Picking them up and Figuring them out: Verb-Particle Constructions, Noise and Idiomaticity

Carlos Ramisch[♣♢], Aline Villavicencio[♣♠], Leonardo Moura[♣] and Marco Idiart[♡]
[♣]Institute of Informatics, Federal University of Rio Grande do Sul (Brazil)
[♢]GETALP Laboratory, Joseph Fourier University - Grenoble INP (France)
[♠]Department of Computer Sciences, Bath University (UK)
[♡]Institute of Physics, Federal University of Rio Grande do Sul (Brazil)
{ceramisch,avillavicencio,lfsmoura}@inf.ufrgs.br, idiart@if.ufrgs.br

Abstract

This paper investigates, in a first stage, some methods for the automatic acquisition of verb-particle constructions (VPCs) taking into account their statistical properties and some regular patterns found in productive combinations of verbs and particles. Given the limited coverage provided by lexical resources, such as dictionaries, and the constantly growing number of VPCs, possible ways of automatically identifying them are crucial for any NLP task that requires some degree of semantic interpretation. In a second stage we also study whether the combination of statistical and linguistic properties can provide some indication of the degree of idiomaticity of a given VPC. The results obtained show that such combination can successfully be used to detect VPCs and distinguish idiomatic from compositional cases.

1 Introduction

Considerable investigative effort has focused on the automatic identification of Multiword Expressions (MWEs), like compound nouns (*science fiction*) and phrasal verbs (*carry out*) (e.g. Pearce (2002), Evert and Krenn (2005) and Zhang et al. (2006)). Some of them employ language and/or type dependent linguistic knowledge for the task, while others employ independent statistical methods, such as Mutual Information and Log-likelihood (e.g. Pearce (2002) and, Zhang et al. (2006)), or even a combination of them (e.g. Baldwin (2005) and Sharoff (2004)), as basis for helping to determine whether a given sequence of words is in fact an MWE. Although some research aims at developing methods for dealing with MWEs in general (e.g. Zhang et al. (2006), Ramisch et al. (2008)), there is also some work that deals with specific types of MWEs (e.g. Pearce (2002) on collocations and Villavicencio (2005) on verb-particle constructions (VPCs)) as each of these MWE types has distinct distributional and linguistic characteristics.

VPCs are combinations of verbs and particles, such as *take off* in *Our plane took off late*, that due to their complex characteristics and flexible nature, provide a real challenge for NLP. In particular, there is a lack of adequate resources to identify and treat them, and those that are available provide only limited coverage, in face of the huge number of combinations in use. For tasks like parsing and generation, it is essential to know whether a given VPC is possible or not, to avoid for example using combinations that sound unnatural or ungrammatical to native speakers (e.g. *give/lend/?grant out* for *the conveying of something to someone or some place* - (Fraser, 1976)).[1] Thus, the knowledge of which combinations are possible is crucial for precision grammar engineering. In addition, as the semantics of VPCs varies from the idiomatic to the more compositional cases, methods for the automatic detection and handling of idiomaticity are very important for any NLP task that involves some degree of semantic interpretation such as Machine Translation (in this case avoiding the problem of producing an unrelated translation for a source sentence). Automatic methods for the identification of idiomaticity in MWEs have been

[1] See Baldwin et al. (2004) for a discussion of the effects of multiword expressions like VPCs on a parser's performance.

CoNLL 2008: Proceedings of the 12th Conference on Computational Natural Language Learning, pages 49–56
Manchester, August 2008

proposed using a variety of approaches such as statistical, substitutional, distributional, etc. (e.g. McCarthy et al. (2003), Bannard (2005) and Fazly and Stevenson (2006)). In particular, Fazly and Stevenson (2006) look at the correlation between syntactic fixedness (in terms of e.g. passivisation, choice of determiner type and pluralisation) and non-compositionality of verb-noun compounds such as *shoot the breeze*.

In this work we investigate the automatic extraction of VPCs, looking into a variety of methods, combining linguistic with statistical information, ranging from frequencies to association measures: Mutual Information (MI), χ^2 and Entropy. We also investigate the determination of compositionality of VPCs verifying whether the degree of semantic flexibility of a VPC combined with some statistical information can be used to determine if it is idiomatic or compositional.

This paper starts with a brief description of VPCs, research on their automatic identification and determination of their semantics (§ 2). We then explain the research questions and the assumptions that serve as the basis for the application of statistical measures (§ 3) on the dataset (§ 4). Our methods and experiments are then detailed (§ 5), and the results obtained are analysed (§ 6). We conclude with a discussion of the contributions that this work brings to the research on verb-particle constructions (§ 7).

2 Verb-Particle Constructions in Theory and Practice

Particles in VPCs are characterised by containing features of motion-through-location and of completion or result in their core meaning (Bolinger, 1971). VPCs can range from idiosyncratic or semi-idiosyncratic combinations, such as *get on* (in e.g. *Bill got on well with his new colleagues*), to more regular ones, such as *tear up* (e.g. in *In a rage she tore up the letter Jack gave her*). A three way classification is adopted by (Dehé, 2002) and (Jackendoff, 2002), where a VPC can be classified as compositional, idiomatic or aspectual, depending on its sense. In compositional VPCs the meaning of the construction is determined by the literal interpretations of the particle and the verb. These VPCs usually involve particles with directional or spatial meaning, and these can often be replaced by the appropriate directional PPs (e.g. *carry in* in *Sheila carried the bags in/into the house* Dehé (2002)). Idiomatic VPCs, on the other hand, cannot have their meaning determined by interpreting their components literally (e.g. *get on*, meaning *to be on friendly terms with someone*). The third class is that of aspectual VPCs, which have the particle providing the verb with an endpoint, suggesting that the action described by the verb is performed completely, thoroughly or continuously (e.g. *tear up* meaning *to tear something into a lot of small pieces*).

From a syntactic point of view, a given combination can occur in several different subcategorisation frames. For example, *give up* can occur as an intransitive VPC (e.g. in *I give up! Tell me the answer*), where no other complement is required, or it may occur as a transitive VPC which requires a further NP complement (e.g. in *She gave up alcohol while she was pregnant*). Since in English particles tend to be homographs with prepositions (*up, out, in*), a verb followed by a preposition/particle and an NP can be ambiguous between a transitive VPC and a prepositional verb (e.g. *rely on*, in *He relies on his wife for everything*). Some criteria that characterise VPCs are discussed by Bolinger (1971):[2]

C1 In a transitive VPC the particle may come either before or after the NP (e.g. *He backed up the team* vs. *He backed the team up*). However, whether a particle can be separated or not from the verb may depend on the degree of bonding between them, the size of the NP, and the kind of NP. This is considered by many to be sufficient condition for diagnosing a VPC, as prepositions can only appear in a position contiguous to the verb (e.g. **He got the bus off*).

C2 Unstressed personal pronouns must precede the particle (e.g. *They ate it up* but not **They ate up it*).

C3 If the particle precedes a simple definite NP, the particle does not take the NP as its object (e.g. in *He brought **along** his girlfriend*) unlike with PP complements or modifiers (e.g. in *He slept **in** the hotel*). This means that in the first example the NP is not a complement of the particle *along*, while in the second it is.

[2]The distinction between a VPC and a prepositional verb may be quite subtle, and as pointed out by Bolinger, many of the criteria proposed for diagnosing VPCs give different results for the same combination, frequently including unwanted combinations and excluding genuine VPCs.

In this paper we use the first two criteria, therefore the candidates may contain noise (in the form of prepositional verbs and related constructions).

VPCs have been the subject of a considerable amount of interest, and some analysis has been done on the subject of productive VPCs. In many cases the particle seems to be compositionally adding a specific meaning to the construction and following a productive pattern (e.g. in *tear up*, *cut up* and *split up*, where the verbs are semantically related and *up* adds a sense of completion to the action of these verbs). Fraser (1976) points out that semantic properties of verbs can affect their ability to combine with particles: for example, *bolt/cement/clamp/glue/paste/nail* are semantically similar verbs where the objects represented by the verbs are used to join material, and they can all combine with *down*. There is clearly a common semantic thread running through this list, so that a new verb that is semantically similar to them can also be reasonably assumed to combine with *down*. Indeed, frequently new VPCs are formed by analogy with existing ones, where often the verb is varied and the particle remains (e.g. *hang on, hold on* and *wait on*). Similarly, particles from a given semantic class can be replaced by other particles from the same class in compositional combinations: *send up/in/back/away* (Wurmbrand, 2000). By identifying classes of verbs that follow patterns such as these in VPCs, we can help in the identification of a new unknown candidate combination, using the degree of productivity of a class to which the verb belongs as a back-off strategy.

In terms of methods for automatic identification of VPCs from corpora, Baldwin (2005) proposes the extraction of VPCs with valence information from raw text, exploring a range of techniques (using (a) a POS tagger, (b) a chunker, (c) a chunk grammar, (d) a dependency parser, and (e) a combination of all methods). Villavicencio (2005) uses the Web as a corpus and productive patterns of combination to generate and validate candidate VPCs. The identification of compositionality in VPCs is addressed by McCarthy et al. (2003) who examine the overlap of similar words in an automatically acquired distributional thesaurus for verb and VPCs, and by Bannard (2005) who uses a distributional approach to determine when and to what extent the components of a VPC contribute their simplex meanings to the interpretation of the VPC. Both report a correlation between some of the measures and compositionality judgements.

3 The Underlying Hypotheses

The problem of the automatic detection and classification of VPCs can be summarised as, for a given VPC candidate, to answer to the questions:

Q1 Is it a real VPC or some free combination of verb and preposition/adverb or a prepositional verb?

Q2 If it is a true VPC, is it idiomatic or compositional?

In order to answer the first question, we use two assumptions. Firstly, we consider that the elements of a true VPC co-occur above chance. The greater the correlation between the verb and the particle the greater the chance that the candidate is a true VPC. Secondly, based on criterion C1 we also assume that VPCs have more flexible syntax and are more productive than non-VPCs. This second assumption goes against what is usually adopted for general MWEs, since it is the prepositional verbs that allow less syntactic configurations than VPCs and are therefore more rigid (§ 2). To further distinguish VPCs from prepositional verbs and other related constructions we also verify the possibility of the particle to be immediately followed by an indirect prepositional complement (like in *The plane took off from London*), which is a good indicator/delimiter of a VPC since in non-VPC constructions like prepositional verbs the preposition needs to have an NP complement. Therefore, we will assume that a true VPC occurs in the following configurations, according to Villavicencio (2005) and Ramisch et al. (2008):

S1 VERB + PARTICLE + DELIMITER, for intransitive VPCs;

S2 VERB + NP + PARTICLE + DELIMITER, for transitive split VPCs and;

S3 VERB + PARTICLE + NP + DELIMITER, for transitive joint VPCs.

In order to answer Q2, we look at the link between productivity and compositionality and assume that a compositional VPC accepts the substitution of one of its members by a semantically related term. This is in accordance to Fraser (1976), who shows that semantic properties of

verbs can affect their ability to combine with particles: for example verbs of hunting combining with the resultative *down* (*hunt/track/trail/follow down*) and verbs of cooking with the aspectual *up* (*bake/cook/fry/broil up*), forming essentially productive VPCs. Idiomatic VPCs, however, will not accept the substitution of one of its members by a related term (e.g. *get* and its synonyms in *get/*obtain/*receive over*), even if at first glance this could seem natural. In our experiments, we will consider that a VPC is compositional if it accepts: the replacement of the verb by a synonym, or of the preposition by another preposition. Summarising our hypothesis, we get:

- For Q1: Is the candidate syntactically flexible, i.e. does it allow the configurations S1 through S3?
  - NO: non-VPC
  - YES: VPC
- For Q2: Is the candidate semantically flexible, allowing the substitution of a member by a related word?
  - NO: idiomatic VPC
  - YES: compositional VPC

4 Data Sources

To generate a gold standard, we used the Baldwin VPC candidates dataset (henceforth Baldwin CD)[3], which contains 3,078 English VPC candidates annotated with information about idiomaticity (14.5% are considered idiomatic). We further annotated this dataset with information about whether each candidate is a genuine VPC or not, where a candidate is consider *genuine* if it belongs to at least one of a set of machine-readable dictionaries: the Alvey Natural Language Tools (ANLT) lexicon (Carroll and Grover, 1989), the Comlex lexicon (Macleod and Grishman, 1998), and the LinGO English Resource Grammar (ERG) (Copestake and Flickinger, 2000)[4]. With this criterion 81.8% of them are considered genuine VPCs.

To gather information about the candidates in this work we employ both a fragment of 1.8M sentences from the British National Corpus (BNC Burnard (2000)) and the Web as corpora. The BNC fragment is used to calculate the correlation measures since they require a corpus with known size. The Web is used to generate frequencies for the entropy measures, as discussed in § 5.2. Web frequencies are approximated by the number of pages containing a candidate and indexed by Yahoo Search API. In order to keep the searches as simple and self-sufficient as possible, no additional sources of information are used (Villavicencio, 2005). Therefore, the frequencies are quite conservative in the sense that by employing inflected forms of verbs, potentially much more evidence could be gathered.

For the generation of semantic variational patterns, we use both Wordnet 3.0 (Fellbaum, 1998) and Levin's English Verb Classes and Alternations (Levin, 1993). Wordnet is organised as a graph of concepts, called *synsets*, linked by relations of synonymy, hyponymy, etc. Each synset contains a list of words that represent the concept. The verbs in a synset and its synonym synsets are used to generate variations of a VPC candidate. Likewise we use Levin's classes, which define 190 fine-grained classes for English verbs, based on their syntactic and semantic features.

It is important to highlight that the generation of the semantic variations strongly relies on these resources. Therefore, cross-language extension would depend on the availability of similar tools for the target language.

5 Carrying out the experiments

Our experiments are composed of two stages, each one consisting of three steps (corresponding to the next three sections). The first stage filters out every candidate that is evaluated as not being a VPC, while the second one intends to identify the idiomatic VPCs among the remaining candidates of the previous stage.

5.1 Generating candidates

For each of the 3,078 items in the Baldwin CD we generated 2 sets of variations, syntactic and semantic, and we will refer to these as *alternative forms* or *variations* of a candidate.

The syntactic variations are generated using the patterns S1 to S3 described in section 3. Following the work of Villavicencio (2005) 3 frequently used prepositions *for*, *from* and *with* are used as delimiters and we search for NPs in the form of pronouns like *this* and definite NPs like *the boy*. The use of alternative search patterns also helps to give an in-

[3]This dataset was provided by Timothy Baldwin for the MWE2008 Workshop.

[4]Version of November 2001.

dication about the syntactic distribution of a candidate VPC, and consequently if it has a preferred syntactic realisation. For instance, for *eat up* and the delimiter *with*, we propose a list of Web search queries for its respective variations v_i, shown with their corresponding Web frequencies in table 1.[5]

Variation (v_i)	Frequency ($n_{Yahoo}(v_i)$)
eat up with	49200
eat the * up with	2240
eat this up with	1120
eat up the * with	3110

Table 1: Distribution of syntactic variations for the candidate *eat up*.

For the semantic variations, in order to capture the idiomaticity of VPCs we generate the alternative forms by replacing the verb by its synonym verbs as follows:

WNS Wordnet Strict variations. When using Wordnet, we consider any verb that belongs to the same synset of the candidate as a synonym.

WNL Wordnet Loose variations. This is an indirect synonymy relation capturing any verb in Wordnet that belongs either to the same synset or to a synset that is synonym of the synset in which the candidate verb is contained.

Levin These include all verbs in the same Levin class as the candidate.

Multiword synonyms are ignored in this step to avoid noisy search patterns, (e.g. **eat up up*). The examples for these variations are shown in table 2 for the candidate *act in*.

Wordnet and Levin are considered ambiguous resources because one verb is potentially contained in several synsets or classes. However, as Word Sense Disambiguation is not within the scope of this work we employ some heuristics to select a given sense for the candidate verb. In order to test the effect of frequency, the first heuristic adopts the first synset in the list, as Wordnet organises synsets in descending order of frequency (denoted as *first*). To study the influence of the number of synonyms, the second and third heuristics use respectively the biggest (*max*) and smallest (*min*) synsets. The last heuristic is the union of all synonyms (*all*). These heuristics are indicated using a subscript notation, where e.g. WNS_{all} symbolizes the WNS variations set using the union of all synsets as disambiguation heuristic. Finally, we generated two additional sets of candidates by replacing the particle by one of the 48 prepositions listed in the ANLT dictionary (*prep*) and also by one of 9 chosen locative prepositions (*loc-prep*). It is important to also verify possible variations of the preposition because compositional VPCs combine productively with one or more groups of particles, e.g. locatives, and present consequently a wider probability distribution among the variations, while an idiomatic VPC presents a higher frequency for a chosen preposition.

[5]The Yahoo wildcard used in these searches matches any word occurring in that particular position.

Variation (v_i)	Source	$n_{Yahoo}(v_i)$
act in	—	2690
playact in	WNS	0
play in	WNS	167000
behave in	WNL	98
do in	WNL	24600
pose in	$Levin$	1610
qualify in	$Levin$	358
rank in	$Levin$	706
rate in	$Levin$	16700
serve in	$Levin$	2240

Table 2: Distribution of syntactic variations for the candidate *eat up*.

5.2 Working the statistical measures out

The classifications of the candidate VPCs are done using a set of measures: the frequencies of the VPC candidates and of their individual words, their Mutual Information (MI), χ^2 and Entropies. We calculate the MI and χ^2 indices of a candidate formed by a verb and a particle based on their individual frequencies and on their co-occurrence in the BNC fragment.

The *Entropy* measure is given by

$$H(V) = -\sum_{i=1}^{n} p(v_i) \ln [\, p(v_i) \,]$$

where

$$p(v_i) = \frac{n(v_i)}{\sum\limits_{\forall\ v_j \in V} n(v_j)}$$

is the probability of the variation v_i to occur among the set of all possible variations V =

$H(V) \leq 0.001081$
$n_{BNC}(p) \leq 51611$
$n_{Yahoo}(v_{transitive}) \leq 1$
$n_{Yahoo}(v) \leq 2020000000 : yes$
$n_{Yahoo}(v) > 2020000000$
$\chi^2 \leq 25.99$
...

Figure 1: Fragment of the decision tree that filters out non-VPCs.

$\{v_1, v_2, \ldots, v_n\}$, and $n(v_i)$ is the Web frequency for the variation v_i.

The entropy of a probability distribution gives us some clues about its shape. A very low entropy is a sign of a heterogeneous distribution that contains a peak. On the other hand, a distribution that presents uniformity will lead to a high entropy value.

The interest of $H(V)$ for the detection of VPCs is in that true instances are more likely to not prefer a canonical form, more widely distributing probabilities over all alternative syntactic frames (S1 to S3), while non-VPCs are more likely to choose one frame and present low frequencies for the proposed variations.

For the semantic variations, the entropy is calculated from a set V of variations generated by the Wordnet synset, Levin class and preposition substitutions described in § 5.1. The interpretation of the entropy at this point is that high entropy indicates compositionality while low entropy indicates idiomaticity, since compositional VPCs are more productive and distribute well over a class of verbs or a class of prepositions and idiomatic VPCs prefer a specific verb or preposition.

5.3 Bringing estimations together

Once we got a set of measures to predict VPCs and another to predict their idiomaticity/compositionality, we would like to know which measures are useful. Therefore, we combine our measures automatically by building a decision tree with the `J48` algorithm, a version of the traditional entropy-based `C4.5` algorithm implemented in the Weka package.[6]

6 Weighting the results up

The first stage of our experiments applied to the 3,078 VPC candidates generated a decision tree using 10-fold cross validation that is partially reproduced in figure 1. From these, 2,848 candidates were considered genuine VPCs, with 2,419 true positives, 100 false negatives and 429 false positives. This leads to a recall of 96% of the VPCs being kept in the list with a precision of 84.9%, and an f-measure of 90.1%. We interpret this as a very positive result since although some false negatives have been filtered out, the remaining candidates are now less noisy.

Figure 1 shows that the entropy of the variations is the best predictor since it is at the root of the tree. We can also see that there are several types of raw frequencies being used before a correlation measure appears (χ^2). We can conclude that the frequency of each transitive, intransitive and split configurations are also good predictors to detect false from true VPCs. At this point, MI does not seem to contribute to the classification task.

For our second stage, we generated Wordnet synonym, Levin class and preposition variations for a list of the 2,867 VPC candidates classified as genuine cases. We also took into account the proportion of synonyms that are MWEs (*vpc-syn*) and the proportion of synonyms that contain the candidate itself (*self-syn*).

In order to know what kind of contribution each measure gives to the construction of the decision tree, we used a simple iterative algorithm that constructs the set U of useful attributes. It first initialises U with all attributes, then calculates the precision for each class (yes and no)[7] on a cross validation using all attributes in U. For each attribute $a \in U$, it ignores a and recalculates precisions. If both precisions decrease, the contribution of a is positive, if both increase then a is negative, else its contribution remains unknown. All features that contribute negatively are removed from U, and the algorithm is repeated until there is no negative attribute left.

The step-by-step execution of the algorithm can be observed in table 3, where the inconclusive steps are hidden. We found out that the optimal features are $U^* = \{$*self-syn*, $H($*prep*$)$, $H(Levin_{first})$, $H(WNS_{first})$, $H(WNS_{min})$, $H(Levin_{max})$, $H(Levin_{min}).\}$ The *self-syn* information seems to be very important, as without it precisions of both classes decrease considerably

[6] http://www.cs.waikato.ac.nz/ml/weka/

[7] We use the precision as a quality estimator since it gives a good idea of the amount of work that a grammar engineer or lexicographer must perform in order to clear the list from false positives.

#	Ignored	Precision No	Precision Yes	+/−
	1^{st} iteration			
0	—	86.6%	54.9%	
1	*vpc-syn*	**86.7%**	**56.6%**	−
2	*self-syn*	85.2%	28.7%	+
4	H(*loc-prep*)	**86.7%**	**56.1%**	−
6	$H(WNS_{max})$	**87.5%**	**57.4%**	−
9	$H(WNLfirst)$	**86.7%**	**57.9%**	−
10	$H(WNL_{max})$	**86.7%**	**57.8%**	−
11	$H(WNL_{min})$	**86.9%**	**57.6%**	−
16	$H(Levin_{all})$	**86.7%**	**55.1%**	−
	2^{nd} iteration			
17	—	87.7%	60.3%	
18	H(*prep*)	87.6%	59.2%	+
21	$H(WNS_{all})$	**87.8%**	**61.6%**	−
22	$H(WNL_{all})$	**87.8%**	**61.0%**	−
23	$H(Levin_{first})$	87.5%	60.2%	+
	3^{rd} iteration			
26	—	87.8%	61.9%	
27	$H(WNS_{first})$	87.8%	61.9%	±
28	$H(WNS_{min})$	87.7%	61.1%	+
29	$H(Levin_{max})$	87.8%	61.6	±
30	$H(Levin_{min})$	87.7%	61.5%	+

Table 3: Iterative attributes selection process. Precision in each class is used as quality estimator.

(experiment #2).

All entropies of the WNL heuristics are of little or no utility. This could probably be explained by either the choice of simple WSD heuristics for selecting synsets, or because the indirect synonymy information is too far related to the original verb to be used in variational patterns. Inspecting the generated variations, we notice that most of the synonym synsets are related to secondary senses or very specific uses of a verb and are thus not correctly disambiguated.

In what concerns the WNS sets, only the smallest and first synset were kept, suggesting again that it may not be a good idea to maximise the synonyms set and for future work, we intent to establish a threshold for a synset to be taken into account. In addition, we can also infer a positive contribution of the frequency of a sense with the choice of the first synset returned by Wordnet resulting in a reasonable WSD heuristic (which is compatible with the results by McCarthy et al. (2004)).

On the other hand, the algorithm selected the first, the smallest and the biggest of the Levin's sets. This probably happens because the majority of these verbs belongs only to one or two, but never to a great number of classes. Since the granularity of the classes is coarser than for synsets, the heuristics often offer four equal or very close entropies and thus redundant information. As an overall result, the last iteration shown in table 3 indicates a precision of 61.9% for the classifier in detecting idiomatic VPCs, that is to say that we automatically retrieved 176 VPCs where 67 are false positives and 109 are truly idiomatic. This value is a quality estimator for the resulting VPCs that will potentially be used in the construction of a lexicon. Recall of idiomatic VPCs goes from 16.7% to 24.9%.

7 Conclusions

One of the important challenges for robust natural language processing systems is to be able to successfully deal with Multiword Expressions and related constructions. We investigated the identification of VPCs using a combination of statistical methods and linguistic information, and whether there is a correlation between the productivity of VPCs and their semantics that could help us detect if a VPC is idiomatic or compositional.

The results confirm that the use of statistical and linguistic information to automatically identify verb-particle constructions presents a reasonable way of improving coverage of existing lexical resources in a very simple and straightforward manner. In terms of grammar engineering, the information about compositional candidates belonging to productive classes provides us with the basis for constructing a family of fine-grained redundancy rules for these classes. These rules are applied in a constrained way to verbs already in the lexicon, according to their semantic classes. The VPCs identified as idiomatic, on the other hand, need to be explicitly added to the lexicon, after their semantic is determined. This study can also be complemented with the results of investigations into the semantics of VPCs, as discussed by both Bannard (2005) and McCarthy et al. (2003).

In addition, the use of clustering methods is an interesting possibility for automatically identifying clusters of productive classes of both verbs and of particles that combine well together.

Acknowledgments

This research was partly supported by the CNPq research project *Recuperação de Informações Multilíngües* (CNPq Universal 484585/2007-0).

References

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

Baldwin, Timothy. 2005. Deep lexical acquisition of verb-particle constructions. *Computer Speech and Language*, 19(4):398–414.

Bannard, Colin J. 2005. Learning about the meaning of verb-particle constructions from corpora. *Computer Speech and Language*, 19(4):467–478.

Bolinger, Dwight. 1971. *The phrasal verb in English*. Harvard University Press, Harvard, USA.

Burnard, Lou. 2000. User reference guide for the British National Corpus. Technical report, Oxford University Computing Services.

Carroll, John and Claire Grover. 1989. The derivation of a large computational lexicon of English from LDOCE. In Boguraev, B. and E. Briscoe, editors, *Computational Lexicography for Natural Language Processing*. Longman.

Copestake, Ann and Dan Flickinger. 2000. An open-source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*.

Dehé, Nicole. 2002. *Particle verbs in English: syntax, information structure and intonation*. John Benjamins, Amsterdam/Philadelphia.

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

Fazly, Afsaneh and Suzanne Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *EACL*. The Association for Computer Linguistics.

Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May.

Fraser, Bruce. 1976. *The Verb-Particle Combination in English*. Academic Press, New York, USA.

Jackendoff, Ray. 2002. English particle constructions, the lexicon, and the autonomy of syntax. In N. Dehé, R. Jackendoff, A. McIntyre and S. Urban, editors, *Verb-Particle Explorations*. Berlin: Mouton de Gruyter.

Levin, Beth. 1993. *English Verb Classes and Alternations: a preliminary investigation*. University of Chicago Press, Chicago and London.

Macleod, Catherine and Ralph Grishman. 1998. Comlex syntax reference manual, Proteus Project.

McCarthy, Diana, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL 2003 workshop on Multiword expressions*, pages 73–80, Morristown, NJ, USA. Association for Computational Linguistics.

McCarthy, Diana, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 279. Association for Computational Linguistics.

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

Ramisch, Carlos, Paulo Schreiner, Marco Idiart, and Aline Villavicencio. 2008. An evaluation of methods for the extraction of multiword expressions. In *Proceedings of the LREC Workshop - Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 50–53, Marrakech, Morocco, June.

Sharoff, Serge. 2004. What is at stake: a case study of russian expressions starting with a preposition. pages 17–23, Barcelona, Spain.

Villavicencio, Aline. 2005. The availability of verb-particle constructions in lexical resources: How much is enough? *Journal of Computer Speech and Language Processing*, 19(4):415–432.

Wurmbrand, S. 2000. The structure(s) of particle verbs. Ms., McGill University.

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