Natural Language Engineering: page 1 of 31. © 2007 Cambridge University Press
doi:10.1017/S135132490600444X

A general feature space for automatic verb classification

ERIC JOANIS,[†] SUZANNE STEVENSON
and DAVID JAMES

Department of Computer Science, University of Toronto, 6 King's College Road,
Toronto, Ontario, Canada, M5S 3H5
e-mail: {joanis,suzanne,james}@cs.toronto.edu

(*Received 26 July 2005; revised 3 June 2006*)

Abstract

Lexical semantic classes of verbs play an important role in structuring complex predicate information in a lexicon, thereby avoiding redundancy and enabling generalizations across semantically similar verbs with respect to their usage. Such classes, however, require many person-years of expert effort to create manually, and methods are needed for automatically assigning verbs to appropriate classes. In this work, we develop and evaluate a feature space to support the automatic assignment of verbs into a well-known lexical semantic classification that is frequently used in natural language processing. The feature space is *general* – applicable to any class distinctions within the target classification; *broad* – tapping into a variety of semantic features of the classes; and *inexpensive* – requiring no more than a POS tagger and chunker. We perform experiments using support vector machines (SVMs) with the proposed feature space, demonstrating a reduction in error rate ranging from 48% to 88% over a chance baseline accuracy, across classification tasks of varying difficulty. In particular, we attain performance comparable to or better than that of feature sets manually selected for the particular tasks. Our results show that the approach is generally applicable, and reduces the need for resource-intensive linguistic analysis for each new classification task. We also perform a wide range of experiments to determine the most informative features in the feature space, finding that simple, easily extractable features suffice for good verb classification performance.

1 Introduction

Improved automatic text understanding requires detailed linguistic information about the words that comprise the text, motivating a recent wealth of work on lexical acquisition methods. Particularly crucial is knowledge about predicates, typically verbs, which communicate both the event being expressed and how participants are related to the event, i.e., "who did what to whom". Much research has focused on automatically learning properties of verbs from text corpora, such as their subcategorization (Brent 1993; Briscoe and Carroll 1997), argument roles (Riloff

[†] Current affiliation: Interactive Language Technologies Group, Institute for Information Technology, National Research Council Canada, A1330-101 St-Jean-Bosco Street, Gatineau, Quebec, Canada J8Y 3G5.

and Schmelzenbach 1998), selectional preferences (Resnik 1996), and lexical semantic classification (Dorr and Jones 1996; Lapata and Brew 1999, 2004; Schulte im Walde 2000; Merlo and Stevenson 2001; Korhonen and Briscoe 2004).

Lexical semantic classification of verbs has been shown to be important in a wide range of language tasks, such as lexical resource construction (Korhonen 2002; Villavicencio 2005), spoken dialog processing (Swift 2005), natural language generation for machine translation (Habash, Dorr and Traum 2003), semantic parsing (Shi and Mihalcea 2005), and semantic (participant) role labelling (see, e.g., Gildea and Jurafsky 2002, among many others since). These tasks rely on a class structure for verbs that captures their syntactic and semantic commonalities. However, manual creation of verb classifications comes at exorbitant expense, requiring large amounts of human time and expertise. Automatic classification of verbs is therefore necessary for extending existing lexicons, such as VerbNet (Kipper, Dang and Palmer 2000; Kipper, Korhonen, Ryant and Palmer 2006), PropBank (Palmer, Gildea and Kingsbury 2005), or FrameNet (Baker, Fillmore and Lowe 1998), or, in the future, bootstrapping verb lexicons in new languages.

Earlier work has generally followed one of two approaches for devising statistical, corpus-based features for use in type-based automatic verb classification.[1] The method of Schulte im Walde (e.g., Schulte im Walde 2000, and later work) uses a single kind of feature – verb subcategorization frames – to capture verb behaviour. These features are very general, in that they apply equally to any verb classes, but are limited in the properties of verb behaviour that they tap into, and are expensive to extract, requiring either a parser or a parsed corpus.[2] By contrast, the approach of Merlo and Stevenson (2001) uses a small collection of features that are specific to the particular class distinctions investigated, requiring expert linguistic analysis to devise. The deeper linguistic analysis allows their feature set to cover a variety of indicators of verb semantics, beyond that of frame information. Moreover, approximations to the deeper properties are devised which require only access to a chunked corpus, thereby avoiding the need for expensive resources that may be unavailable in most languages (see Merlo, Stevenson, Tsang and Allaria 2002).

Our goal is to develop an approach that combines the advantages of each of these earlier verb classification methods. Specifically, we aim to develop a feature space that is *general* – applicable to any class distinctions within the target classification; *broad* – tapping into a variety of semantic features of the classes; and *inexpensive* – requiring no more than a part-of-speech (POS) tagger and chunker.[3]

[1] In type-based classification, a verb is placed into the (existing) lexical semantic class that typically corresponds to its predominant usage, as reflected in statistical features over its occurrences in a corpus. We return to the issue of token-based classification, as in Lapata and Brew (2004), in Section 9 on related work.

[2] Schulte im Walde (2003) explores the additional use of selectional preference features, but they do not generally improve classification performance. In any case, such features are even more resource intensive, since they rely further on an ontology for encoding selectional information.

[3] We assume that chunkers generally rely on less extensive grammatical knowledge than parsers or parsed corpora, since the latter typically require significantly more manual effort in the form of annotation and/or grammar development. Of course, some particular chunker may result from more extensive resource usage than some particular parser, but

Our target lexical semantic classes are those described by Levin (1993) (hereafter Levin). These classes have been incorporated into the computational lexicon of VerbNet (Kipper *et al.* 2000), which has recently been further extended (Kipper *et al.* 2006). Levin classes group together verbs that share both a common semantics (such as manner of motion, or change of state), and a set of syntactic *alternations*. An alternation refers to the alternative mappings of the semantic arguments of a verb to syntactic positions, as in (i):

i. a. *I loaded <u>the truck</u>* ***with hay***.

b. *I loaded* ***hay*** *<u>onto the truck</u>*.

To accomplish our goal of developing general features that tap into a broad range of semantic properties, we devise our feature set based on an analysis of the range of possible alternations that were used in the definition of the Levin classes. We further require that the underlying properties we identify be mapped to statistical features that are easily extractable from a POS tagged and chunked corpus.

Since we analyze verb class distinctions at a general level, we need only do the linguistic analysis once, rather than having to do an individual analysis for every set of classes that we want to distinguish. Moreover, it is worth emphasizing that we do not base our features directly on an analysis of the existing Levin classes themselves (as was done, e.g., in Dorr and Jones 1996). We instead analyze the possible alternations that are the primary determinants of the verb classification. By basing our features directly on the alternations themselves, rather than on the existing classes (which employ a particular combination of alternations), we have devised a general feature space which in principle is useful for any Levin-type verb classification task.[4]

We demonstrate the applicability of the feature space to distinctions among 14 classes, including a total of 835 verbs – a larger scale investigation than has been undertaken in any previous verb classification experiments. (By comparison, Schulte im Walde (2003) uses 168 verbs in 43 classes, while Merlo and Stevenson (2001) use 59 verbs in 3 classes.) To preview our results, on our suite of 11 experimental classification tasks, we achieve reductions in error rate ranging from 48% to 88% over a chance baseline accuracy, averaging 61% over all the tasks, demonstrating the potential of the approach for a wide range of classes. Furthermore, the performance compares favourably with sets of features hand-selected for the particular class distinctions, supporting our claim that a general feature space can avoid time-consuming expert linguistic analysis for each task considered. We also perform a number of experiments designed to identify the most informative features of the feature space, and show that simple counts over syntactic arguments are sufficient in many cases.

The rest of the paper describes the analysis underlying the features in our general feature space, the classes chosen for demonstrating its effectiveness, and

our intention here is to refer to the general case in which chunkers are less resource intensive than parsers.

[4] And indeed, it may be useful for other approaches to predicate classification as well (such as FrameNet), to the extent that their distinctions have expression in syntactic behaviour.

Table 1. *Feature groups with number of features of each type*

§ in text	Feature category	Number of features
2.1	Syntactic slots	77
	Use of pronouns	6
2.2	Slot overlap	41
2.3	Passive	2
	POS of the verb	6
	Aux, modal, Adv	13
	Derived forms	3
2.4	Animacy of NPs	76

our experimental materials, procedures, and results. We conclude with a discussion of related work, as well as limitations and planned extensions of our approach.

2 The feature space

Lexical semantic classes as in Levin (1993) form a shallow hierarchy of groups of verbs with shared meaning. Such classes also share syntactic patterns, because of constraints on the mapping of semantic arguments to syntactic positions (Pinker 1989; Levin 1993). Thus syntactic properties can be useful in verb classification, since they reflect underlying semantic properties of the verbs. In particular, the various alternations of a verb, such as those in example (i) above, in which semantic arguments may appear in different syntactic configurations, are especially useful in distinguishing verb classes. The importance of the differing semantic (thematic) roles assigned by verbs for distinguishing the classes has also been recognized (Merlo and Stevenson 2001).

Our general feature space was designed to tap into the differences between classes both in the alternations they allow, and in the semantic roles assigned. Because we aimed for a set of features that would be applicable to all Levin (or Levin-type) classes, and not just our particular experimental classes, we analyzed the full set of alternations that Levin uses to characterize the verb classes. These alternations, given in Part I of Levin's book, determine the possible distinctions that we need to detect in our classification experiments. For each type of alternation (or set of related alternations), we determined syntactic indicators that would be observable in a corpus. It is worth emphasizing that, in devising our feature space, we did not use the enumeration of the classes themselves, in Part II of Levin. The idea was to determine *possible* class distinctions, rather than distinctions relevant to particular classes, to ensure the generality of the method.

The feature space, which includes 224 distinct features, is summarized in Table 1, and described in detail in the rest of this section. All frequency counts referred to are normalized by the total number of occurrences of the verb (or, in some cases, a relevant subset).

Table 2. *Examples of alternations*

a.	*The ice melted.*	*The children played.*
b.	*The sun melted the ice.*	*The children played chess.*

2.1 Syntactic slot features

The first set of features encodes the frequency of the syntactic positions (what we will refer to as "slots") that potentially contain verbal arguments. These are not subcategorization frame frequencies, since we gather counts over each of the slots – subject, direct and indirect object, and prepositional phrases[5] – independently of their cooccurrence with other slots. (The one exception is subject, which we count overall, as well as separately in its transitive and intransitive usages, because of the importance of transitivity alternations in Levin.) These counts of separate slots are easier to extract than counts of entire syntactic frames; in Section 8 below, we compare the performance of our individual slot features against full subcategorization frame counts to explore their relative informativeness.

Prepositional phrases are treated somewhat differently from other slots, since the particular preposition used is often indicative of the relation of a PP argument to the verb. We include a separate feature for each of 51 high frequency prepositions (those that occur at least 10,000 times in our experimental corpus, the BNC), as well as for 19 groups of closely related prepositions (e.g., one group consists of *between*, *in between*, *among*, *amongst*, *amid*, and *amidst*). See Joanis (2002) for the full list of prepositions and groups. We also maintain a feature for the number of usages of a verb with any prepositional phrase, and a feature for the average number of prepositions per use of a verb.

Levin includes a number of alternations in which a syntactic slot contains a specific word or class of words, in particular pronouns. One case is the use of reflexive pronouns (e.g., *Jill dressed hurriedly/Jill dressed herself hurriedly,* Levin §1.2.3). Another case is the use of semantically empty constituents (*it* and *there*): a contentful argument that in one alternant surfaces as subject or object, instead appears in another position, and an expletive pronoun fills the vacated subject or object slot (e.g., *A problem developed/There developed a problem*, Levin §6.1). To approximate these uses, we add features to count reflexive pronouns in object position; *it* in the subject, intransitive subject and direct object positions; and *there* in the subject and transitive subject positions.

2.2 Slot overlap features

Using syntactic slot information alone misses potentially important properties of alternations, since verbs from different classes may occur in the same syntactic frames but undergo different mappings of their arguments to the positions. For example, consider the sentences in Table 2. Both verbs occur in alternating intransitive (a)

[5] We do not currently consider sentential complements, since few of the verbs included in Levin occur with these.

and transitive (b) syntactic frames. However, the verb *melt* undergoes a causative alternation, in which the semantic argument that appears in the intransitive subject position appears in the transitive object position, while the verb *play* undergoes an unspecified object alternation, in which the semantic argument appearing as the transitive object is optional. To address this, we follow earlier work which has used a measure of overlap over nouns in syntactic slots as an indicator of participation in an alternation (Stevenson and Merlo 1999; McCarthy 2000; Tsang and Stevenson 2004). The idea is that since the same semantic argument may occur in two different slots in a pair of alternating frames, the degree to which those two slots contain the same entities is an indication of the verb's participation in the alternation.

In our feature space, we consider the overlap between each pair of slots that corresponds to an alternation used by Levin to characterize the classes (i.e., §1–8 in Part I of Levin). For each alternation in which a semantic argument occurs in one slot in one usage of the verb, and in a different slot in the alternant usage, we add a feature with the measure of overlap in noun (lemma) usage between those two slots for the verb. For example, given the alternation exemplified by *The sky cleared / The clouds cleared from the sky* (Levin §2.3.5), we add an overlap feature for the subject slot and the object-of-*from* slot. We calculate this overlap using the method of Merlo and Stevenson (2001).[6] When appropriate, we consider prepositions as a group, e.g., when the alternation refers to a general locative or directional specification, rather than to a specific preposition.

2.3 Tense, voice, and aspect features

Verb meaning and alternations also interact in interesting ways with voice, tense, and aspect. For example, Levin (§5) notes several alternations that depend on the passive voice, and so we count passive use of each verb.[7] Some alternations in Levin indirectly depend on the tense of the verb (e.g., the middle voice is usually in the present tense), so we include a set of features that encode the proportion of occurrence of the six POS tags that verbs can take in the Penn tagset: VB, VBP, VBZ, VBG, VBD, and VBN. We also further augment the feature space to count verb uses which occur with an adverb or with each specified auxiliary or modal (as

[6] Let S_X be the set of lemmas occurring in slot X for a given verb, MS_X the multiset of lemmas occurring in slot X, and w_X the number of times lemma w occurs in MS_X. Then we define the overlap between slots A and B as follows:

$$\text{overlap}(A, B) = \frac{\sum_{w \in S_A \cap S_B} \max(w_A, w_B)}{|MS_A| + |MS_B|}$$

For example, if $\{a, a, a, b, b, c\}$ is the multiset of lemmas occurring as subject of a verb, and $\{a, b, b, b, b, d, d, e\}$ is the multiset of lemmas occurring as direct object of the same verb, then overlap(*subject*, *object*) $= (3 + 4)/(6 + 8) = 0.5$. For additional discussion, see Joanis (2002).

[7] Specifically, we include two passive features, which normalize counts of passive use differently: over all verb uses, and over simple past and past/passive participle verb uses only. The latter is for compatibility with Merlo and Stevenson (2001), in which the ambiguity of the simple past and passive participle was relevant to their particular classes under investigation.

well as counting together all occurrences with the use of a modal), since these uses also interact with voice and aspect. In addition, there are alternations that involve derived forms of the verb (Levin §5.4, §7.1,2), hence we include a set of features that measure the frequency of a verb used as a noun (base form only) or as an adjective (in both past participle and present participle forms).

2.4 The animacy feature

We mentioned above that differences in semantic roles assigned to arguments can also distinguish classes of verbs. For example, in order to capture the distinction between Agent and Theme subjects across their target classes, Merlo and Stevenson (2001) used a feature that estimated the proportion of subjects that were animate entities. We generalize this approach to consider the animacy of each of our syntactic slots, since this property is relevant to more than just the Agent/Theme distinction (e.g., Experiencer vs. Stimulus in subject or object positions; Recipient vs. Goal in indirect object or prepositional object positions). We count as animate all personal pronouns other than *it* (following Stevenson *et al.* 1999; cf. Aone and McKee 1996), as well as proper NPs labelled as "person" by our chunker (Abney 1991).[8] Other general feature differences across semantic roles are more challenging to detect automatically (e.g., volitionality or independent existence, from Dowty 1991). We leave to future work further generalization of the idea of distinctive features among semantic roles.

2.5 Comments on estimating the features

Dorr and Jones (1996) showed that 98% of the classes in Levin could be uniquely identified by "perfect knowledge" of the argument frames each class permits. In practice, however, this is not as optimistic a result as it may seem with respect to classifying individual verbs, since many of the alternations mentioned with a particular class in Levin are annotated with the restriction of "most verbs" or "some verbs." Indeed, this is the case for a number of our experimental classes. Clearly, some tendency for the class to undergo an alternation must be captured, rather than a binary decision that holds for all verbs in a class. We address this by using statistical features that indicate proportion of occurrence of a feature.

Moreover, in gleaning information from a corpus as we do, we do not have access to the perfect knowledge used by Dorr and Jones (1996) about which argument frames a verb allows. For example, a verb may simply not appear in the corpus with a particular argument that it allows, or may appear too infrequently to be distinguishable by the use of that argument. Prepositional arguments pose additional difficulties. Even if a verb appears with a preposition that can introduce an argument, we cannot distinguish such an occurrence of the PP from an adjunct use using our

[8] Abney's chunker includes crude named entity recognition: if a proper noun phrase starts with one of the five titles *Mr., Ms., Mrs., Miss.,* or *Dr.*, or if it contains a known first name (from a list of 178 English first names), it is labelled as a person's name, and we count it as animate.

simple processing tools. Our extraction of subjects is also limited, and the detection of indirect objects is highly inaccurate.

Our other features suffer from similar deficiencies. Given the extraction tools we use, the relevant uses of pronouns are indistinguishable from other usages (i.e., unrestricted uses of the reflexives, or non-expletive uses of *it* or *there*). Animacy, too, is clearly a rough approximation, using only pronouns and a subset of proper nouns in the estimates. The overlap measure is a crude approximation to alternation behaviour; detecting alternations, however, is a thorny problem in itself, on which only limited success has been achieved using more resource intensive methods (McCarthy 2000; Tsang and Stevenson 2004). Finally, the tense, voice and aspect features are perhaps the cleanest in terms of accurate estimation, but have the most indirect relation to the class distinctions. In all cases, we assume that the set of features, taken together, will be useful even given a certain level of noise.

3 Experimental verb classes

For the experimental investigation of our feature space, we select a number of verb classes as test cases for evaluating the method. Both for practical reasons and to make the results of general interest, the classes must neither be too small nor contain mostly infrequent verbs. Moreover, we also desire experimental classes that show a range of syntactic and semantic distinctions, to evaluate the robustness of our feature space. That is, our feature space is intended to be capable of distinguishing any Levin-type classes, and so we must choose experimental comparisons that exhibit a variety of class differences.

We therefore manually select pairs (or, in two cases, triples) of classes to represent a range of distinctions that exist among the classes in general. For example, some of the pairs/triples are syntactically dissimilar (which should be easier for our surface-level features), while others show little syntactic distinction across the classes. The pairs/triples also show more or less semantic variability, in terms of their assignment of semantic roles to arguments (which again may influence the effectiveness of our features). These syntactic and semantic distinctions are described along with the description of the pairs/triples of classes below.

Six pairs and two triples form the basis for eight basic classification tasks, using a total of 15 classes. (Note that some classes occur in more than one task.) These are listed in Table 3 at the end of this section, along with their Levin class numbers and the number of verbs in each class. We also form experimental sets including a larger number of these classes, to investigate the ability of the features to distinguish among a wider range of classes in a single classification experiment. The following describes both the basic classification tasks and the multiway tasks created from them.

3.1 Basic classification tasks

Below are examples from each experimental pair/triple of classes, with descriptions of the syntactic and semantic similarities and differences of each.

1) Benefactive verbs versus Recipient verbs.

Mary baked... a cake for Joan/Joan a cake.

Mary gave... a cake to Joan/Joan a cake.

These dative alternation verbs differ in the preposition used and in the semantic role of its object (and the corresponding indirect object): the object of *for* is a Beneficiary, and the object of *to* a Recipient.

2) *Admire* verbs versus *Amuse* verbs.

I admire Jane.

Jane amuses me.

These are the two classes of psychological state verbs. They differ in the mapping of semantic roles to syntactic positions: for *Admire* verbs, the Experiencer argument is the subject and the Stimulus argument is the object; for *Amuse* verbs, the mapping is reversed (Stimulus subject and Experiencer object).

3) *Run* (manner-of-motion) verbs versus Sound Emission verbs.

*Kids ran in the room./*The room ran with kids.*

Birds sang in the trees./The trees sang with birds.

Both classes are intransitive activity verbs; interestingly, the Sound Emission verbs can be used in a manner-of-motion sense (*The truck rumbled down the street*). As shown in the example sentences here, the classes differ in some of the prepositional alternations they allow.

4) Light and Substance Emission verbs versus Sound Emission verbs.

The jewels sparkled./The fountain gushed.

The hinges squeaked.

These are three very similar classes of verbs – subclasses of the Emission class – that allow most of the same alternations, with each allowing one or two alternant forms that the others do not allow (or allow in only limited cases). For example, Light and Substance Emission verbs do not allow the manner-of-motion reading that Sound Emission verbs allow (see the discussion of the classes in task 3 above). Substance Emission verbs participate in the substance/source alternation (*The well gushed valuable oil/Valuable oil gushed from the well*) that Sound Emission verbs do not allow (**The hinges squeaked irritating noise/*Irritating noise squeaked from the hinges*), and only some Light Emission verbs allow (**The jewels sparkled green light/?Green light sparkled from the jewels*, but *The strobe flashed green light/Green light flashed from the strobe*). Because the Light and Substance Emission classes are relatively small, we merge these two classes for the comparison with Sound Emission verbs.

5) *Cheat* verbs versus *Steal* and *Remove* verbs.

*I cheated... Jane of her money/*the money from Jane.*

*I stole/removed... *Jane of her money/the money from Jane.*

These semantically related classes – subclasses of the Verbs of Removing class – differ in the prepositional alternants they allow.

6) *Wipe* verbs versus *Steal* and *Remove* verbs.

Wipe... the dust/the dust from the table/the table.

*Steal/Remove... the money/the money from the bank/*the bank.*

The *Wipe* verbs are another subclass of the Verbs of Removing. The classes here generally allow the same syntactic frames, but differ in the possible semantic role assignment. For example, although both classes allow a transitive frame, the Location argument can appear as the direct object of *Wipe* verbs but not of *Steal* and *Remove* verbs, as shown in the third alternant continuation above.

7) *Spray/Load* versus *Fill* versus Other Verbs of Putting (several related classes).
I loaded... hay on the wagon/the wagon with hay.
*I filled... *hay on the wagon/the wagon with hay.*
*I put... hay on the wagon/*the wagon with hay.*
These three classes – all subclasses of Verbs of Putting – also differ in prepositional alternants. Note, however, that the options for *Spray/Load* verbs overlap with both of the other two types of verbs.

8) Optionally Intransitive: *Run* versus Change of State versus "Object Drop".
The trainer jumped the lion through the hoop/The lion jumped through the hoop.
The chef melted the butter in the pan/The butter melted in the pan.
Mary baked a cake for Joan/Mary baked for Joan.
These are the three classes of Merlo and Stevenson (2001), which we investigate here for comparison to their results. All are optionally intransitive but assign different semantic roles to their arguments. (Note that the Object Drop verbs are a grouping formed by Merlo and Stevenson (2001) and are a superset of the Benefactives in task 1 above.)

These eight basic classification tasks vary in difficulty. As mentioned in Section 2.5, many alternations are noted in Levin's class descriptions as only applying to some of the verbs. This is true for a number of the possible frames for the Emission verbs in task 4, as well as for the *Wipe* verbs in contrast to the other Verbs of Removing in task 6. One might expect these kinds of distinctions to be among the hardest we consider. Since many of these classes are distinguished by PP use, we also expect that the inability (mentioned earlier) to distinguish arguments from adjuncts may cause difficulty in some classification tasks. For example, in task 2, only the *Admire* verbs can take a PP_{for} argument (*I admired Jane for her honesty/*I amused Jane for my/her hilarity*). However, such an argument cannot be distinguished from a PP_{for} adjunct, as in *I amused Jane for the money*. We return to the issue of the relative importance of arguments and adjunct PPs in the discussion of our experimental results.

3.2 Additional multiway tasks

In addition to the eight basic classification tasks described above, we also experiment with larger sets of experimental classes. We add the following three multiway tasks to explore how well our feature space scales to multiple class distinctions:

9) a 6-way task combining the Verbs of Removing (*Cheat*, *Steal–Remove*, and *Wipe*) and the Verbs of Putting (*Spray/Load*, *Fill*, and "Other Verbs of Putting"), all of which undergo similar alternations of locative arguments;

Table 3. *Verb classes (see Section 3), their Levin class numbers, and the number of experimental verbs in each (see Section 4.2)*

Verb class	Levin class number	# of verbs
Benefactive	26.1, 26.3	49
Recipient	13.1, 13.3	33
Admire	31.2	39
Amuse	31.1	157
Run	51.3.2	85
Sound Emission	43.2	63
Light and Substance Emission	43.1, 43.4	41
Cheat	10.6	30
Steal and *Remove*	10.5, 10.1	50
Wipe	10.4.1, 10.4.2	42
Spray/Load	9.7	42
Fill	9.8	65
Other Verbs of Putting	9.1–6	59
Change of State	45.1–4	200
Object Drop	26.1, 26.3, 26.7	64

10) an 8-way task that adds to the above 6-way task the *Run* and Sound Emission verbs, which also undergo locative alternations;
11) a 14-way task including all the classes (except Benefactive, which is a subset of Object Drop).

Of these, the 6-way task has the fewest class distinctions to make, but these are among the hardest given the similarity of the prepositional alternations allowed by these six classes.

4 Experimental materials

4.1 Corpus

To estimate all features, we use the 100M word British National Corpus (BNC, Burnard 2000), which consists of samples of recent British English ranging over a wide spectrum of domains, including both fiction and non-fiction. Since it is a general corpus, we do not expect any strong domain-specific bias in verb usage. The corpus is tagged with the CLAWS POS tagset, which we map to the Penn tagset required by the chunker we use in further processing (that of Abney 1991).

4.2 Verb selection

We selected our experimental verbs as follows. We started with a list of all the verbs in the selected classes from Levin, removing any verb that did not occur at least 100 times in the BNC. Because we assign a single class label to each verb in our experiments, we removed any verb that belonged to more than one of the classes in our selected pairs/triples of classes for experimental comparison. We also removed any verb that we deemed to be overly polysemous (belonging to six or more Levin classes).

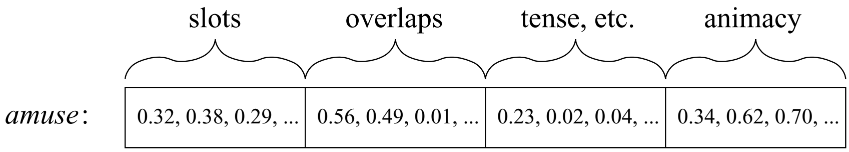

Fig. 1. Sample verb vector with normalized feature values.

Table 3 above shows the number of verbs in each class at the end of this process. Of these verbs, 20 from each class were randomly selected to use as training data.[9] Our unseen test verbs then consist of all verbs remaining after we remove the current training data, as well as any additional training data used in the previous studies we build on – namely Merlo and Stevenson (2001) and Joanis and Stevenson (2003) – since the latter also cannot count as "unseen," even if unused in current training.

4.3 Feature extraction

We use Abney's (1991) chunker, called SCOL, to extract the verb group and syntactic slots on which our features are based. SCOL allows us to extract a subject and the first object (typically the direct object) of a verb with reasonable confidence, and to extract prepositional phrases potentially associated with the verb. However, it does not identify a second object with the verb, which we also require. When a second object appears, it is the direct object, and the first object is an indirect object. We use TGrep2 (Rohde 2002) to identify potential second objects by assuming that when an object is followed by a noun phrase which SCOL has left unattached, a double-object frame is being used. This method of double-object frame identification has very low precision (.22 on a random sample), but we found on development data that it produces useful features despite the high level of noise.

From the extracted information, we calculate all the features described in Section 2, yielding a vector of 224 normalized counts for each verb, which forms the input to the machine learning system. Figure 1 shows an example feature vector for the verb *amuse*.

5 Machine learning method

For all of our experiments, we use Chang and Lin's (2001) LIBSVM library for support vector machines (SVMs), with the default settings recommended by Hsu, Chang and Lin (2003):[10]

- We use the standard C-SVC Cost-Based Support Vector Classification model.
- We use a radial basis function kernel, $e^{-\gamma|u-v|^2}$.

[9] With two exceptions: for the Benefactive and Object Drop classes, the smallest of the classes used in Merlo and Stevenson (2001), we biased our training verb selection to verbs used in their study, to leave more unseen verbs for testing.

[10] In previous work, we used the C5.0 decision tree induction system, but in preliminary experiments for this work, we found that SVMs generally outperformed C5.0 on our tasks.

- For tasks involving more than two classes, we perform multiway classification via the 'one-against-one' methodology: we build a set of binary classifiers to compare each pair of classes and determine the global winner via majority vote.[11] (In the case of a tie, the winner is selected arbitrarily, following Hsu and Lin 2002.)

All the 2-way and 3-way experiments have balanced training sets, with 20 verbs per class. In all cases, each verb has a single correct classification within the 2 or 3-way task even if it is otherwise ambiguous. However, we combine data from several 2-way and 3-way experiments to create the multiway experiments, in which some verbs then have more than one possible class. To avoid ambiguity, we exclude any verb from a multiway task if it occurs in more than one class in that task. After removing such verbs, we are left with unbalanced training sets. We correct for this imbalance by adjusting the penalties for incorrect classification so that our trained classification model has a uniform prior distribution of classes in all tasks.

For each classification task and feature set, we take the following three steps:

I. Preprocess Training Data:

We designate a feature as 'missing' for a particular verb when its calculation involves division by zero. We handle a missing value by replacing it with the 60% trimmed mean value of that feature for all data points in the current task – that is, we remove the top 30% and the bottom 30% of values, then calculate the mean on the remaining data points for that feature.

In SVM training, features with greater variance can dominate the classification (Hsu *et al.* 2003); we thus scale all features to have similar variance. We adjust the location and scale of our features by subtracting the 60% trimmed mean and dividing by the mean absolute deviation. To reduce the influence of outliers, we also transform our features using an arctan transformation, which reduces the range of extreme values while having little effect on others (Sarle 2002). We use trimmed mean and mean absolute deviation as our estimators of location and scale because they are robust to skewness and outliers (Iglewicz 1983). Finally, we removed any features whose value was constant over all training verbs, since these cannot provide any information to the classifier.

II. Select Optimal Parameters

LIBSVM uses two input parameters: a regularization parameter, C, and a scaling parameter for the kernel, γ. We consider a similar range of values to those recommended by Hsu *et al.* (2003):

$C = 2^{-5}, 2^{-3}, \ldots, 2^{15}, 2^{17}$

$\gamma = 2^{-17}, 2^{-15}, \ldots, 2^{1}, 2^{3}$

For each task and feature set, we select the combination of C and γ that offer the best accuracy on the preprocessed training set according to 10-fold cross-validation

[11] Rifkin and Klautau (2004) show that the 'one-against-one' and the 'one-against-all' methodologies have equivalent empirical performance, but Hsu and Lin (2002) show that the former is implemented more efficiently in the LIBSVM package, so we retain it for this practical reason.

Table 4. *Primary experimental results*

	Experimental task	Baseline %acc[a]	*N* test verbs	All features %acc[a]	Levin-derived %acc[a]
1)	Benefactive/Recipient	50.0	33	86.4	84.1
2)	*Admire/Amuse*	50.0	152	93.9	87.0
3)	*Run*/Sound Emission	50.0	90	86.8	85.7
4)	Light,Subst/Sound E.	50.0	60	75.0	67.5
5)	*Cheat/Steal–Remove*	50.0	36	76.5	62.7
6)	*Wipe/Steal–Remove*	50.0	46	80.4	86.0
	Average (2-way tasks)	50.0		83.2	78.8
7)	*Spray/Fill*/Putting	33.3	101	65.6	67.7
8)	Optionally Intrans.	33.3	222	74.2	69.0
	Average (3-way tasks)	33.3		69.9	68.4
9)	6 Locative Classes	16.7	136	63.1	57.0
10)	8 Locative Classes	12.5	215	61.7	58.5
11)	All 14 Classes	7.1	496	58.4	56.6
	Average (≥6-way tasks)	12.1		61.1	57.4

[a] %acc is per cent macro-averaged accuracy.

with 100 repeats. (We use stratified cross-validation, in which the 10 subsets of data have the same proportion of class members as the full dataset.)

III. Train and Test SVM Classifier

For each task (and feature set), we train a classifier on the full set of preprocessed training data and evaluate its performance on the unseen test data, to which the same preprocessing has been applied. (That is, missing features in the test data are replaced with the 60% trimmed mean calculated on the training data, and all features are then transformed using the same training data trimmed mean, mean absolute deviation, and arctan function.)

Since we use balanced training sets and a uniform random baseline, our accuracy metric should assign equal weight to each class. We therefore report macro-averaged accuracy, which assigns equal weight to each class, regardless of the skew in the distribution of test verbs.[12]

6 Experimental results using the full feature space

6.1 Primary experimental results

We perform eleven classification tasks: eight 2-way and 3-way tasks (Section 3.1) and three multiway tasks formed from those (Section 3.2). The experimental tasks are shown in the first column of Table 4. Note that we do not create one classifier which is applied to different test data sets; rather, a separate classifier is defined for each task using only the training data for the verb classes under investigation. Given

[12] Some authors refer to macro-averaged accuracy as average per-class accuracy or macro-averaged recall.

that there is no recognized informed baseline for the task of verb classification, we (as in work by others) use a chance baseline for comparison of our results. Since in all cases the prior probability is uniform (see Section 5), the baseline (chance) performance is $1/k$ for a task discriminating k classes. This baseline is shown in the second column of Table 4, while the third column gives the number of test verbs for each task.

Our first set of experiments uses all our features (as listed in Table 1); the results are shown in the fourth column of Table 4. In all cases, test performance shows a substantial improvement, of 25% to 51%, over the baseline, with a reduction in error rate ranging from 48% to 88%. The classifiers perform quite well at all tasks, and even the 14-way task far exceeds the baseline. Indeed, it is worth noting that even tasks that we predicted to be very difficult (the Emission verbs and the *Wipe/Steal–Remove* classes) had good overall performance, of 75.0% and 80.4%, respectively (tasks 4 and 6 in the table).

We set out to show that our general feature space could perform comparably to features manually devised for the specific classification task by linguistic experts. For one task, we can make such a comparison – the three optionally intransitive classes investigated in Merlo and Stevenson (2001). On our new test verbs in this task, our feature space outperformed their hand-crafted features, 74.2% (task 8 in the table) versus 57.8%.[13] This result reveals a potential advantage of the general feature set, which includes a broad range of indicators of syntactic behaviour. These indicators may help to discriminate classes even if they are not conceived of as core distinctions between the classes.

For the other classification tasks, however, no manually derived feature space exists. We instead compare the general feature space to subsets of our own features (called the Levin-derived subsets) which are hand-selected through an analysis of the classes in Levin. For each class, we systematically identify the subset of features indicated by the class description given in Levin. To do this, we note each alternation mentioned by Levin for the class (both grammatical and ungrammatical ones) and add features corresponding to those alternations. (Features corresponding to grammatical alternations are expected to have relatively high values, and those corresponding to ungrammatical alternations are expected to have relatively low values.) For each experimental task (comparison of a set of classes), the Levin-derived subset is the union of these subsets of features for all the classes in the task. As an example, Table 5 lists the Levin-derived subset of features thus selected for the Benefactive and Recipient classes; the union of these features is used in the task which distinguishes these two classes.

Note that this approach of selecting a subset of our features based on a manual analysis of the classes does not address the possibility that a linguist may devise a feature for which we have no analogue. However, it does capture the important

[13] Note that applying C5.0 – the learner used by Merlo and Stevenson (2001) – to the same new verbs in this task yields a somewhat smaller difference in performance. Using our feature space, C5.0 yields an accuracy of 68.9%; with the features of Merlo and Stevenson (2001), it achieves 59.3%.

Table 5. *Levin-derived subsets for Benefactive and Recipient classes*

Class	Levin-derived subset of our feature space
Benefactive verbs	**Slot features**: dir. object, ind. object, PP$_{for}$, PP$_{from}$, PP$_{into}$, PP$_{out\ of}$; **overlap features**: (subject, dir. object), (intr. subject, dir. object), (trans. subject, intr. subject), (trans. subject, PP$_{from}$), (trans. subject, PP$_{out\ of}$), (dir. object, PP$_{from}$), (dir. object, PP$_{into}$), (dir. object, PP$_{out\ of}$), (ind. object, PP$_{for}$); **animacy features**: subject, trans. subject, dir. object, ind. object, PP$_{for}$, PP$_{into}$, PP$_{out\ of}$.
Recipient verbs	**Slot features**: dir. object, ind. object, PP$_{to}$, PP$_{with}$, PP-group$_{behind}$, PP-group$_{near}$;[a] **overlap features**: (subject, dir. object), (intr. subject, dir. object), (dir. object, PP$_{with}$), (ind. object, PP$_{to}$); **animacy features**: subject, trans. subject, dir. object, ind. object, PP$_{to}$.

[a] As described in Section 2.1, we treat some highly related prepositions as a group.

challenge faced by our method in having to automatically select the relevant features for a task from our large feature space.

The results for the Levin-derived feature sets are given in the fifth column of Table 4. Comparing the test accuracies in the fourth and fifth columns, we see that the general feature space performs better than the Levin-derived subsets on most tasks. The performance benefit of the general feature space is an average of 3.5% across all tasks. The full feature space outperforms the subsets by 5% or more on 5 of the 11 tasks, while the subsets show a performance advantage of 5% or more on only one task.

6.2 Discussion of results using the full feature space

The above results taken together demonstrate that our general feature space can be used successfully for automatic verb classification. Moreover, the results support our hypothesis that the need for time-consuming expert analysis for each new classification task can be reduced through the careful design of general features based on a higher level analysis of the class structure.

It is also worth pointing out that it is not simply the number of features that is the key to successful classification. Although the Levin-derived sets contain 18 to 74 features (compared to 224 in the full feature set), in these and our remaining experiments, we find no consistent correlation between performance and number of features used in the task.

7 Experiments by feature groups

While the results above demonstrate that our feature space is effective, it is important also to determine which parts of it contribute the most useful information. For this purpose, we divide our feature space into the four major feature groups, summarized in Table 6.

Table 6. *Feature groups*

Group	Number of features	Description
Slots	83	Syntactic slots and use of pronouns (§2.1)
Overlaps	41	Slot overlap features (§2.2)
Tense	24	Tense, voice and aspect features (§2.3)
Animacy	76	Animacy features (§2.4)

7.1 One feature group at a time

For each feature group and each of our eleven tasks, we follow the methodology described above, but using only the features in that group. This set of experiments allows us to assess how much information each group of features provides for the classification tasks. The results, shown in Table 7, indicate that the slot features on their own perform on average very close to the full feature space. Although there is some loss in performance, only in two tasks is the decrease 5% or more. This result shows that the mapping of semantic arguments to syntactic positions, noted by Pinker (1989) and others, and used by Levin (1993) as the basis for her classes, is reflected to a high degree in statistics over the slots occurring with a verb. This result is encouraging for the ease of use and portability of our method, since these features are relatively simple (compared to the overlap and animacy features) and have ready analogues in other languages (i.e., either the same kind of slot features, or perhaps features that indicate the case assigned to NPs).

The other feature groups, although outperforming the baseline (with only one exception), prove to be much less informative to the machine learning system. Even the overlap features in general do not perform very well on their own, though they are specifically intended to capture alternations – the basis of Levin's classification, and therefore the basis of our experimental classes. One exception is that the overlap features do well for the Emission verbs (task 4); given that these classes allow mostly the same alternations, this may indicate that perhaps some difference in the frequency of alternation use is being detected. We can also see that the tense, voice and aspect features do reasonably well for the *Wipe* vs. *Steal–Remove* task (task 6). Levin notes that use of the gerund is characteristic of *Wipe* verbs with an unspecified object (*Brian was wiping*), so perhaps this distinctive behaviour is detected by the tense features. This feature group also does well for the Putting verbs (task 7), but here the role they play is less obvious.

Generally, the feature groups other than slots mostly perform the same for the multiway experiments, but performance varies more in the 2-way and 3-way tasks. These features may get at finer distinctions among minimal pairs of classes, with their contribution swamped by the slot features when many more classes are involved.

7.2 Removing each feature group

The experiments described above tell us how much each feature group can accomplish on its own, but tell us little about how much information each group

Table 7. *Results by feature groups: each feature group used alone*

					Only the features in group			
	Experimental task	Baseline %acc[a]	*N* test verbs	General f. space %acc[a]	Slots %acc[a]	Overlaps %acc[a]	Tense %acc[a]	Animacy %acc[a]
1)	Benefactive/Recipient	50.0	33	86.4	84.1	65.9	72.7	75.0
2)	*Admire/Amuse*	50.0	152	93.9	91.1	75.8	78.8	81.3
3)	*Run*/Sound Emission	50.0	90	86.8	82.7	60.4	64.4	70.1
4)	Light,Subst/Sound E.	50.0	60	75.0	70.0	72.5	57.5	56.2
5)	*Cheat/Steal–Remove*	50.0	36	76.5	70.8	55.4	54.6	40.0
6)	*Wipe/Steal–Remove*	50.0	46	80.4	82.9	66.5	75.4	58.1
	Average (2-way tasks)	50.0		83.2	80.3	66.1	67.2	63.4
7)	*Spray/Fill*/Putting	33.3	101	65.6	65.6	50.8	62.3	43.1
8)	Optionally Intrans.	33.3	222	74.2	74.2	56.8	62.3	65.7
	Average (3-way tasks)	33.3		69.9	69.9	53.8	62.3	54.4
9)	6 Locative Classes	16.7	136	63.1	59.7	34.1	33.8	33.0
10)	8 Locative Classes	12.5	215	61.7	63.4	32.8	36.2	31.3
11)	All 14 Classes	7.1	496	58.4	57.8	26.3	30.4	28.2
	Average (≥6-way tasks)	12.1		61.1	60.3	31.1	33.5	30.8

[a] %acc is per cent macro-averaged accuracy.

contributes to performance in combination with other features. For each feature group, we also ran our set of experiments with all but the features in that group.

The results, shown in Table 8, confirm the predominant and crucial role of the syntactic slot features. Removing the slot features decreases performance by at least 10% in 5 of the tasks, and by at least 5% in 4 more tasks. The average decrease over all tasks is more than 10%. Removal of the overlap features decreases performance by at least 5% in 3 tasks, and removal of tense features in only one task. Removal of animacy features has an even smaller influence on performance on individual tasks.

In looking at the averages across 2-way, 3-way, and multi-way tasks, removal of slots decreases performance by about 8%, 7%, and 16%, respectively. Removal of any other feature group yields less than 3% difference in average performance in all cases. This shows that for the feature groups other than slots, any decrease in performance arising from their removal in some task is balanced by an increase in performance in another task.

7.3 Zooming in on prepositions

Given the importance of the slot features and the preponderance of prepositional slots, we hypothesized that prepositions might play a central role among our features. We investigated the impact of prepositions further by creating the "Prepositions" subgroup, which contains only the 72 preposition features. We ran our eleven tasks with only the preposition features and also with all but the preposition features. Table 9 shows the results of these experiments, with the results also for the slot features alone and all but the slot features repeated for easier comparison. (Recall that the preposition features are also included in the slot features.)

For the 2-way tasks, prepositions alone are sometimes a bit better and sometimes a bit worse than the entire slot group, but overall the performance using only prepositions is similar to that of the slot features, and only somewhat below that of the entire feature space. In contrast, most of the larger tasks clearly benefit from the other slot features, with substantial reduction in accuracy when only prepositions are used. For some tasks (such as distinguishing the 3 optionally intransitive classes, where transitivity is crucial), having the frequencies of subject and object slots is very informative.

7.4 Discussion of feature group results

These experiments demonstrate that a set of easily extractable features, namely the estimates of syntactic slot frequencies, enable us to obtain good results in English verb classification. The features justified by a deeper linguistic analysis, such as the overlap and animacy features, contribute little to the overall performance of our feature space. However, every feature group makes a contribution (sometimes substantial) to at least some of the tasks, and we cannot rule out the potential role of the features other than slots for other verbs and verb classes than those investigated here.

Table 8. *Results by feature groups: all features but the group*

					All features except those in group			
Experimental task		Baseline %acc[a]	*N* test verbs	General f. space %acc[a]	Slots %acc[a]	Overlaps %acc[a]	Tense %acc[a]	Animacy %acc[a]
1)	Benefactive/Recipient	50.0	33	86.4	75.0	84.1	84.1	86.4
2)	*Admire/Amuse*	50.0	152	93.9	91.6	95.4	94.6	94.2
3)	*Run*/Sound Emission	50.0	90	86.8	80.0	84.7	88.4	85.7
4)	Light,Subst/Sound E.	50.0	60	75.0	73.8	68.8	77.5	76.2
5)	*Cheat/Steal–Remove*	50.0	36	76.5	55.8	67.7	71.5	76.5
6)	*Wipe/Steal–Remove*	50.0	46	80.4	71.0	82.9	80.4	82.9
	Average (2-way tasks)	50.0		83.2	74.5	80.6	82.8	83.6
7)	*Spray/Fill*/Putting	33.3	101	65.6	60.1	67.2	62.1	64.6
8)	Optionally Intrans.	33.3	222	74.2	66.0	70.8	73.6	72.9
	Average (3-way tasks)	33.3		69.9	63.0	69.0	67.8	68.8
9)	6 Locative Classes	16.7	136	63.1	43.3	56.7	60.7	60.9
10)	8 Locative Classes	12.5	215	61.7	50.6	61.7	62.4	59.5
11)	All 14 Classes	7.1	496	58.4	42.1	59.7	55.8	55.0
	Average (≥6-way tasks)	12.1		61.1	45.3	59.4	59.6	58.5

[a] %acc is per cent macro-averaged accuracy.

Table 9. *Results by feature groups: zooming in on prepositions*

					Subset of features			
	Experimental task	Baseline %acc[a]	*N* test verbs	General f. space %acc[a]	Slots %acc[a]	Preps %acc[a]	All but Pr. %acc[a]	All but Sl. %acc[a]
1)	Benefactive/Recipient	50.0	33	86.4	84.1	90.9	81.8	75.0
2)	*Admire/Amuse*	50.0	152	93.9	91.1	86.2	90.3	91.6
3)	*Run*/Sound Emission	50.0	90	86.8	82.7	83.9	79.0	80.0
4)	Light,Subst/Sound E.	50.0	60	75.0	70.0	70.0	75.0	73.8
5)	*Cheat/Steal–Remove*	50.0	36	76.5	70.8	67.7	70.8	55.8
6)	*Wipe/Steal–Remove*	50.0	46	80.4	82.9	81.0	76.5	71.0
	Average (2-way tasks)	50.0		83.2	80.3	80.0	78.9	74.5
7)	*Spray/Fill*/Putting	33.3	101	65.6	65.6	65.6	56.1	60.1
8)	Optionally Intrans.	33.3	222	74.2	74.2	62.8	70.1	66.0
	Average (3-way tasks)	33.3		69.9	69.9	64.2	63.1	63.0
9)	6 Locative Classes	16.7	136	63.1	59.7	55.4	47.6	43.3
10)	8 Locative Classes	12.5	215	61.7	63.4	55.0	46.5	50.6
11)	All 14 Classes	7.1	496	58.4	57.8	52.6	47.5	42.1
	Average (≥6-way tasks)	12.1		61.1	60.3	54.3	47.2	45.3

[a] %acc is per cent macro-averaged accuracy.

It is somewhat surprising that the slot features alone perform as well as they do. We noted in discussing limitations of the features (see Section 2.5) that, although perfect knowledge of argument frames is sufficient for determining a class, the slot features are only a noisy approximation to such knowledge. One possibility is that this weakness of slot features is actually one of their strengths: by including all the constituents that occur with a verb (at least, as detected by the chunker we use), the slot features capture not only the core arguments of a verb, but its pattern of usage with adjunct PPs as well. For example, Schulte im Walde (2003) found that using all PPs, rather than just argument PPs, improved performance of clustering German verbs; this indicates that adjunct PPs may be highly informative in distinguishing the fine-grained semantics of verbs. However, in further experimentation, we do not see support for this explanation for the good performance of our slot features; we return to this point in Section 8.3.

Features related to preposition usage comprise the majority of the slot features – 72 out of 83. While we find that the prepositional features alone are comparable to the full set of slot features in the 2-way tasks, this is not the case for the 3-way and multiway experiments. Although few in number, the slot features other than prepositional features tap into alternation behaviour that is important to at least some verb class distinctions.

We conclude that using the full set of slot features is warranted, and even the inclusion of other relatively inexpensive features (such as the tense group and a version of animacy using pronoun estimates) is likely to increase the generalizability of the feature space to novel verbs and classes.

8 Experiments using subcategorization frames

Our approach was founded on the principle of devising features that would be easily extractable. We wanted to avoid expensive resources (such as parsers or parsed corpora that rely on extensive grammars), so that the method could be applied in languages for which such tools or corpora are unavailable. One consequence of this stance was that we gathered counts over individual syntactic slots extracted using a chunker, rather than determining more sophisticated features such as subcategorization frame distributions. We have shown that our slot features alone perform almost as well as the full feature space. In this section, we investigate the performance of the slot features compared to subcat frame distributions, to see if the latter are more informative in classification.

8.1 Subcat distribution features

We collect subcat frame statistics for all our verbs as follows. The BNC is processed using Briscoe and Carroll's (1993) Robust Accurate Statistical Parsing (RASP) system and Briscoe and Carroll's (1997) subcat frame extraction software, to obtain a distribution over the subcat frames for each verb.[14] Of the 163 possible subcat

[14] We thank Anna Korhonen, University of Cambridge, for providing this software, and Vivian Tsang, University of Toronto, for running it on the BNC.

Table 10. *Results using Briscoe and Carroll's (1997) subcat distributions*

	Experimental task	Baseline %acc[a]	*N* test verbs	Slots %acc[a]	Subcat frames %acc[a]
1)	Benefactive/Recipient	50.0	33	84.1	65.9
2)	*Admire/Amuse*	50.0	152	91.1	73.1
3)	*Run*/Sound Emission	50.0	90	82.7	76.8
4)	Light,Subst/Sound E.	50.0	60	70.0	56.2
5)	*Cheat/Steal–Remove*	50.0	36	70.8	60.8
6)	*Wipe/Steal–Remove*	50.0	46	82.9	75.2
	Average (2-way tasks)	50.0		80.3	68.0
7)	*Spray/Fill*/Putting	33.3	101	65.6	41.5
8)	Optionally Intrans.	33.3	222	74.2	62.5
	Average (3-way tasks)	33.3		69.9	52.0
9)	6 Locative Classes	16.7	136	59.7	40.1
10)	8 Locative Classes	12.5	215	63.4	42.0
11)	All 14 Classes	7.1	496	57.8	31.8
	Average (≥6-way tasks)	12.1		60.3	38.0

[a] %acc is per cent macro-averaged accuracy.

frames described in Briscoe and Carroll's (1997) subcat dictionary, 103 are identified in this approach as occurring at least once in the BNC, giving us a fairly detailed distribution of subcat frames for our verbs.

We use the resulting subcat distributions as features for each of our eleven tasks, applying exactly the same methodology as before. As we can see in Table 10, the subcat distributions perform substantially worse than the individual slot features. Counting full subcategorization frames requires sophisticated tools – a parser and complex subcat frame extraction software in this experiment – but this heavy machinery does not help in our tasks. These results confirm earlier findings by Sarkar and Tripasai (2002) that parsing is not necessary for the estimation of useful features in classifying verbs. This suggests that it is worth exploring similar types of easily extractable features in other languages as well, rather than assuming that sophisticated processing tools are necessary to make headway in verb classification.

However, one possible explanation for the poor performance of the subcat frames described above is that they ignore the identity of particular prepositions used in the frames. In Section 7.3, we showed that preposition information is a very important contributor to the performance of our general feature space. Moreover, Schulte im Walde (2003) demonstrated that when using subcat frames in clustering German verbs, the identity of prepositions played an important role. However, we cannot simply add the specific preposition information to Briscoe and Carroll's (1997) subcat frames, for the same reason that they do not include it: taking it into account in all the frames where prepositions occur would yield thousands of different frames to consider, creating sparse data problems.[15]

[15] It is feasible for Schulte im Walde (2003) to add preposition information to her subcat frames, because she uses fewer subcat frames (39) and fewer prepositions (30) than we consider.

8.2 Simple subcat features with preposition information

In order to determine the potential of a method that combines subcat frame information (as opposed to individual slot information) with knowledge of specific prepositions, we instead use a set of very coarse subcat frames that we can obtain from the chunker we use. These frames include S-V, S-V-O, S-V-O-O, S-V-PP, S-V-O-PP, and S-V-O-O-PP (S = subject, V = verb, O = object, PP = prepositional phrase). Note that we take only the first PP that occurs with the verb, on the assumption that subsequent PPs are much more likely to be adjuncts.[16] Given this set of 6 basic subcat frames, we experiment on our 11 classification tasks with different levels of refinement for further specifying the PP in the 3 frames that include one. In all cases, the values for a verb across the set of subcat frames forms a probability distribution (i.e., each occurrence of the verb is considered to occur in exactly one of the frame possibilities).

- In the first set of experiments, “SCF-Basic”, we use only the 6 initial subcat frames above as the feature set, with no information about the particular preposition used in a PP.
- In the second set of experiments, “SCF-PP-Groups”, we subdivide PPs into 39 different cases based on the preposition. We first consider the 19 preposition groups we used in our general feature space; we then take the 19 prepositions that occur at least 10,000 times in the BNC and are not already contained in one of the groups. An annotation of “other” is used for prepositions not covered by these choices. Each subcat frame with a PP is annotated with one of these 39 designations. With the three frames that do not include a PP, this yields 120 subcat features.
- The final set of experiments, “SCF-PP-Individual”, first considers individually all 51 prepositions which occur at least 10,000 times in the BNC, and then uses the groups only for remaining lower-frequency prepositions. In this case we have a choice of 51 individual prepositions, 9 groups, and “other”, for further annotation of frames with a PP. Adding the three frames that don’t include a PP, we obtain 186 subcat features.

As we see in Table 11, the results confirm our findings (and reinforce those of Schulte im Walde 2003) that knowledge of prepositions – whether by group or individually – contributes crucial information in verb classification. As expected, SCF-Basic, which uses simple subcat frames with no individuated PP knowledge, has very poor performance. The SCF-PP-Groups and SCF-PP-Individual results are both about the same as the results we obtain with the slot features, even though knowledge of prepositions in the SCF features is limited to that of the first PP occurring with the verb. (Recall that we include all PPs occurring with a verb in the extraction of our slot features.) Interestingly, SCF-PP-Groups performs similarly

[16] Of course, adjunct PPs may be useful as well, as we discuss below, but in this particular experiment we are trying to include only likely arguments, as in Briscoe and Carroll’s (1997) subcat frames. A future area of exploration would be to try to determine the relative contribution of argument versus adjunct PPs to verb classification.

Table 11. *Results using our simple subcat distributions*

					Simple subcat distributions		
Experimental task		Baseline %acc[a]	*N* test verbs	Slots %acc[a]	SCF-Basic %acc[a]	SCF-PP-Gr. %acc[a]	SCF-PP-Ind. %acc[a]
1)	Benefactive/Recipient	50.0	33	84.1	78.3	79.6	84.1
2)	*Admire/Amuse*	50.0	152	91.1	63.4	84.7	82.7
3)	*Run*/Sound Emission	50.0	90	82.7	71.1	85.3	81.2
4)	Light,Subst/Sound E.	50.0	60	70.0	52.5	73.8	73.8
5)	*Cheat/Steal–Remove*	50.0	36	70.8	56.9	75.8	71.9
6)	*Wipe/Steal–Remove*	50.0	46	82.9	71.5	83.5	83.5
	Average (2-way tasks)	50.0		80.3	65.6	80.4	79.5
7)	*Spray/Fill*/Putting	33.3	101	65.6	42.9	73.5	64.7
8)	Optionally Intrans.	33.3	222	74.2	47.8	65.4	75.8
	Average (3-way tasks)	33.3		69.9	45.4	69.4	70.2
9)	6 Locative Classes	16.7	136	59.7	37.3	68.5	62.0
10)	8 Locative Classes	12.5	215	63.4	38.6	65.7	63.2
11)	All 14 Classes	7.1	496	57.8	25.4	54.2	58.1
	Average (≥6-way tasks)	12.1		60.3	33.8	62.8	61.1

[a] %acc is per cent macro-averaged accuracy.

to SCF-PP-Individual, which indicates that the information summarized by our preposition groups is useful in our tasks.

8.3 Discussion of subcat frame results

We started this section by asking whether more sophisticated subcategorization frame features could outperform the slot features found to be the most informative of our feature space. Although features encoding subcat distributions drawn from a parsed corpus are shown to be inferior to our slot features, we speculate that this is due to a lack of knowledge of the prepositions used. Our alternative subcat features, based on simple subcat frames augmented with knowledge of the first preposition used, perform roughly equivalently to our original slot features, which include all PPs used with a verb.

In Section 7.4, we suggested that a possible reason for the effectiveness of our slot features may be that they capture not only argument usage (used in the definition of the classes), but information about adjuncts occurring with the verbs as well. While adjuncts are not used in characterizing the classes, it is possible they may contain highly distinguishing information correlated with semantic class. However, our results in this section suggest that detection of adjunct usage patterns is not the reason for the success of the slot features – or if it is, at least it is not a necessary factor in achieving similar performance. Our simple subcat frames count only the first PP used with an occurrence of a verb, eliminating most adjuncts from consideration.

9 Related work

Much research over the past decade has investigated means for automatically learning the complex knowledge about verbs needed for machine text analysis. As an example of the importance of this topic, a recent international workshop was devoted to work on determining the nature, representation, learning and use of verb class information (Erk, Melinger and Schulte im Walde 2005). Moreover, the classes of Levin (1993) – in their original conception, and in their instantiation in the computational lexicon of VerbNet – have become a standard lexical resource in NLP (e.g., Shi and Mihalcea 2005; Swift 2005). However, reliable automatic verb classification has remained an elusive goal.

Dorr and Jones (1996) show that perfect knowledge of the allowable syntactic frames for a verb enable an accuracy of 98% in assigning verbs to Levin classes. Following the work of Merlo and Stevenson (2001) and Schulte im Walde (2000), we instead approximate such knowledge through statistical corpus analysis, allowing for easier extensibility to new classes. Furthermore, rather than using a class-by-class analysis as in Dorr and Jones (1996) or Merlo and Stevenson (2001), our features are determined through an analysis of the possible alternations for verbs independent of their class assignment. This leads to a more general set of features, as in Schulte im Walde (2000), but unlike hers, ours do not depend on an expensive resource such as a full parser.

Our study generalizes the slot overlap feature of Merlo and Stevenson (2001) as an indicator of membership in a verb class, on the assumption that slot overlap is correlated with alternation behaviour. Indeed, McCarthy (2000) confirms this relation, showing that a slot overlap feature similar to ours can be a useful indicator for a given alternation. However, our overlap features, while useful in some tasks, are not highly informative overall. McCarthy achieves an improvement in alternation detection by using an alternative similarity measure over the content of the two slots that alternate, and Tsang and Stevenson (2004) present a method that further improves such an approach. These results indicate that other formulations of slot similarity should be pursued in future verb classification work, to extend the usefulness of overlap features.

Schulte im Walde (2000, 2003) and Schulte im Walde and Brew (2002) achieve promising results using unsupervised clustering of verbs in English and German, according to general syntactic frame statistics. Also using unsupervised learning, Oishi and Matsumoto (1997) cluster Japanese verbs automatically into hundreds of semantic classes, which they then combine into a network of 38 classes using linguistic knowledge and semi-automated processing. Their approach, like ours, uses a combination of syntactic frame and aspectual features, but a limited set. Mayol, Boleda and Badia (2005) also use a small number of features based on linguistic analysis in clustering Catalan verbs. They obtain very good results (an F-measure of .87, on a task with a baseline of .65), but the target classes are syntactic, rather than semantic as in ours and others' work noted here. In any case, our work has aimed to extend this type of hand-picked feature space to achieve a generality that will limit the need for human expert input in devising features for individual classification tasks.

Our results here have used supervised learning. On similar experimental tasks, using our general feature space, Stevenson and Joanis (2003) achieve mean accuracies of .73, .53, and .38, respectively on the 2-way, 3-way, and multiway tasks, using a semi-supervised approach based on a small seed set of verbs. These results show the potential of general features such as ours and Schulte im Walde's to contribute to verb classification and discovery, although the decrease in performance compared to supervised methods shows that some human intervention will likely remain a necessity. Indeed, in adding new classes to extend Levin's classification and VerbNet, Korhonen and colleagues took a semi-automatic approach which first manually identifies new alternations and potential classes, and then semi-automatically determines when certain criteria are met for including the classes and the verbs belonging to them (Korhonen and Briscoe 2004; Kipper *et al.* 2006). To create useful classes, unsupervised verb class discovery will require some sort of manual identification of linguistically relevant properties to guide the process.

Other unsupervised work has gone beyond the task of verb clustering to that of verb argument structure induction. Rooth, Riezler, Prescher, Carroll and Beil (1999) use an EM-based technique to induce a lexicon which groups verbs into classes that reflect the mappings of semantic arguments to syntactic positions. Gildea (2002) extends their probabilistic model by explicitly modelling alternation behaviour, but finds that the Rooth *et al.* (1999) model, which indirectly captures verb alternation information, performs best of five EM-based models that he tests. Although both of these pieces of work are limited in the semantic roles and alternations considered, they highlight the importance of moving beyond verb classification and clustering, to automatically filling in additional syntactic and semantic information for the acquired classes.

Moving beyond type-based verb classification, recent research has also explored token-based verb classification, assigning a class label to individual verb instances (Lapata and Brew 2004; Girju, Roth and Sammons 2005). In other work, Swier and Stevenson (2004) use probabilistic knowledge of a verb's possible classes to help guide semantic role labelling of its arguments. Both of these tasks assume that the possible classes for a verb are known. Thus far, our approach is restricted to providing a single predominant class for a verb. Future work must extend this notion to fuzzy or probabilistic assignment of verbs to classes, in order to adequately support the differential classification of verb instances. However, even the current approach of assigning a single class label intended to capture the predominant usage of a verb may prove useful, in analogy to work on predominant sense classification, which has been shown to be a useful component of word sense disambiguation (McCarthy, Koeling, Weeds and Carroll 2004).

10 Conclusion

Complex information about verbs is required to support semantic processing of predicates and their relation to constituents in a sentence, i.e., to determine "who did what to whom". Automatic verb classification is a means for leveraging expensive hand-coding of such information, but appropriate features must be devised and must be demonstrated to be useful. We develop a general feature space based on a high-level analysis of the verb classes in Levin (1993) (a standard NLP resource), mapping linguistic properties of the classes to features extractable from a chunked corpus. These general and easily extractable features achieve very good accuracies in verb classification across a range of class distinctions. By basing our features on the possible argument alternations that define the classes, rather than on specific classes, we aim our feature space to be useful not only for those verb classes already identified, but for any additional classes that allow for the same alternations. We also believe it will be straightforward to extend the feature space to cover additional alternations not in the scope of Levin (such as those involving sentential arguments), through a generalization of the existing features to new syntactic positions.

We have performed a wide range of experiments to identify which types of features are most informative in the verb classification task. Our conclusion, based on our full suite of experiments, is that the syntactic information about core constituents occurring with a verb is most important to verb classification. Moreover, knowledge of the frequent prepositions used, and of groups of less frequent prepositions, is essential to the success of such features. However, we reiterate that these results are based on a particular set of test verbs and classes. Our other features – overlap of arguments across slots; tense, voice, and aspect; animacy of individual arguments – are useful in some tasks, suggesting that they may play a key role when the scope of experimentation is broadened, as will be necessary in future work.

This investigation has focused on English verb classes, and our feature space is limited in being motivated by alternations in English. It thus lacks grammatical features (e.g., the case of nouns, or other rich morphological properties) used in other languages to mark arguments or indicate the relation between them. Interestingly, Merlo *et al.* (2002) show that some of the hand-crafted features of Merlo and

Stevenson (2001) are useful in Italian for classifying the same verb classes in that language. We think a more general feature space such as ours will have even more potential for crosslinguistic applicability: We have devised a mapping of Levin's analysis of the variation in expression of arguments across classes in English to a general set of features. To the extent that Levin's analysis is grounded in general principles concerning the linking of semantic arguments to their syntactic expression, our feature space is an initial step in capturing variation in the expression of arguments across languages.

Acknowledgments

We gratefully acknowledge the financial support of NSERC of Canada, Bell University Labs, and the University of Toronto. Our colleagues in the computational linguistics group at Toronto were very helpful with comments on earlier instantiations of this work. We also thank four anonymous reviewers for their generous time in providing detailed comments and suggestions that helped us to improve the paper.

References

Abney, S. (1991) Parsing by chunks. In: Berwick, R., Abney, S. and Tenny, C. (eds.), *Principle-Based Parsing*. Kluwer Academic.

Aone, C. and McKee, D. (1996) Acquiring predicate-argument mapping information in multilingual texts. In: Boguraev, B. and Pustejovsky, J. (eds.), *Corpus Processing for Lexical Acquisition*, pp. 191–202. MIT Press.

Baker, C. F., Fillmore, C. J. and Lowe, J. B. (1998) The Berkeley FrameNet Project. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING/ACL-1998)*, pp. 86–90.

Brent, M. (1993) From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Linguistics*, **19**(3): 243–262.

Briscoe, T. and Carroll, J. (1993) Generalised probabilistic LR parsing of natural language (corpora) with unification-based grammars. *Computational Linguistics*, **19**(1): 25–60.

Briscoe, T. and Carroll, J. (1997) Automatic extraction of subcategorization from corpora. *Proceedings of the Fifth ACL Conference on Applied Natural Language Processing (ANLP-97)*, pp. 356–363, Washington, DC.

Burnard, L. (ed.) (2000) *British National Corpus User Reference Guide*. URL: `http://www.natcorp.ox.ac.uk/World/HTML/urg.html`.

Chang, C.-C. and Lin, C.-J. (2001) LIBSVM: A library for support vector machines. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Dorr, B. J. and Jones, D. (1996) Role of word sense disambiguation in lexical acquisition: Predicting semantics from syntactic cues. *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, pp. 322–327, Copenhagen, Denmark.

Dowty, D. R. (1991) Thematic proto-roles and argument selection. *Language*, **67**(3): 547–619.

Erk, K., Melinger, A. and Schulte im Walde, S. (eds.) (2005) *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*. Saarbrücken, Germany.

Gildea, D. (2002) Probabilistic models of verb-argument structure. *Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002)*, pp. 308–314, Taipei, Taiwan.

Gildea, D. and Jurafsky, D. (2002) Automatic labeling of semantic roles. *Computational Linguistics*, **28**(3): 245–288.

Girju, R., Roth, D. and Sammons, M. (2005) Token-level disambiguation of verbnet classes. (Erk *et al.* 2005), pp. 56–61.

Habash, N., Dorr, B. J. and Traum, D. (2003) Hybrid natural language generation from lexical conceptual structures. *Machine Translation*, **18**(2): 81–128.

Hsu, C.-W., Chang, C.-C. and Lin, C.-J. (2003) A practical guide to support vector classification, July. URL: `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`.

Hsu, C.-W. and Lin, C.-J. (2002) A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, **13**(2): 415–425.

Iglewicz, B. (1983) Robust scale estimators and confidence intervals for location. In: Hoaglin, D. C., Mosteller, M. and Tukey, J. W. (eds.), *Understanding Robust and Exploratory Data Analysis*. Wiley.

Joanis, E. (2002) Automatic verb classification using a general feature space. Master's thesis, Department of Computer Science, University of Toronto.

Joanis, E. and Stevenson, S. (2003) A general feature space for automatic verb classification. *Proceedings of the Tenth Conference of the European Chapter of the Association for Computational Linguistics (EACL-03)*, pp. 163–170, Budapest, Hungary.

Kipper, K., Dang, H. T. and Palmer, M. (2000) Class based construction of a verb lexicon. *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000)*, Austin, TX.

Kipper, K., Korhonen, A., Ryant, N. and Palmer, M. (2006) A large-scale extension of VerbNet with novel verb classes. *Proceedings of the 12th EURALEX International Congress*, Turin, Italy.

Korhonen, A. (2002) Semantically motivated subcategorization acquisition. *Proceedings of the ACL Workshop on Unsupervised Lexical Acquisition*, pp. 51–58.

Korhonen, A. and Briscoe, T. (2004) Extended lexical-semantic classification of english verbs. *Proceedings of the HLT/NAACL Workshop on Computational Lexical Semantics*, pp. 38–45.

Lapata, M. and Brew, C. (1999) Using subcategorization to resolve verb class ambiguity. In: Fung, P. and Zhou, J. (eds.), *Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora (EMNLP/VLC-99)*, pp. 266–274.

Lapata, M. and Brew, C. (2004) Verb class disambiguation using informative priors. *Computational Linguistics*, **30**(1): 45–73.

Levin, B. (1993) *English verb classes and alternations: A preliminary investigation*. University of Chicago Press.

Mayol, L., Boleda, G. and Badia, T. (2005) Automatic learning of syntactic verb classes. (Erk *et al.* 2005), pp. 92–97.

McCarthy, D. (2000) Using semantic preferences to identify verbal participation in role switching alternations. *Proceedings of the First Conference of the North American Chapter of the ACL (NAACL-2000)*, pp. 256–263, Seattle, WA.

McCarthy, D., Koeling, R., Weeds, J. and Carroll, J. (2004) Finding predominant senses in untagged text. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-2004)*, pp. 280–287, Barcelona, Spain.

Merlo, P. and Stevenson, S. (2001) Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, **27**(3): 373–408.

Merlo, P., Stevenson, S., Tsang, V. and Allaria, G. (2002) A multilingual paradigm for automatic verb classification. *Proceedings of the 40th Annual Meeting of the ACL*, pp. 207–214, Philadelphia, PA.

Oishi, A. and Matsumoto, Y. (1997) Detecting the organization of semantic subclasses of Japanese verbs. *Int. J. Corpus Linguistics*, **2**(1): 65–89.

Palmer, M., Gildea, D. and Kingsbury, P. (2005) The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, **31**(1): 71–106.

Pinker, S. (1989) *Learnability and cognition: the acquisition of argument structure*. MIT Press.

Resnik, P. (1996) Selectional constraints: an information-theoretic model and its computational realization. *Cognition*, **61**(1–2): 127–159.

Rifkin, R. and Klautau, A. (2004) In defense of one-vs-all classification. *J. Machine Learning Res.* **5**(Jan): 101–141.

Riloff, E. and Schmelzenbach, M. (1998) An empirical approach to conceptual case frame acquisition. *Proceedings of the Sixth Workshop on Very Large Corpora (WVLC-98)*, pp. 49–56, Montreal, Canada.

Rohde, D. L. T. (2002) TGrep2 user manual version 1.3. Available with the TGrep2 package at `http://tedlab.mit.edu/~dr/Tgrep2/`.

Rooth, M., Riezler, S., Prescher, D., Carroll, G. and Beil, F. (1999) Inducing a semantically annotated lexicon via EM-based clustering. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 104–111, College Park, MD.

Sarkar, A. and Tripasai, W. (2002) Learning verb argument structure from minimally annotated corpora. *Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002)*, pp. 864–869, Taipei, Taiwan.

Sarle, W. S. (2002) Should I nonlinearly transform the data? *Neural Network FAQ, part 2 of 7: Learning*. Periodic posting to the Usenet newsgroup `comp.ai.neural-nets`, URL: `ftp://ftp.sas.com/pub/neural/FAQ.html`.

Schulte im Walde, S. (2000) Clustering verbs semantically according to their alternation behaviour. *Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000)*, pp. 747–753, Saarbrücken, Germany.

Schulte im Walde, S. (2003) Experiments on the choice of features for learning verb classes. *Proceedings of the Tenth Conference of the European Chapter of the Association for Computational Linguistics (EACL-2003)*, pp. 315–322, Budapest, Hungary.

Schulte im Walde, S. and Brew, C. (2002) Inducing German semantic verb classes from purely syntactic subcategorisation information. *Proceedings of the 40th Annual Meeting of the ACL*, pp. 223–230, Philadelphia, PA.

Shi, L. and Mihalcea, R. (2005) Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing. In: Gelbukh, A. (ed.), *Computational Linguistics and Intelligent Text Processing; Sixth International Conference, CICLing 2005, Proceedings*, Lecture Notes in Computer Science, vol 3406, pp. 100–111, Mexico City, Mexico.

Stevenson, S. and Joanis, E. (2003) Semi-supervised verb class discovery using noisy features. *Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-2003)*, pp. 71–78, Edmonton, Canada.

Stevenson, S. and Merlo, P. (1999) Automatic verb classification using grammatical features. *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL-99)*, pp. 45–52, Bergen, Norway.

Stevenson, S., Merlo, P., Kariaeva, N. and Whitehouse, K. (1999) Supervised learning of lexical semantic verb classes using frequency distributions. *Proceedings of SigLex99: Standardizing Lexical Resources*, pp. 15–22, College Park, MD.

Swier, R. and Stevenson, S. (2004) Unsupervised semantic role labelling. *Proceedings of the 2004 Conference on Emperical Methods in Natural Language Processing*, pp. 95–102, Barcelona, Spain.

Swift, M. (2005) Towards automatic verb acquisition from VerbNet for spoken dialog processing. (Erk *et al.* 2005), pp. 115–120.

Tsang, V. and Stevenson, S. (2004) Calculating semantic distance between word sense probability distributions. *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004)*, pp. 81–88, Boston, MA.

Villavicencio, A. (2005) The availability of verb-particle constructions in lexical resources: How much is enough? *Computer Speech and Language, Special Issue on Multiword Expressions*, **19**(4): 415–432.