	**F-Score**	**Prec**	**Recall**	**F-Score**
N	0.570	0.619	0.593	0.533	0.655	0.588	0.502	0.545	0.523
D	0.627	0.619	0.623	0.537	0.465	0.498	0.533	0.520	0.527
I	0.443	0.391	0.415	0.550	0.452	0.496	0.583	0.521	0.550
	Experiment 3b Noisy			**Experiment 3b Clean**			**Experiment 3b Combined**		
	Prec	**Recall**	**F-Score**	**Prec**	**Recall**	**F-Score**	**Prec**	**Recall**	**F-Score**
N	0.536	0.627	0.578	0.548	0.689	0.611	0.522	0.624	0.568
D	0.664	0.589	0.625	0.519	0.439	0.476	0.564	0.480	0.518
I	0.544	0.536	0.540	0.525	0.425	0.470	0.565	0.549	0.557
	Experiment 3c Noisy			**Experiment 3c Clean**			**Experiment 3c Combined**		
	Prec	**Recall**	**F-Score**	**Prec**	**Recall**	**F-Score**	**Prec**	**Recall**	**F-Score**
N	0.579	0.619	0.598	0.579	0.682	0.623	0.535	0.632	0.579
D	0.667	0.655	0.661	0.576	0.523	0.549	0.579	0.511	0.543
I	0.547	0.507	0.526	0.550	0.452	0.496	0.553	0.514	0.533

Table 5. Performance statistics for experiments 3 a,b,c

	Experiment 4a - Noisy			Experiment 4b - Noisy		
	P	**R**	**F**	**P**	**R**	**F**
N	0.522	0.649	0.578	0.525	0.635	0.757
D	0.549	0.503	0.525	0.511	0.432	0.469
I	0.600	0.411	0.488	0.515	0.466	0.489
	Experiment 4a - Clean			**Experiment 4b – Clean**		
	P	**R**	**F**	**P**	**R**	**F**
N	0.586	0.576	0.581	0.578	0.627	0.602
D	0.670	0.702	0.686	0.671	0.667	0.669
I	0.540	0.493	0.515	0.533	0.464	0.496
	Experiment 4a - Combined			**Experiment 4b – Combined**		
	P	**R**	**F**	**P**	**R**	**F**
N	0.538	0.583	0.560	0.544	0.583	0.563
D	0.566	0.554	0.560	0.564	0.560	0.562
I	0.520	0.465	0.491	0.520	0.458	0.487

Table 6. Performance statistics for experiments 4 a and b

	Experiment 5 - Noisy		
	P	**R**	**F**
N	0.675	0.703	0.689
D	0.715	0.762	0.738
I	0.566	0.435	0.492
	Experiment 5 - Clean		
	P	**R**	**F**
N	0.618	0.689	0.652
D	0.610	0.535	0.570
I	0.480	0.493	0.486
	Experiment 5 - Combined		
	P	**R**	**F**
N	0.683	0.665	0.674
D	0.647	0.619	0.633
I	0.515	0.592	0.551

Table 7. Performance statistics for experiment 5

9 Future Work

The work discussed here is a precursor to the rich, and perhaps more computationally difficult task of sense-disambiguation of VPCs. While we have gone some way to computing differences in the semantics of different VPCs, there is a far greater level of sophistication required before all the semantic properties of these anomalous constructs can be computed.

There is also scope for an improvement of the results presented here, through a deeper linguistic analysis of the structure of these VPCs, in particular looking at the features of the arguments.

We have also identified other semantic properties of VPCs (such as the dichotomy of whether the verb and preposition are being used in a literal or metaphoric sense), which in the future, may also form an independent basis for computational classification tasks. Whether these tasks can be performed to any high level of accuracy is left as an open question.

Proceedings of the Australasian Language Technology Workshop 2004
Macquarie University, Sydney, December 8th, 2004.

Accepted after full paper review

10 References

Abeillé, A. 1988. Light verb constructions and extraction out of NP in a tree adjoining grammar. In *Papers of the 24th Regional Meeting of the Chicago Ling. Soc.*

Abeillé, A. 1995. The flexibility of French idioms: A representation with Lexicalised Tree Adjoining Grammar. In M. Everaert, E-J. van der Linden, A. Schenk, and R. Schreuder, editors, *Idioms: Structural and Psychological Perspectives*, chapter 1. Lawrence Erlbaum Associates.

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

Bannard, C., Baldwin, T. and Lascarides, A.. 2003. A Statistical Approach to the Semantics of Verb-Particles. In *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan, pp. 65-72.

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

Barkema, H. 1994. The idiomatic, syntactic and collocational characteristics of received NPs: some basic statistics. *Hermes* 13: 19-40. Aarhus School of Business.

Basili, R., Pazienza, M. and Velardi, P. 1993. Semi-automatic extraction of linguistic information for syntactic disambiguation. *Applied Artificial Intelligence*, 7:339-64.

Blaheta, D. and Johnson, M. 2001. Unsupervised learning of multi-word verbs. In *Proc. of the ACL/EACL 2001 Workshop on the Computational Extraction, Analysis and Exploitation of Collocations*, pp 54-60.

Brill, E. 1995. "Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging", *Computational Linguistics*. 21:543-65.

Calzolari, N., Fillmore, C., Grishman, R., Ide, N., Lencu, A., MacLeod, C. and Zampolli, A. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC 2002), pp 1934-40*

Copestake, A., Lambeau, F., Villavicencio, A., Bond, F., Baldwin, T., Sag, I.A. and Flickinger, D. 2002. Multiword expressions: linguistic precision and reusability. In *Proc. Third conference on Language Resources and Evaluation (LREC-2002)*, pp. 1941—1947. Las Palmas, Canary Islands.

Dixon, R. M. W. 1982. The Grammar of English Phrasal Verbs. *Australian Journal of Linguistics*, 2:149-247.

Kaalep, K. and Muischnek, K. 2002. Using the text corpus to create a comprehensive list of phrasal verbs. In *Proc. of the 3rd International conference on Language Resources and Evaluation (LREC 2002)*, pp 101-5.

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

Krovetz, R. and Croft, W.B. 1989. Word sense disambiguation using machine-readable dictionaries. In *Proc. Annual ACM Conference on Research and Development in Information Retrieval 1989.*

Lin, D. 1999. Automatic Identification of Non-compositional Phrases. In *Proc. 37th Annual Meeting of the ACL*, pp 317-24.

Lindstromberg, S. 1998. English Prepositions Explained. Amsterdam: John Benjamins.

Maynard, D. and Ananiadou, S. 1999. Identifying contextual information for multi-word term extraction. In *5th International Congress on Terminology and Knowledge Engineering (TKE 99)*, pp 212-21.

Maynard, D. and Ananiadou, S. 1998. Acquiring contextual information for term disambiguation. In *Proc. of 1st Workshop Computational Terminology, Computerm '98*, Montreal, Canada.

Meyer, G. A. 1975. The Two-Word Verb, A Dictionary of the Verb-Prepositional Phrases in American English. The Hague: *Mouton.*

Patrick, J., Palko, D., Munro, R., Zappavigna, M. 2003. Inferring semantic structure from format, Computing Arts: Digital Resources in the Humanities, (ed) C. Cole & H. Craig, Sydney: The Uni of Sydney, pp150-168.

McCarthy, D., Keller, B., Carroll, J. 2003, Detecting a Continuum of Compositionality in Phrasal Verbs. In *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan, pp. 65-72.

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

O'Hara, T and Wiebe J. 2003. Preposition Semantic Classification via PENN TREEBANK and FRAMENET. In *Proc. of Computational Natural Language Learning-2003 (CoNLL-03).*

Sag, I.A., Baldwin, T., Bond, F., Copestake, A. and Flickinger, D. 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)*, pp 1-15, Mexico City, Mexico.

Smadja, F. 1993. Retrieving collocations from text: Xtract. In *Computational Linguistics*, 19(1):143-78.

Wehrli, E. 1998. Translating idioms. In *Proc. of the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics: COLING/ACL-98*, pp 1388-92, Montreal, Canada.

Yarowsky, D. 1992. Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. In *Proceedings, COLING-92.* pp 454-460.

Yarowsky, D. 1995. Unsupervised Word Sense Disambiguation Rivalling Supervised Methods. *Proc 33rd Ann. Meet.of the ACL.*

Accepted after full paper review