How to Pick out Token Instances of English Verb-Particle Constructions

Su Nam Kim and Timothy Baldwin
CSSE, University of Melbourne, VIC 3010, Australia

June 20, 2009

Abstract. We propose a method for automatically identifying individual instances of English verb-particle constructions (VPCs) in raw text. Our method employs the RASP parser and analysis of the sentential context of each VPC candidate to differentiate VPCs from simple combinations of a verb and prepositional phrase. We show that our proposed method has an F-score of 0.974 at VPC identification over the Brown Corpus and Wall Street Journal.

Keywords: verb-particle construction, multiword expression, identification

1. Introduction

This paper describes a method for identifying English **verb-particle constructions** (i.e. **VPCs**).[1] VPCs (e.g. *take off* and *battle on*) are a type of multiword expression (hereafter, MWE), that is they are lexical items that are made up of multiple simplex words and display lexical, syntactic, semantic and/or statistical idiosyncrasies (Sag et al., 2002; Calzolari et al., 2002; Bannard, 2003; McCarthy et al., 2003; Widdows and Dorow, 2005; Baldwin and Kim, 2009). As with other MWEs, VPCs present significant problems for natural language processing (hereafter, NLP) in terms of fluency in generation and robustness in parsing.

VPCs are verbal MWEs that are made up of a verb and obligatory particle(s), usually in the form of an intransitive preposition (e.g. *skive off* and *look up*: Dehe, Baldwin and Kim (2002, 2009)). For the purpose of this paper, we follow Baldwin (2005a) in adopting the simplifying assumption that VPCs: (a) consist of a head verb and a unique prepositional particle (e.g. *hand in*, *walk off*); and (b) are either transitive (e.g. *hand (the report) in, put on (a jumper)*) or intransitive (e.g. *battle on*). A defining characteristic of transitive VPCs is that they can generally occur with either joined (e.g. *He put on the sweater*) or split (e.g. *He put the sweater on*) word order. In the case that the object is pronominal, however, the VPC must occur in split word order (c.f. **He handed in it*)

[1] VPCs are found in a number of languages, including English, German and Dutch, but in this paper, we target English VPCs exclusively; VPCs are also commonly termed "phrasal verbs" in the literature.

Kim, Su Nam and Timothy Baldwin (to appear) How to Pick out Token Instances of English Verb-Particle Constructions, Journal of Language Resources and Evaluation.

(Huddleston and Pullum, 2002; Villavicencio, 2003b). The semantics of the VPC can either derive transparently from the semantics of the head verb and particle (e.g. *walk off*) or be significantly removed from the semantics of the head verb and/or particle (e.g. *look up*); analogously, the selectional preferences of VPCs can mirror those of their head verbs or alternatively diverge markedly. The syntax of the VPC can also coincide with that of the head verb (e.g. *walk off*) or alternatively diverge (e.g. *lift off*).

VPCs relate closely to prepositional verbs (Jackendoff, 1973; O'Dowd, 1998; Huddleston and Pullum, 2002; Baldwin, 2005b), which are similarly made up of a verb and preposition, but the preposition is transitive and selected for by the verb (e.g. *refer to*, *look for*). It is possible to differentiate transitive VPCs[2] from prepositional verbs via the variable word order of the particle and object NP with transitive VPCs, as outlined above (Bolinger, 1976; Jackendoff, 1973; Fraser, 1976; Lidner, 1983; O'Dowd, 1998; Dehe et al., 2001; Jackendoff, 2002; Huddleston and Pullum, 2002; Baldwin, 2005b).

The key intuition underlying our proposed method is that in contexts where there is syntactic ambiguity for a given verb–preposition combination, it is possible to resolve the ambiguity via the selectional preferences of the verb vs. the VPC. For example, in the sentence *Kim ran in the room*, the object of the VPC *run in* (in the sense of "drive carefully to avoid damaging a new engine") tend to be MACHINERY whereas the object of *in* as an adjunct of the simple verb *run* will tend to be of type PLACE. *Room* is semantically incompatible with the VPC analysis semantics, suggesting a verb-PP analysis. In contexts where there is a strong lexico-syntactic preference for a VPC analysis (e.g. *look it up*) or verb-PP analysis (e.g. *put it on the table*), on the other hand, syntactic parsers which are attuned to verb subcategorisation and preposition valence are highly adept at predicting the correct analysis. Based on this observation, our method takes the form of post-processing over the output of a probabilistic parser with a symbolic backbone, and attempts to identify and correctly disambiguate instances of syntactic ambiguity based on selectional preferences. The main contribution of this work is to demonstrate the utility of syntactic and semantic features for VPC identification.

In this paper, we exclusively focus on the task of VPC **identification**, that is the detection of individual VPC **token** instances in corpus data (Li et al., 2003). This contrasts with the more widely-researched task of VPC **extraction**, where the objective is to arrive

[2] Prepositional verbs are obligatorily transitive, so there is no ambiguity with intransitive VPCs.

at an inventory of VPC **types**/lexical items based on analysis of token instances in corpus data (Baldwin and Villavicencio, 2002; Baldwin, 2005a). The basic intuition behind the proposed identification method is that the selectional preferences of VPCs over predefined argument positions[3] provide insight into whether a verb and preposition in a given sentential context combine to form a VPC (e.g. *Kim handed in the paper*) or alternatively constitute a verb-PP (e.g. *Kim walked in the room*). That is, we seek to identify individual preposition token instances as intransitive prepositions (i.e. prepositional particles) or transitive prepositions based on analysis of the governing verb.

The remainder of the paper is structured as follows. Section 2 surveys the literature on VPC identification/extraction. Section 3 outlines the basic motivation behind our method, and Section 4 provides a detailed description of how this intuition is applied in our method and the resources used in this research. Section 5 outlines the data sets used in our experimentation, and Section 6 contains detailed evaluation of the proposed method. Section 7 discusses the effectiveness of our approach. Finally, Section 8 summarizes the paper and outlines future work.

2. Related Work

In this section, we survey relevant past research on VPCs, focusing on the extraction/identification of VPCs and the prediction of the compositionality/productivity of VPCs.

For VPC extraction and identification, Baldwin and Villavicencio (2002) proposed a method for extracting VPCs using a POS tagger, chunk parser, full syntactic parser and a combination of all three. The output of the method is a simple list of VPCs, which Baldwin (2005a) extended to propose a method for extracting VPCs with valence information for direct application in a grammar. Baldwin (2005a) followed Villavicencio (2003a) in assuming that VPCs: (a) have a unique prepositional particle, and (b) are either simple transitive or intransitive. Baldwin (2005a) achieved an extraction F-score of 74.9% and 89.7% for intransitive and transitive VPCs, respectively, over the British National Corpus.

Li et al. (2003) performed VPC identification based on hand-crafted regular expressions over the context of occurrence of verb–preposition pairs. The paper reports a performance between 95.8% and 97.5%. Although these results are impressive, the adaptability of the method to new domains and languages is questionable, and the method is not

[3] Focusing exclusively on the subject and object argument positions.

directly applicable to other types of MWEs such as light verb constructions (Grefenstette and Teufel, 1995; Stevenson et al., 2004) or determinerless PPs (Baldwin et al., 2006; van der Beek, 2005).

In Fraser (1976) and Villavicencio (2003b), it is argued that the semantic properties of verbs can determine the likelihood of their occurrence with different particles. Bannard et al. (2003), McCarthy et al. (2003) and Kim and Baldwin (2007) proposed methods for estimating the compositionality of VPCs based largely on distributional similarity and semantic similarity of the head verb and VPC. O'Hara and Wiebe (2003) proposed a method for disambiguating the semantics of prepositions in verb-PPs. Cook and Stevenson (2006) classified the semantics of particles in VPCs using linguistic features. Katz and Giesbrecht (2006) built on the research of Baldwin et al. (2003) in identifying token instances of non-compositional MWEs (particularly verb–noun idioms) in German using Latent Semantic Analysis, and further attempted to measure the compositionality of MWEs. While our interest is in VPC identification—a fundamentally syntactic task—we draw on the style of shallow semantic processing employed in these methods in modeling the semantics of VPCs relative to their base verbs.

3. Selectional Preferences

Divergences in VPC and simplex verb semantics are often reflected in differing selectional preferences, as manifested in patterns of noun co-occurrence. That is, when verbs co-occur with particles to form VPCs, their meaning can be significantly different from the semantics of the head verb in isolation.

(1) and (2) illustrate the difference in the selectional preferences of the verb *put* in isolation as compared with the VPC *put on*.[4]

(1) *put* = "place"
EX: *Put the book on the table.*
ARGS: $book_{\text{OBJ}}$ = "book, publication, object"
ANALYSIS: verb-PP

[4] All sense definitions are derived from WordNet 2.1, based on the first sense of each word; note that all examples are based on corpus examples, but simplified for expository purposes.

(2) *put on* = "wear"
EX: *Put on the coat*.
ARGS: $coat_{\text{OBJ}}$ = "garment, clothing"
ANALYSIS: VPC

Put on is generally used in the context of "wearing" something, with object nouns such as *sweater* and *coat*, whereas *put* in isolation has less sharply defined selectional restrictions and can occur with any noun. In terms of the word senses of the head nouns of the object NPs, the VPC *put on* tends to co-occur with objects which have the semantics of CLOTHING. On the other hand, the simplex verb *put* in isolation tends to be used with a broader range of both concrete and abstract objects, and prepositional phrases containing NPs with the semantics of PLACE.

Also, as observed above, the valence of a VPC can differ from that of the head verb. (3) and (4) illustrate two different senses of *take off* with intransitive and transitive valence, respectively. Note that *take* cannot occur as a simplex intransitive verb.

(3) *take off* = "lift off"
EX: *The airplane takes off.*
ARGS: $airplane_{\text{SUBJ}}$ = "airplane, aeroplane"
ANALYSIS: VPC

(4) *take off* = "remove"
EX: *They take off the cape*.
ARGS: $they_{\text{SUBJ}}$ = "person, individual"
$cape_{\text{OBJ}}$ = "garment, clothing"
ANALYSIS: VPC

In (3), the intransitive *take off* co-occurs with a subject of semantic class AEROPLANE. In (4), on the other hand, the transitive *take off* has an object noun of class CLOTHING.

From the above, we can observe that head nouns in the subject and object argument positions can be used to distinguish VPCs from simplex verbs with prepositional phrases (i.e. verb-PPs).

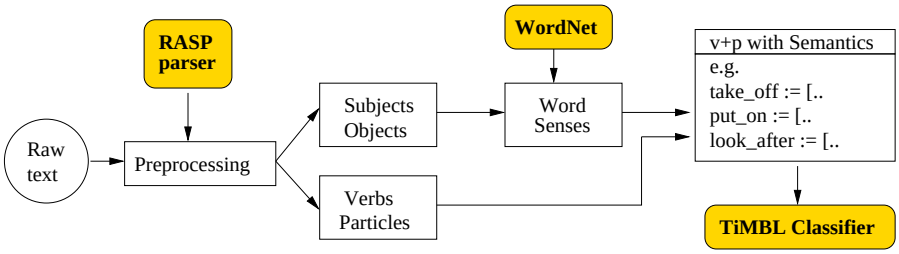

Figure 1. System Architecture

4. Approach, Architecture and Resources

The distinguishing features of our approach are: (i) it tackles the task of VPC identification rather than VPC extraction, and (ii) it uses both syntactic and semantic features, employing the WordNet 2.1 senses of the subject and/or object(s) of the verb. In the sentence *He put the coat on the table*, e.g., to distinguish the VPC *put on* from the verb *put* occurring with the prepositional phrase *on the table*, we identify the senses of the head nouns of the subject and object(s) of the verb *put* (i.e. *he* and *coat*, respectively). That is, VPCs are identified by looking at the semantics of the head nouns of the subject and/or object of a given verb (either VPC or verb in isolation).

Figure 1 depicts the complete process used to distinguish VPCs from verb-PPs.

First, we parse all sentences in a given corpus using the RASP parser (Briscoe and Carroll, 2002), and identify verbs and prepositions in the RASP output. This is a simple process of checking the POS tags in the most-probable parse, and for both particles (tagged `RP`) and transitive prepositions (tagged `II`), reading off the governing verb from the dependency tuple output. We also retrieve the head nouns of the subject and object(s) of each verb directly from the dependency tuples. The RASP output contains dependency tuples derived from the most probable parse, each of which includes a label identifying the nature of the dependency (e.g. `SUBJ` or `DOBJ`), the head word of the modifying constituent, and the head of the modified constituent. Note that we parameterise RASP to output the single best parse for each sentence in grammatical relations format, not to use verb subcategorisation frame probabilities, and not use its in-built list of VPCs. McCarthy et al. (2003) evaluated the precision of RASP at identifying VPCs to be 87.6% and the recall to be 49.4%, based on the gold-standard POS tags in the Wall Street Journal section of the Penn Treebank 2.0

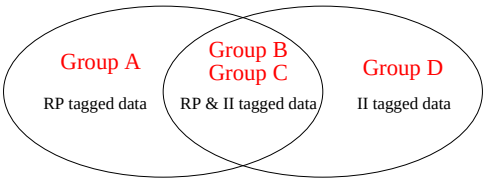

Figure 2. Classification of data in the RASP output

(Marcus et al., 1993). To better understand the baseline performance of RASP, we analysed all false-positive examples tagged with `RP` and false-negative examples tagged with `II`, relative to the gold-standard data in the Penn Treebank. See Section 5.1 for details.

Based on the RASP output, we next obtain the lexical semantics of the head nouns based on WordNet 2.1 (Fellbaum, 1998), using the first sense for that word in SemCor (Landes et al., 1998). The final feature representation for each VPC and verb-PP takes the form of the verb lemma, preposition, and WordNet class of the subject and/or object(s). For the training instances only, we additionally generate separate instances for each of the first- to third-level hypernyms of the first sense.

Having extracted all the features, we then separate it into test and training data, and use TiMBL v5.1 (Daelemans et al., 2004) to learn a classifier.

5. Data Collection

5.1. Data Classification

The evaluation data is made up of sentences containing prepositions tagged as either `RP` or `II`. Based on the output of RASP, the sentences are divided into four groups, as detailed in Figure 2.

Group A contains the verb–preposition token instances tagged exclusively as VPCs (i.e. the preposition is never tagged as `II` in combination with the given head verb). Group B contains the verb–preposition token instances identified as VPCs by RASP where there were also instances of that same combination identified as verb-PPs. Group C contains the verb–preposition token instances identified as verb-PPs by RASP where there were also instances of that same combination identified as VPCs. Finally, group D contains the verb-preposition combinations which were tagged exclusively as verb-PPs by RASP. We focus particularly on disambiguating verb–preposition token instances falling into

Table I. Error rate and inter-annotator agreement for each group

	Group A	Group B	Group C	Group D
False positive rate (FPR)	0.041	0.040	–	–
False negative rate (FNR)	–	–	0.102	0.034
Inter-annotator agreement	0.952	0.996	0.933	0.992

groups B and C, where RASP has identified an ambiguity for that particular combination. We do not further classify token instances in group D, on the grounds that: (a) for high-frequency verb–preposition combinations, RASP was unable to find a single instance warranting a VPC analysis, suggesting it had high confidence in its ability to correctly identify instances of this lexical type; and (b) for low-frequency verb–preposition combinations where the confidence of there definitively not being a VPC usage is low, the token sample is too small to disambiguate effectively and the overall impact would be negligible even if we tried. In addition, during evaluation, we look exclusively at the performance of VPC identification. As a result, we focus particularly on data in groups B and C.

Naturally, the output of RASP is not error-free, i.e. VPCs may be parsed as verb-PPs and vice versa. In particular, other than the results of McCarthy et al. (2003) for identifying VPCs, we had no a priori sense of RASP's ability to distinguish VPCs and verb-PPs. Therefore, we manually checked the false-positive and false-negative rates in all four groups (as defined relative to the gold-standard annotation in the Penn Treebank) and obtained the performance of the parser with respect to VPCs. The verb-PPs in groups A and B are false-positives while the VPCs in groups C and D are false-negatives (we consider the VPCs to be positive examples).

To calculate the number of incorrect examples, two human annotators independently checked each verb–preposition instance.[5] Table I details the rate of false-positives and false-negative examples in each data group, as well as the inter-annotator agreement (calculated over the entire group).

[5] The reason we chose to hand-check the instances rather than simply using the gold-standard POS tags in the original Brown Corpus and Wall Street Journal (which distinguish between particles and transitive prepositions) was that the POS tags were found to be highly unreliable.

Table II. The number of VPC and verb-PP token instances occurring in groups A, B and C at varying frequency cut-offs

Frequency	Type	Group A	Group B	Group C	Total
$f \geq 1$	VPC	5,223	1,312	0	6,535
	V-PP	0	0	995	995
$f \geq 5$	VPC	3,787	1,108	0	4,895
	V-PP	0	0	217	217

Table III. Breakdown of subject and object head nouns in groups A&B, and group C (NN = noun, P-PRN = personal pronoun, and D-PRN = demonstrative pronoun)

Group	Common NN	P-PRN	D-PRN	Proper NN	*who*	*which*	*what*
A&B	7,116	629	127	156	94	32	11
C	1,239	79	1	18	6	0	0

5.2. Collection

We combined together the 6,535 (putative) VPCs and 995 (putative) verb-PPs from groups A, B and C, as identified by RASP over the corpus data. Table II shows the number of VPC tokens in groups A and B, and the number of verb-PPs in group C. $f \geq 1$ is the number of tokens which occur at least once, and $f \geq 5$ is the number of tokens which occur five or more times. Note that the number of (ambiguous) verb-PP tokens which occur repeatedly (in group C) is much less than that of VPCs (in groups A and B).

From the sentences containing VPCs and verb-PPs, we retrieved a total of 8,165 nouns in the subject and/or object positions—including pronouns (e.g. *I, he, she*), proper nouns (e.g. *CITI, Canada, Ford*) and demonstrative pronouns (e.g. *one, some, this*)—which occurred as the head noun of a subject or object of a VPC in group A or B. We similarly retrieved 1,343 nouns for verb-PPs in group C. Table III shows the distribution of different noun tokens across these two sets.

We found that about 10% of the nouns are pronouns (P-PRN or D-PRN), proper nouns or WH words (*who*, *which* or *what*). In evaluation, we test three strategies for dealing with pronouns, proper nouns and WH words: (1) pronouns are manually resolved to the WordNet class of

their antecedents and proper nouns are replaced by their hypernyms; (2) all pronouns and proper nouns are left unresolved; and (3) only proper nouns are replaced by their hypernyms.

For pronouns, we manually resolved the antecedent and took this as the head noun. When *which* is used as a relative pronoun, we identified if it was co-indexed with an argument position of a VPC or verb-PP, and if so, manually identified the antecedent, as illustrated in (5).

(5) **EX:** *Tom likes the* [*books*] *which* [*he*] *sold off.*
ARGS: he_{SUBJ} = "person"
$which_{\text{OBJ}}$ = "book"

With *what*, on the other hand, we were generally not able to identify an antecedent, in which case the argument position was left without a word sense (for detailed discussion, see Section 7).

(6) *Tom didn't look up* [*what*] *to do.*

(7) [*What*] *went on?*

For proper nouns, we identified their common noun hypernym based on manual disambiguation, as the coverage of proper nouns in WordNet is (intentionally) poor. Examples of proper nouns and their common noun hypernyms are: *CITI* → BANK, *Canada* → COUNTRY, and *Smith* → HUMAN.

We generate a unique instance for each VPC and verb-PP token instance. We additionally identify hypernyms (up to) three levels up the WordNet hierarchy from the first sense of each noun argument.[6] This is intended as a crude form of smoothing for closely-related word senses which occur in the same basic region of the WordNet hierarchy, and enable the determination of suitable selectional preference classes in WordNet.

Finally, we randomly selected 80% of the instances to use as training data and the remaining 20% as test data based on parser output. The total number of training instances, before and after performing hypernym expansion using WordNet, is indicated in Table IV.

[6] The choice of 3 levels was made empirically.

Table IV. The number of training instances

Training Instances	Group A	Group B	Group C
Before expansion	5,223	1,312	995
After expansion	24,602	4,158	5,985

6. Evaluation

We separately evaluated the three different strategies for resolving pronouns and proper nouns (full manual resolution, no manual resolution, and manual resolution for proper nouns only). Note that our focus is exclusively on VPC identification, and hence we do not present explicit results for verb-PP token identification.

Due to the differing amounts of data in A, B and C, we experiment with four different combinations of data from each, based on differing frequency thresholds over the training data. In the first two datasets, we include only instances from groups B and C (i.e. token instances of types with both VPC and V-PP instances), including all VPC instances from C (i.e. a frequency threshold of $f \geq 1$), and either all V-PP instances ($f \geq 1$) or only V-PP instances with a token frequency of 5 or greater ($f \geq 5$) from B. In the next two datasets, we additionally include unambiguous VPCs from group A to boost the number of positive training instances, either taking all VPC instances ($f \geq 1$) or only those instances with a token frequency of 5 or greater ($f \geq 5$). The reason we always use all V-PP token instances ($f \geq 1$) from C is that the V-PPs tend to have low token frequencies in this set. Note that in all cases, we include all test instances, irrespective of frequency, such that the precision, recall and F-score under the different experimental settings are directly comparable.

As our baseline for VPC identification, we use the raw output of RASP.

6.1. Experiment with Fully Resolved Noun Semantics

Table V shows the results of our method over the Brown Corpus and Wall Street Journal using manually-resolved pronouns and proper nouns, in terms of VPC identification. As mentioned above, we evaluate relative to different combinations of data from A, B, and C, with different thresholds. The performance of RASP in identifying VPCs is calcu-

Table V. VPC identification results with fully resolved pronouns and proper nouns

Data	Frequency	Precision	Recall	F-score
RASP	—	.959	.955	.957
B+C	$f \geq 1$	.948	.958	.952
	$f \geq 5$	.955	.979	.966
A+B+C	$f \geq 1$	.962	.962	.962
	$f \geq 5$	.964	.983	**.974**

lated based on human judgement over all token instances in groups B and C. When RASP identifies a verb and particle correctly, we consider it to have identified the VPC correctly irrespective of whether the argument structure is correct or not. Also, we ignore ambiguity between particles and adverbs (e.g. *hand* <u>*out*</u> vs. *walk* <u>*out*</u>), leading to higher performance than that reported in McCarthy et al. (2003).

Table V shows that the performance over high-frequency data from groups A, B and C is the highest (F-score = 0.974). As a general trend, the best results are achieved over the high-frequency VPCs, including data from A. Encouragingly, we achieve a slightly higher result than the 0.958–0.975 claimed by Li et al. (2003) with relatively little manual intervention (to resolve the semantic class of pronouns and proper nouns).

6.2. Experiment without resolving pronouns or proper nouns

We next repeat the experiment using the same data set as above but without manual resolution of the antecedents of pronouns and proper nouns. Here, every pronoun and proper noun (and common noun not found in WordNet) is represented not as a synset but as a coarse-grained feature describing the noun type (common noun, pronoun, or proper noun). Common nouns are automatically assigned WordNet synsets as before, whereas pronouns and proper nouns are sub-classified into the human and non-human classes. All of these features are automatically derived, and based on POS tags and dictionaries.

Our interest in this experiment is to determine the relative drop when we take away the rich ontological semantics we manually annotated in the first experiment.

Table VI. VPC identification results without resolving pronouns or proper nouns

Data	Frequency	Precision	Recall	F-score
RASP	—	.959	.955	.957
B+C	$f \geq 1$	.936	.958	.946
	$f \geq 5$	.940	.956	.948
A+B+C	$f \geq 1$	.949	.969	**.959**
	$f \geq 5$	.951	.966	.958

Table VI shows the results without manually resolving pronouns and proper nouns. Due to the relative sparsity of semantic information, the performance of this method is below that of the manually-resolved nouns in our first experiment, but it still achieved a slightly better F-score than the RASP parser (an F-score of 0.959 vs. 0.957).

6.3. Experiment with partially resolved proper nouns

Our third experiment is identical to the previous two experiments except that proper nouns are (partially) resolved using WordNet, in that if a proper noun is found in WordNet it is resolved in an identical manner to common nouns, and if not we fall back to the human vs. non-human binary distinction from our second experiment. As such, this experiment still requires no manual effort to resolve the semantics of head nouns, but lacks semantics for pronouns and proper nouns which do not occur in WordNet.

Our expectation is that despite WordNet having poor coverage of proper nouns, we will still manage to retrieve word senses of many commonly-occurring proper nouns automatically. Note that around 28% of the proper noun token instances in our data were found in WordNet.

Table VII describes the performance of our method with partially-resolved semantics for proper nouns. The F-score is almost identical to that for unresolved semantics (Experiment 2), suggesting that the primary gain in performance in Experiment 1 was for pronouns rather than proper nouns.

Table VII. VPC identification results with partially resolved proper nouns

Data	Frequency	Precision	Recall	F-score
RASP	—	.959	.955	.957
B+C	$f \geq 1$	.938	.960	.948
	$f \geq 5$	.938	.957	.947
A+B+C	$f \geq 1$	.951	.967	**.959**
	$f \geq 5$	.951	.966	.958

Table VIII. VPC identification results with hypernym expansion (4WS) vs. simple senses (1WS)

Frequency	WSD	Precision	Recall	F-score
$f \geq 1$	4WS	.962	.962	.962
	1WS	.958	.969	.963
$f \geq 5$	4WS	.964	.983	**.974**
	1WS	.950	.973	.962

6.4. Experiment with and without hypernym expansion

Finally, we compare the performance of the proposed method with manual sense resolution and hypernym expansion (4WS), to that with manual sense resolution but without hypernym expansion (1WS). Note that for all experiments reported so far, we have used hypernym expansion, and as such, the numbers for hypernym expansion are identical to those from Table V. The results, presented in Table VIII, suggest that using hypernyms improves performance over frequent verb–preposition combinations.

7. Analysis and Discussion

We proposed an automatic method to identify English VPCs based on the selectional preferences of different argument positions. We experimented with three different strategies for resolving the semantics of pronouns and proper nouns, and found that while an oracle coreference resolution and proper noun interpretation system improved perfor-

Table IX. Performance of different parsers at VPC identification over the Brown Corpus and Wall Street Journal

Parser	Precision	Recall	F-score
RASP	.959	.955	.957
fnTBL	.703	.632	.668
Charniak	.659	.694	.676
minipar	.364	.429	.397

mance slightly, the relative increment over a fully-automated method with partial coverage is slight. Overall, our method exceeded the performance reported in Li et al. (2003) and the RASP baseline. This suggests both that selectional preferences can boost the performance of VPC identification, and that it is possible to capture selectional preferences in a supervised learning framework with no or little manual intervention.

We also proposed a naive method for modelling noun semantics which doesn't rely on a word sense disambiguation system or hand tagging. The method proved superior to simple lexical probabilities (as are used by RASP), and gained from semantic smoothing via three levels of hypernyms. Since McCarthy et al. (2004) found that 54% of word tokens are used with their first (or default) sense and the performance of supervised word sense disambiguation (WSD) systems are hovering around 60-70%, a simple first-sense WSD system has room for improvement, but is sufficient to acquire the word senses of nouns without manual word sense disambiguation.

Our method takes the form of a postprocessing step after parsing, with all experiments based on the RASP parser. Clearly the performance of the post-processing is predicated on the quality of the parser output, as we rely on the parser to identify the argument structure and head nouns. To evaluate the relative performance of RASP at VPC identification relative to other existing parsers, we evaluated a full text chunk parser based on fnTBL (Ngai and Florian, 2001), the Charniak treebank parser (Charniak, 2000) and minipar (Lin, 1993) over the same task. Note that we did not retrain any of the parsers, just as we did not retrain RASP in our original experiments.

Table IX shows the VPC identification performance of the three parsers, relative to the performance for RASP. RASP outperformed

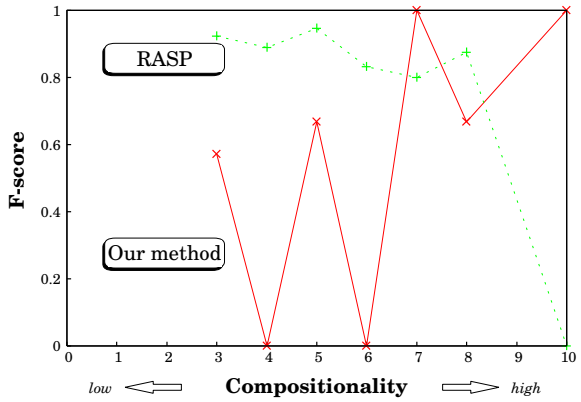

Figure 3. The relationship between VPC compositionality and VPC identification F-score

all three parsers, suggesting that it was a well-chosen parser for the task at hand.

To check the correlation between the compositionality of each VPC and our ability to identify its token instances, we took 117 VPCs of varying semantic compositionality and analysed the relative ability of our method to identify the token instances of each. For this, we used the data set of McCarthy et al. (2003), which provides compositionality judgements for VPC types based on three annotators, on a scale of 0 to 10 (0 = non-compositional, 10 = fully compositional).

Figure 3 is a graph of the F-score for both RASP and our method at different levels of compositionality, for those VPCs in our data set which also occur in the data of McCarthy et al. (2003). From the graph, we see that we our method actually depletes the VPC identification F-score over low-compositionality VPCs, but greatly increases performance over high-compositionality VPCs, with the combined effect being a modest increase in F-score. The reason for this is that low-compositionality VPCs (e.g. *drag on*) are often easy for parsers to identify, as their subcategorisation properties diverge from the simplex verb or there is no corresponding simplex verb at all (c.f. *chicken out*).[7] Here, RASP performs predictably well. High-compositionality VPCs (e.g. *call in*), on the other hand, tend to be less easy to distinguish from V-PPs based on syntax alone, and the semantic modelling underlying our method comes to the fore. The plummets in F-score (to zero!) for

[7] Note that no compositionality 0–2 instances were observed in our data to be able to track this trend to the level of full non-compositionality.

our method over VPCs with compositionality 4 and 6 are a slight cause for concern. In practice, the numbers of VPCs at stake are tiny (3 and 6 tokens, respectively), so the effect is largely due to data sparseness. In the latter case (compositionality = 6), all token instances correspond to the single VPC type of *come over*, where *over* wasn't commonly observed as a particle elsewhere in the data, causing the classifier to misclassify all instances.

Given a reliable method for predicting the compositionality of a given verb–preposition combination, we could consider evoking our method only for high-compositionality VPCs, and more effectively hybridising our method with the raw RASP outputs. Current research on compositionality prediction, however, is far from reliable (McCarthy et al., 2003; Baldwin et al., 2003), making this an unrealistic expectation at present.

Clearly there are more possibilities for exploiting semantic features than what we have explored in this paper. As future research, we are particularly interested in including distributional similarity and semantic features for other argument types, as well as evaluating the proposed method over a broader set of constructions (including non-MWEs). Additionally, it would be interesting to conduct experiments over different domains (including the WSJ as a standalone corpus) to determine the impact of domain on our results.

From our method, we found several factors that require further study. In manually analysing the data, some data instances were missing explicit head nouns, leading to nouns without word senses. If only a small number of token instances is available, missing word senses could influence the performance of the method since the classifier relies on training data to disambiguate VPCs against verb-PPs. Particular instances of missing nouns are imperative and abbreviated sentences such as the following:

(8) *Come in.*

(9) *(How is your cold?) Broiled out.*

Another factor is the lack of word sense data, particularly in WH questions, where it is often non-trivial to identify the antecedent noun to specify the noun semantics:

(10) *What do I hand in?*

(11) *You can add up anything.*

Also, the method is clearly dependent on the base performance of RASP, and any improvement in the base parser has the potential

to improve our method (or in the extreme case, make our method redundant!). We observed that among the false positive VPCs, there were occurrences of the particle occurring before the verb. An example of this is with the sentence:

(12) Help me up, I feel kind of stiff.

from which RASP identified the VPC *feel up*. Linguistically speaking, the particle must always appear after the verb (except with non-selected adverbial uses of prepositions such as *up he got*), a constraint which could be built into RASP.

Another example of the particle being attached to the wrong verb, e.g. in the following sentence:

(13) Lucy drew out the chair and sat down.

RASP identified the VPC *draw down*. One again here, a constraint on the degree of separation between the verb and its particle (similarly to Baldwin (2005a)) could prevent such misanalyses.

A common cause of false negatives was copular sentences with particles such as:

(14) Power is back on.

The particle usage here is unproblematic, but the classifier was unable to predict that VPCs must incorporate non-copular verbs. It would be relatively easy to filter these out through the addition of extra features or post-processing over the verb lemma.

8. Conclusions

In this paper, we proposed a method to identify VPCs automatically from raw text data. We first used the RASP parser to identify verb-preposition token instances as possible VPCs or verb-PPs. Then, we extracted the argument structure for each verb and derived the word senses of the subject and/or object head nouns. Finally, we built a supervised classifier using TiMBL v5.1 to relabel false positive VPCs as verb-PPs and vice versa. Over a small data set extracted from the Brown Corpus and Wall Street Journal, our classifier achieved an F-score of 0.974 for the task of VPC identification. We also tested the proposed method over various representations of noun semantics, and showed that automatic methods can near the performance of methods which assume full coreference resolution and proper noun interpretation. Finally, we demonstrated a direct correlation between the degree

of compositionality and the ability of our method to correctly identify VPCs.

The main advantage of our method is that it is fully automated and makes active use of existing resources. We suggest that our proposed approach is a reliable, stable method for automatic VPC identification.

Acknowledgements

This research was carried out in part with support from Australian Research Council grant no. DP0663879.

References

Baldwin, T.: 2005a, 'The Deep Lexical Acquisition of English Verb-particles'. *Computer Speech and Language, Special Issue on Multiword Expressions* **19**(4), 398–414.

Baldwin, T.: 2005b, 'Looking for Prepositional Verbs in Corpus Data'. In: *Proceedings of the 2nd ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*. Colchester, UK, pp. 115–126.

Baldwin, T., C. Bannard, T. Tanaka, and D. Widdows: 2003, 'An Empirical Model of Multiword Expression Decomposability'. In: *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*. Sapporo, Japan, pp. 89–96.

Baldwin, T., J. Beavers, L. Van Der Beek, F. Bond, D. Flickinger, and I. A. Sag: 2006, 'In Search of a Systematic Treatment of Determinerless PPs'. In: P. Saint-Dizier (ed.): *Syntax and Semantics of Prepositions*. Springer.

Baldwin, T. and S. N. Kim: 2009, 'Multiword Expressions'. In: N. Indurkhya and F. J. Damerau (eds.): *Handbook of Natural Language Processing*. Boca Raton, USA: CRC Press, 2nd edition.

Baldwin, T. and A. Villavicencio: 2002, 'Extracting the unextractable: A case study on verb-particles'. In: *Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002)*. Taipei, Taiwan, pp. 98–104.

Bannard, C.: 2003, 'Statistical Techniques for Automatically Inferring the Semantics of Verb-Particle Constructions'. Master's thesis, University of Edinburgh.

Bannard, C., T. Baldwin, and A. Lascarides: 2003, 'A Statistical Approach to the Semantics of Verb-Particles'. In: *Proceedings of the ACL2003 Workshop on Multiword Expressions: analysis, acquisition and treatment*. Sapporo, Japan, pp. 65–72.

Bolinger, D.: 1976, *The Phrasal Verb in English*. Boston, USA: Harvard University Press.

Briscoe, T. and J. Carroll: 2002, 'Accurate Statistical Annotation of General Text'. In: *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*. Las Palmas, Canary Islands, pp. 1499–1504.

Calzolari, N., C. Fillmore, R. Grishman, N. Ide, A. Lenci, C. MacLeod, and A. Zampolli: 2002, 'Towards Best Practice for Multiword Expressions in Computational Lexicons'. In: *Proceedings of the 3rd International Conference on*

Language Resources and Evaluation (LREC 2002). Las Palmas, Canary Islands, pp. 1934–1940.

Charniak, E.: 2000, 'A Maximum Entropy-Based Parser'. In: *Proceedings of the 1st Annual Meeting of the North American Chapter of Association for Computational Linguistics*. Seattle, USA, pp. 132–139.

Cook, P. and S. Stevenson: 2006, 'Classifying Particle Semantics in English Verb-Particle Constructions'. In: *Proceedings of the ACL-2006 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*. Sydney, Australia, pp. 45–53.

Daelemans, W., J. Zavrel, K. van der Sloot, and A. van den Bosch: 2004, 'TiMBL: Tilburg Memory Based Learner, version 5.1, Reference Guide'.

Dehe, N.: 2002, *Particle Verbs in English: Syntax, Information Structure and Intonation*. Amsterdam, Netherlands/Philadelphia, USA: John Benjamins Publishing.

Dehe, N., R. Jackendoff, A. McIntyre, and S. Urban (eds.): 2001, *Verb-Particle Explorations*. Berlin, Germany / New York, USA: Mounton de Gruyter.

Fellbaum, C. (ed.): 1998, *WordNet, An Electronic Lexical Database*. Cambridge, Massachusetts, USA: MIT Press.

Fraser, B.: 1976, *The Verb-Particle Combination in English*. The Hague: Mouton.

Grefenstette, G. and S. Teufel: 1995, 'A Corpus-based Method for Automatic Identification of Support Verbs for Nominalizations'. In: *Proceedings of the 7th European Chapter of Association of Computational Linguistics (EACL-1995)*. Dublin, Ireland, pp. 98–103.

Huddleston, R. and G. K. Pullum: 2002, *The Cambridge Grammar of the English Language*. Cambridge, UK: Cambridge University Press.

Jackendoff, R.: 1973, 'The base rules for prepositional phrases'. In: *A Festschrift for Morris Halle*. New York, USA: Rinehart and Winston, pp. 345–356.

Jackendoff, R.: 2002, *Foundations of Language*. Oxford, UK: Oxford University Press.

Katz, G. and E. Giesbrecht: 2006, 'Automatic Identification of Non-Compositional Multi-Word Expressions using Latent Semantic Analysis'. In: *Proceedings of the ACL-2006 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*. Sydney, Australia, pp. 28–35.

Kim, S. N. and T. Baldwin: 2007, 'Detecting Compositionality of English Verb-Particle Constructions using Semantic Similarity'. In: *Proceedings of Conference of the Pacific Association for Computational Linguistics*. Melbourne, Australia, pp. 40–48.

Landes, S., C. Leacock, and R. I. Tengi: 1998, 'Building Semantic Concordances'. In: C. Fellbaum (ed.): *WordNet: An Electronic Lexical Database*. Cambridge, USA: MIT Press.

Li, W., X. Zhang, C. Niu, Y. Jiang, and R. K. Srihari: 2003, 'An Expert Lexicon Approach to Identifying English Phrasal Verbs'. In: *Proceedings of the ACL2003 Workshop on Multiword Expressions: analysis, acquisition and treatment*. Sapporo, Japan, pp. 513–520.

Lidner, S.: 1983, 'A lexico-semantic analysis of English verb particle constructions with OUT and UP'. Ph.D. thesis, University of Indiana at Bloomington.

Lin, D.: 1993, 'Principle-based Parsing without Overgeneration'. In: *Proceedings of the 31th Association of Computational Linguistics (ACL-1993)*. Columbus, Ohio, USA, pp. 112–120.

Marcus, M. P., B. Santorini, and M. A. Marcinkiewicz: 1993, 'Building a Large Annotated Corpus of English: the Penn Treebank'. *Computational Linguistics* **19**(2), 313–330.

McCarthy, D., B. Keller, and J. Carroll: 2003, 'Detecting a Continuum of Compositionality in Phrasal Verbs'. In: *Proceedings of the ACL2003 Workshop on Multiword Expressions: analysis, acquisition and treatment*. Sapporo, Japan, pp. 73–80.

McCarthy, D., R. Koeling, J. Weeds, and J. Carroll: 2004, 'Finding Predominant Senses in Untagged Text'. In: *Proceedings of the 42nd Annual Meeting of the Association of Computational Linguistics*. Barcelona, Spain, pp. 280–287.

Ngai, G. and R. Florian: 2001, 'Transformation-based Learning in the Fast Lane'. In: *Proceedings of the 2nd Annual Meeting of the North American Chapter of Association for Computational Linguistics (NAACL)*. Pittsburgh, USA, pp. 40–47.

O'Dowd, E. M.: 1998, *Prepositions and Particles in English*. Oxford University Press.

O'Hara, T. and J. Wiebe: 2003, 'Preposition Semantic Classification via Treebank and FrameNet'. In: *Proceedings of the 7th Conference on Natural Language Learning*. Edmonton, Canada, pp. 79–86.

Sag, I. A., T. Baldwin, F. Bond, A. Copestake, and D. Flickinger: 2002, 'Multiword Expressions: A Pain in the Neck for NLP'. In: *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*. Mexico City, Mexico, pp. 1–15.

Stevenson, S., A. Fazly, and R. North: 2004, 'Statistical Measures of the Semi-productivity of Light Verb Constructions'. In: *Proceedings of the 2nd ACL Workshop on Multiword Expressions: Integrating Processing*. Barcelona, Spain, pp. 1–8.

van der Beek, L.: 2005, 'The Extraction of Determinerless PPs'. In: *Proceedings of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*. Colchester, UK, pp. 190–199.

Villavicencio, A.: 2003a, 'Verb-Particle Constructions and Lexical Resources'. In: *Proceedings of the ACL2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*. Sapporo, Japan, pp. 57–64.

Villavicencio, A.: 2003b, 'Verb-Particle constructions in the World Wide Web'. In: *Proceedings of the ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their use in Computational Linguistics Formalisms and Applications*. Toulouse, France.

Widdows, D. and B. Dorow: 2005, 'Automatic Extraction of Idioms using Graph Analysis and Asymmetric Lexicosyntactic Patterns'. In: *Proceedings of ACL2005 Workshop on Deep Lexical Axquisition*. Ann Arbor, Michigan, USA, pp. 48–56.