Verb-Particle Constructions in the World Wide Web

Aline Villavicencio
University of Cambridge Computer Laboratory,
William Gates Building, JJ Thomson Avenue,
Cambridge, CB3 0FD, UK
Aline.Villavicencio@cl.cam.ac.uk

Abstract

In this paper we investigate the phenomenon of verb-particle constructions, discussing their characteristics and their availability for use with NLP systems. Combinations automatically extracted from corpora greatly improve the coverage of available resources. However, the data sparseness problem is particularly acute for these constructions and even using a corpus as large as the British National Corpus, a great proportion of combinations have a very low frequency, while others never occur in it. In this paper we propose using the World Wide Web as a way to validate candidate combinations minimising the problem of data sparseness. This method can be use to extend the coverage of existing lexical resources by validating combinations automatically generated from classes of verbs, and to improve the reliability of those combinations automatically extracted from corpora.

Keywords: Verb-Particle Constructions, Levin's Verbal Classes, World Wide Web.

1 Introduction

In this paper we investigate verb-particle constructions (VPCs) in English and their availability for NLP systems. Due to their complex characteristics and their flexible nature, they provide a challenge for NLP in general. In particular, there is a lack of adequate resources to identify and treat VPCs, and many applications tend to ignore them. However, due to their frequency in natural language interactions, it is clear that successful applications need to deal with them appropriately, if they want to capture natural languages.

VPCs are combinations of verbs and prepositional or adverbial particles, such as *eat up* in *Bob ate up all the chocolate*. In these constructions particles are characterised by containing features of motion-through-location and of completion or result in their core meaning [Bolinger1971]. However, VPCs can range from idiosyncratic or semi-idiosyncratic combinations, such as *give in* (in e.g. *Her son was so determined to get what he wanted that she finally gave in*), to more regular ones, such as *clean up* (in e.g. *He needs to clean up his flat*). Cases of 'idiomatic' VPCs like *give in*, meaning *to agree to what someone wants after a period when you refuse to agree*, where the meaning of the combination cannot be straightforwardly inferred from the meaning of the verb and the particle, fortunately seem to be a small minority [Side1990]. Most cases seem to be more regular, with the particle compositionally adding a specific meaning to the construction and following a productive pattern. Indeed, Side notes that particles in VPCs seem to fall into a set of possible categories, defined according to their meanings in VPCs. For instance, in his analysis of

VPCs involving *off*, which is defined as *indicating distance in time or space, departure, removal, disconnection, separation*, most VPCs considered seem to fit into this category. Examples are *take off* meaning to depart, *cut off* meaning to disconnect and *strain off* to remove. A three way classification is adopted in [Dehé2002], [Emonds1985] and [Jackendoff2002], where a VPC can be classified into compositional, idiomatic or aspectual, depending on its sense. In the compositional VPCs the meaning of the construction is determined by the literal interpretations of the particle and the verb (e.g. *throw out*). Idiomatic VPCs, on the other hand, cannot have their meaning determined by interpreting their components literally (e.g. *go off* meaning 'to explode'). The third class, of aspectual VPCs, 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*). In the investigation described here we focus on compositional and aspectual VPCs.

VPCs have been the subject of a considerable amount of interest, and some investigation has been done on the subject of productive VPCs. Bame [Bame1999] analyses some of these productive cases in the framework of Head-Driven Phrase Structure Grammar: namely those of aspectual and resultative combinations using the particle *up*. For example in *Kim carried the television up* the resultative up indicates that the argument is affected (i.e., at the end of the action the television is *up*). In contrast, the aspectual *up* in *Kim ate the sandwich up* suggests that the action is taken to some conclusion (i.e., the sandwich is totally consumed at the end of the action). Villavicencio and Copestake [Villavicencio and Copestake2002] propose defining a family of lexical rules, organised in a default inheritance hierarchy, to capture productive patterns of verb-particle constructions like these. Fraser points out that semantic properties of verbs can affect their possibilities of combining with particles [Fraser1976]. For example *bolt, cement, clam, glue, paste* and *nail* all are semantically similar verbs where the objects specified by the verbs are used to join material and they can all productively 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*. Moreover, Side notes that frequently new VPCs are formed by analogy with existing ones, with often the verb being varied and the particle remaining (e.g. *hang on, hold on* and *wait on*).

As these works suggest, many VPCs follow productive patterns, where semantically related verbs are combined with a given sense of a particle. By identifying classes of verbs that follow patterns such as these in VPCs, we are able to maximise the use of the information contained in lexical resources. In this way, we can make use of regular patterns to productively generate VPCs from verbs already listed in a lexical resource, according to their verbal classes and the particles with which they can combine. For example, the resultative combinations *walk/run/jump up/down/out/in/away/around* from the motion verbs *walk, run* and *jump* and the directional/locative particles *up*, *down*, *out*, *in*, *away* and *around*. The use of Levin's classification of verbs [Levin1993] to productively generate candidate VPCs from semantically related verbs is a possible alternative to extend the coverage of lexical resources, as suggested by Villavicencio [Villavicencio2003]. The verbal classes seems to be good indicators of productivity in verb-particle constructions. However, the data sparseness problem, which is particularly acute for multiword expressions like VPCs, means that the full contribution made by the candidate VPCs needs yet to be determined, since not even a 100 million word corpus is enough, and many of the combinations proposed cannot be verified. From these combinations some may be valid, but simply do not occur in the corpus, while other are genuinely invalid. In this paper we propose to verify the validity of VPCs automatically generated from classes of verbs by searching for them using the World Wide Web as corpus, in order to minimise the problem of data sparseness, following Grefenstette [Grefenstette1999] and Keller et al. [Keller et al.2002].

We start by discussing some characteristics of VPCs that make them so challenging. Then we analyse the coverage provided by some available lexical resources, and the use of corpora to extend the coverage provided by them in sections 3 and 4. We then discuss Levin's classes of verbs and the combinations they productively generate with the particle *up*, which is the most widely used of particles. Next we address the issue of how these can be validated using the World Wide Web, to avoid the problem of data sparseness, finalising with a discussion of the results obtained and future work.

2 VPCs in a Nutshell

In this section we discuss some of the characteristics that make VPCs so challenging for NLP. VPCs are often highly polysemous, with for instance, eight senses being listed for *make up* in the Collins Cobuild Dictionary of Phrasal Verbs (e.g. *to form something* and *to invent*). They also show syntactic variation, where each combination can take part in several different subcategorisation frames. For example, *add up* can occur as an intransitive verb-particle combination in *It's a few calories here and there, and it all quickly adds up* or as a transitive one in *We need to add these marks up*.

In transitive VPCs, where an NP complement is required, some particles have a fixed position in relation to the verb, such as *come up* in *She came up with the idea*, where the particle is expected immediately after the verb. Thus we cannot have **She came with the idea up*. Other combinations have a more flexible order in relation to the verb, and can equally well occur after another complement or immediately after the verb: e.g. *John ate his cereal up* and *John ate up his cereal*. In the latter, the particle comes before a simple definite NP without taking it as its object (unlike e.g. *It consists of two parts*, which is a prepositional verb). Whether a particle can be separated or not from the verb may depend on the degree of bondage of the particle with the verb, on the size of the NP, and on the kind of NP. Thus, when the NP is an unstressed personal pronoun, in a transitive VPC, it must precede the particle (e.g. *They ate it up* but not **They ate up it*). This is also the case for VPCs subcategorising for other verbal complements, like PPs and sentential complements, where the particle must come immediately after the verb. Besides complements, certain adverbs are also accepted between the verb and the particle, such as *right* in *He came* ***right*** *back*.

3 VPCs and Dictionaries

In this section we analyse some of the lexical resources available for NLP systems, in terms of the VPCs they contain. In table 1 we can see the coverage of phrasal verbs (PVs) in several dictionaries and lexicons: Collins Cobuild Dictionary of Phrasal Verbs (Collins-PV), Cambridge International Dictionary of Phrasal Verbs (CIDE-PV), the electronic versions of the Alvey Natural Language Tools (ANLT) lexicon [Carroll and Grover1989] (which was derived from the Longman Dictionary of Contemporary English, LDOCE), the COMLEX lexicon [Macleod and Grishman1998], and the LinGO English Resource Grammar (ERG) [Copestake and Flickinger2000] version of November 2001. This table shows in the second column the number of PV entries for each of these dictionaries, including not only verb-particle constructions but also prepositional verbs. The third column shows the number of VPC entries (available only for the electronic dictionaries).

These dictionaries have a considerable number of PV entries potentially providing us with a good starting point for handling VPCs. Each of them uses a slightly different set of verbs and particles in its VPCs, and table 2 shows some of their characteristics, where the seventh column shows the

Table 1: Phrasal Verb Entries in Dictionaries

Dictionary	PV Entries	VPC Entries
ANLT	6,439	2,906
CIDE-PV	over 4,500	-
Collins-PV	over 3,000	-
Comlex	12,564	4,039
ERG	533	337

Table 2: VPCs in Dictionaries

Dictionary	Verbs	VPCs Entries	Distinct VPCs	Particles	Verbs in VPCs	Proportion of Verbs in VPCs
ANLT	5,667	2,906	2,250	44	1,135	20%
Comlex	5,577	4,039	1,909	23	990	17.75%
ERG	1,223	337	270	25	176	14.39%
A+C	6,043	-	3,107	44	1,394	23.07%
A∩C	5,201	-	1,052	23	731	14.05%
A+C+E	6,113	-	3,156	45	1,400	22.90%

proportion of verbs used in VPCs from all the verbs in the dictionary. In this table A+C represents the union of ANLT and Comlex, A∩C their intersection and A+C+E the union of ANLT, Comlex and ERG

When we rank the particles according to the frequency with which they occur in the VPCs, we get similar patterns for all of the dictionaries, figure 1. This figure shows the 5 top ranked particles for each of the dictionaries, and for all of them *up* is the particle involved in the largest number of combinations.

In each of these dictionaries only a small proportion of the total number of verbs is used in its VPCs, as can be seen in table 2 and figure 2. For example, only 20% of the verbs listed in the ANLT form at least one VPC. For the other dictionaries this proportion is even lower. These tend to be very widely used and general verbs, such as *come, go, get, put, bring* and *take*. Which of the remaining verbs do not form valid VPCs and which verbs form VPCs that were simply omitted needs to be investigated, and in this paper we attempt to go one step in this direction.

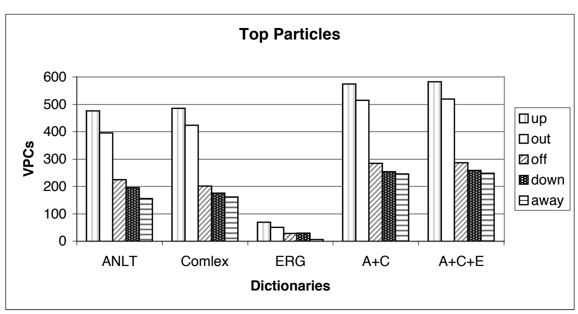

Figure 1: Top Particles in Dictionaries

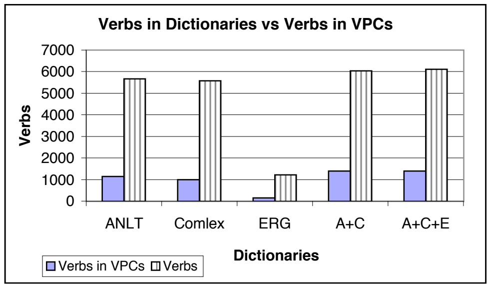

Figure 2: Verbs in VPCs and Verbs in Dictionaries

The number of VPCs listed in each dictionary is shown in table 2, where we can also see the increase in the number of VPCs obtained by the union of the dictionaries. Even though there is a large number of entries already obtained by combining the two largest dictionaries, ANLT and Comlex, a considerable proportion (16%) of the entries in the LinGO ERG lexicon are not listed in any of them (this proportion would increase if we took subcategorization etc into account).[1] Most of these are at least semi-compositional, e.g., *crisp up*, *come together*, *tie on*, and were probably omitted from the dictionaries for that reason,[2] though some others, such as *hack up*, are probably recent coinages. Dictionaries are static resources that tend to list idiosyncratic combinations at the expense of omitting the more productive ones, so we cannot rely only on the combinations they provide.

4 VPCs and Corpora

The number of VPCs is constantly growing, and we need ways of extending the coverage provided by lexical resources. The use of corpora to extract VPCs can contribute to extending their coverage. An investigation of the automatic extraction of combinations from corpora is described in [Baldwin and Villavicencio2002]. In this section we use VPCs extracted from the British National Corpus (BNC), as described in [Bannard et al.2003] and [McCarthy et al.2003], and we compare these VPCs with those contained in the combined dictionaries (A+C+E-VPCs), and how the former can be used to complement the coverage provided by the latter.

The BNC is a 100 million word corpus containing samples of written text from a wide variety of sources, designed to represent as wide a range of modern British English as possible. It includes texts from newspapers, journals, books, and many other sources. Using the methods described in [Baldwin and Villavicencio2002], 8,751 VPC entries were extracted from the BNC. These entries are classified into intransitive and/or transitive VPCs, depending on their subcategorisation frame, and they result in 7,078 distinct VPCs. Some of these entries are not VPCs but rather noise, such as **** *off*, *'s down*, etc. After removing the most obvious cases of noise, there were 7,070 VPCs left. These are formed by 2,542 verbs and 48 particles. McCarthy's method [McCarthy et al.2003] resulted in 4,482 VPCs, after the most obvious cases of noise were removed. They are formed by the combination of 1,999 verbs and 9 particles, among which there are cases of prepositional verbs.

These different extraction methods yielded different sets of VPCs, as we can see in table 3. This

[1]The LinGO ERG lexicon was manually constructed with most of the verb-particle entries being empirically motivated by the Verbmobil corpus. It is thus probably reasonably representative of a moderate-size domain-specific lexicon.

[2]The Cobuild Dictionary explicitly states that literal meanings and combinations are not given for all verbs.

table shows some comparisons, where BNC-1 represents the set of VPCs extracted using the methods described by Baldwin and Villavicencio, BNC-2 those extracted by McCarthy et al., and BNC the union of both. Even though these two methods operate in the same corpus, their results are quite distinct, with one complementing the other.

Table 3: Comparison between VPCs automatically extracted from corpora

Resources	VPC Entries	Verbs	Particles
BNC-1	7,070	2,542	48
BNC-2	4,482	1,999	9
BNC-1 - BNC-2	4,429	956	39
BNC-2 - BNC-1	1,841	413	0
BNC-1 ∩ BNC-2	2,641	1,586	9
BNC	8,911	2,955	48

In terms of the VPCs, by joining A+C+E-VPCs with all the VPCs extracted from the BNC (BNC-VPCs) there is an increase of 209% in the number of VPCs, since from the 8,911 VPCs in BNC, only 2,318 are also in the combined dictionaries, as can be seen in table 4. A considerable number of the extracted VPCs form productive combinations, some containing a more informal or a recent use of verbs (e.g. *hop off*, *kangaroo down* and *skateboard away*). These VPCs provide a useful addition to the information contained in the dictionaries. Thus, when the combined dictionaries are joined with the BNC, there is a major increase in the number of VPCs, to a total of 9,745 distinct combinations.

These methods provide us with a larger set of VPCs and some information about their syntactic behaviour, like their subcategorisation frames. However, they suffer from the problem of data sparseness and a great proportion of the extracted VPCs have a very low frequency. For instance, in BNC-2 40.52% of the combinations occur only once. Among these we have genuine combinations but there are also instances of false positives or noise, and it is difficult to decide which is which on the basis of one occurrence. One possible way of minimising this problem is to use the World Wide Web as an extremely large corpus, since as pointed out by Grefenstette [Grefenstette1999] and Keller et al. [Keller et al.2002] the web is the largest data set available for NLP: in December 2002 the web contained at least 3,033 million pages, which were indexed by the search engine Google, according to the Search Engine Showdown (http://www.searchengineshowdown.com). Several researchers have started to explore this idea, making use of this huge resource to overcome the problem of data sparseness. For instance, Keller et al use the web to obtain frequencies

Table 4: Comparison between VPCs from Combined Dictionaries and those from BNC

Resources	VPC Entries	Verbs	Particles
A+C+E	3,156	1,400	45
BNC	8,911	2,955	48
A+C+E - BNC	834	160	17
BNC - A+C+E	6,593	1,715	20
A+C+E ∩ BNC	2,318	1,240	28
A+C+E+BNC	9,745	3,115	65

for adjective-noun, noun-noun and verb-object bigrams, testing if the web could be used to obtain frequencies for bigrams that are unseen in a given corpus. They suggest that the large amount of data available in the web largely outweighs any problem that may derive from it being unbalanced and containing noise. Grefenstette employs the web to do example-based machine translation of compounds from French into English. The method he employs would suffer considerably from data sparseness if it were only to rely on corpus data, so for compounds that are sparse in the BNC he also obtains frequencies from the web. In the next sections we investigate how the web can help us find evidence to distinguish between valid VPCs and noise in automatically generated or extracted combinations. In order to do that we employ a verbal classification to generate candidate VPCs, which can be used to extend the coverage of the available resources, and how the web can be used to filter and obtain frequencies for the candidate VPCs.

5 VPCs in the Web

We now discuss the possibility of automatically generating candidate VPCs from a set of verbs and particles, and how the validity of the combinations can be determined using the web as a corpus. In this investigation we concentrate on VPCs generated by combining a classification of semantically related verbs and the particle *up*.

5.1 The Candidate VPC Set

Fraser [Fraser1976] noted how semantic properties of verbs can affect their possibilities of combination with particles. For example verbs of hunting and the resultative *down* (*hunt/track/trail/follow down*) and verbs of cooking and the aspectual *up* (*bake/cook/fry/broil up*). As semantic properties of verbs can influence the patterns of combination with particles that they follow, by having a semantic classification of verbs we can investigate how they combine with certain particles. This can be used to extend the coverage of the available resources by generating VPCs from classes of related verbs that follow productive patterns of combinations. One such classification was proposed by Levin [Levin1993], where verbs are grouped into classes according to semantic and syntactic properties, based on the assumption that the syntactic behaviour of verbs is semantically determined. In this section we discuss the possibility of using Levin's classes of verbs to generate candidate verb-particle combinations, as suggested in [Villavicencio2003].

In Levin's classification there are 190 classes and subclasses that capture 3,100 different verbs listed, resulting in 4,167 entries, since each verb can belong to more than one class. For example, the verb *to run* belongs to classes 26.3 (Verbs of Preparing), 47.5.1 (Swarm Verbs), 47.7 (Meander Verbs) and 51.3.2 (Run Verbs). The number of elements in each class varies considerably, so that 60% of all of these classes have more than 10 elements, accounting for 88% of the verbs, while the other 40% of the classes have 10 or less elements, capturing the remaining 22% of the verbs. The 5 larger classes are shown in table 5.

All the combinations formed by Levin's classes and the particle *up* were produced. The combinations were generated by taking each verb and appending the particle to it. It is necessary to test the validity of a candidate VPC, since not all verbs can be combined with particles. For example, Fraser notes the generalisation that stative verbs almost never combine with a particle (e.g. *know, want, hope, resemble*, etc) [Fraser1976]. Some other verbs seem to occur with only one particle (e.g. *chicken out* and *sober up*). Moreover, although there are some cases where it appears reasonable to treat verb-particle combination as fully productive (within fairly finely specified classes), there are also cases of semi-productivity. For instance, many verbs denoting cooking

Table 5: Verb Entries in Levin's Classes

Class	Entries
45.4	257
31.1	220
51.3.2	124
43.2	119
9.9	109

processes can occur with aspectual *up*: e.g., *boil up*, *fry up*, *brew up*, *heat up*. But some other combinations seem odd e.g., ?*sauté up*. This problem of semi-productivity is further discussed in [Villavicencio and Copestake2002]. Nonetheless, some verbal classes (and particles) seem to be good indicators of VPC acceptability. For example, in Class 11.3 (Verbs of Bring and Take), all verbs form valid combinations with all the particles investigated (*in, down, out, up*), according to the combined resources [Villavicencio2003].

5.2 Looking for VPCs in the Web

From the 4,167 verbs listed in Levin's classification the majority, 3,933, are in the combined resources. However, from the 4,167 possible VPCs generated from combining the verbs in Levin's classes with *up*, only 1,674 are in the combined resources (A+C+E+BNC-VPCs). Even though the combined resources have a large number of VPCs, this coverage is still limited. For instance, in a manual analysis of the combinations involving the class of motion verbs, a great proportion of the VPCs are not attested in these resources, even if most of the combinations are considered acceptable by native speakers. It is necessary to establish whether the unattested VPCs genuinely do not form valid combinations, or whether they do not occur due to the data sparseness problem. In this section we discuss how to use the web to verify if the candidate VPCs are genuine on the basis of the frequency with which they occur.

As not all verbs in Levin's classes will form valid VPCs, each of the combination that was unattested in the combined resources was searched in the web using the search engine Google. For each combination searched, Google provided us with a measure of frequency in the form of the number of pages in which that combination appeared. Since we want to be able to identify and exclude the invalid cases, we assume that if a VPC is not attested either in the web or in the combined resources, then it is not a valid VPC. In order to provide a uniform search pattern for all the VPCs, we searched for all of them as intransitive VPCs, which is one of the most common subcategorisation frames for VPCs. We thus abstracted away from the problem of searching instances with fixed/flexible word order and different subcategorisation frames. Furthermore, in order to ensure that only VPCs, and not prepositional verbs were retrieved, we used the following context for the searches: "VERB up for". It ensures that *up* is not followed by an NP, which would be ambiguous between a transitive VPC (Verb Particle NP) and a prepositional verb, where the PP is headed by *up* (Verb PP). Only pages containing this exact term are retrieved. For instance, *slim up*, which was unattested in the combined resources, had 1,500,000 pages retrieved with *slim up for*, and one of them contains the sentence: *Why do we need to spend tax money to convince you to slim up for your own good?*.

Using the web as corpus, a total of 2,094 of the candidate VPCs were considered valid. For these the maximum number of pages retrieved was 9,330,000 for *mail up* and the minimum was 11 for *desprout up*. Among the unattested combinations we have *genuflect up* and *salaam up*. As a result a total of 3,768 VPCs out of the 4,167 candidate VPCs was attested in the combined resources or in the web, corresponding to 90% of the possible candidates. From these, 890 are cases of VPCs containing verbs that were listed in the combined resources but were not used in any VPC listed in them. In terms of the classes 66% of them had all of its candidate VPCs considered valid; 24% had 50% or more of its VPCs as valid. In the remaining 10% of the classes, only 3 of them had no attested VPCs: Classes 37.9 (Advise Verbs), 39.4 (Devour Verbs) and 40.1.3 (Exhale Verbs), which contain a total of 15 VPCs. By joining them with A+C+E+BNC-VPCs we have an increase of 21% in the number of VPCs with a total of 11,796 VPCs.

6 Discussion

In this paper, we investigated the use of a verbal classification to productively generate candidate VPCs , using the web as a way of verifying the validity of VPCs and filtering out unattested cases. Searching the web for candidate VPCs generated from Levin's classes on the basis of their semantic/syntactic interrelations, rather than searching for any possible occurrence of a word followed by a particle, means that we ensured that only genuine verbs were used in the combinations avoiding random noise caused by misspelled words, non-native speakers, pages in other languages, etc. Thus, we used Levin's classes as a means of constraining the possible combinations and the web as a means of filtering unattested VPCs. We used this method to extend the coverage of lexical resources with automatically generated VPCs, and it can also be used to verify which VPCs automatically extracted from corpora that had a low frequency are genuine ones, while at the same time reinforcing their frequencies using the web.

These results suggest that Levin's classes are indeed a good starting point for obtaining productive patterns in verb-particle constructions. This investigation focused only on the particle *up* as a test case, but it is already possible to see an improvement in the coverage of the lexical resources. A more wide investigation using a larger set of verbs and particles and human annotators is envisaged, to extend even further the coverage of existing lexical resources. This investigation will continue to address the question of the great number of the verbal entries in a lexical resource not used in its VPCs, using the web to search for candidate VPCs generated by these verbs. Nonetheless, the results obtained so far are encouraging and confirm that we can straightforwardly extend the coverage of lexical resources by using a semantic classification of verbs to productively generate possible VPCs, and validating them using the web as a very large corpus.

Acknowledgments

I'd like to thank Timothy Baldwin and Diana McCarthy for kindly making available the combinations extracted from the BNC. This research was supported in part by the NTT/Stanford Research Collaboration, research project on multiword expressions.

References

[Baldwin and Villavicencio2002] Timothy Baldwin and Aline 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.

[Bame1999] Ken Bame. 1999. Aspectual and resultative verb-particle constructions with Up. Handout for talk presented at the Ohio State University Linguistics Graduate Student Colloquium, May.

[Bannard et al.2003] Colin Bannard, Timothy Baldwin, and Alex Lascarides. 2003. A statistical approach to the semantics of verb-particles. In Francis Bond, Anna Korhonen, Diana McCarthy, and Aline Villavicencio, editors, *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 65–72, Sapporo, Japan.

[Bolinger1971] Dwight Bolinger. 1971. *The Phrasal Verb in English.* Harvard University Press, Harvard, USA.

[Carroll and Grover1989] John Carroll and Claire Grover. 1989. The derivation of a large computational lexicon of English from LDOCE. In B. Boguraev and E. Briscoe, editors, *Computational Lexicography for Natural Language Processing*. Longman.

[Copestake and Flickinger2000] Ann Copestake 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).*

[Dehé2002] Nicole Dehé. 2002. *Particle Verbs in English: Syntax, Information Structure and Intonation.* John Benjamins, Amsterdam/Philadelphia.

[Emonds1985] Joseph E. Emonds. 1985. *A Unified Theory of Syntactic Categories.* Dordrecht: Foris.

[Fraser1976] Bruce Fraser. 1976. *The Verb-Particle Combination in English.* Academic Press, New York, USA.

[Grefenstette1999] Gregory Grefenstette. 1999. The World Wide Web as a resource for example-based machine translation tasks. In *Proc. of ASLIB, Conference on Translating and the Computer*, London.

[Jackendoff2002] Ray Jackendoff. 2002. English particle constructions, the lexicon, and the autonomy of syntax. In A. McIntyre N. Dehé, R. Jackendoff and S. Urban, editors, *Verb-Particle Explorations*. Berlin: Mouton de Gruyter.

[Keller et al.2002] Frank Keller, Maria Lapata, and Olga Ourioupina. 2002. Using the web to overcome data sparseness. In Jan Hajič and Yuji Matsumoto, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Philadelphia.

[Levin1993] Beth Levin. 1993. *English Verb Classes and Alternations - A Preliminary Investigation.* The University of Chicago Press.

[Macleod and Grishman1998] Catherine Macleod and Ralph Grishman. 1998. Comlex syntax reference manual, Proteus Project. `http://nlp.cs.nyu.edu/comlex`.

[McCarthy et al.2003] Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In Francis Bond, Anna Korhonen, Diana McCarthy, and Aline Villavicencio, editors, *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 73–80, Sapporo, Japan.

[Side1990] Richard Side. 1990. Phrasal verbs: sorting them out. *ELT Journal*, 44(2):144–52.

[Villavicencio and Copestake2002] Aline Villavicencio and Ann Copestake. 2002. Verb-particle constructions in a computational grammar of English. In Jongbok Kim and Stephen Wechsler, editors, *Proceedings of the Ninth International Conference on Head-Driven Phrase Structure Grammar*, Kyung-Hee University, Seoul, Korea. CSLI Publications.

[Villavicencio2003] Aline Villavicencio. 2003. Verb-particle constructions and lexical resources. In Francis Bond, Anna Korhonen, Diana McCarthy, and Aline Villavicencio, editors, *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 57–64, Sapporo, Japan.