Automatic Verb Classification Based on Statistical Distributions of Argument Structure

Paola Merlo*
University of Geneva

Suzanne Stevenson†
University of Toronto

Automatic acquisition of lexical knowledge is critical to a wide range of natural language processing tasks. Especially important is knowledge about verbs, which are the primary source of relational information in a sentence—the predicate-argument structure that relates an action or state to its participants (i.e., who did what to whom). In this work, we report on supervised learning experiments to automatically classify three major types of English verbs, based on their argument structure—specifically, the thematic roles they assign to participants. We use linguistically-motivated statistical indicators extracted from large annotated corpora to train the classifier, achieving 69.8% accuracy for a task whose baseline is 34%, and whose expert-based upper bound we calculate at 86.5%. A detailed analysis of the performance of the algorithm and of its errors confirms that the proposed features capture properties related to the argument structure of the verbs. Our results validate our hypotheses that knowledge about thematic relations is crucial for verb classification, and that it can be gleaned from a corpus by automatic means. We thus demonstrate an effective combination of deeper linguistic knowledge with the robustness and scalability of statistical techniques.

1. Introduction

Automatic acquisition of lexical knowledge is critical to a wide range of natural language processing (NLP) tasks (Boguraev and Pustejovsky 1996). Especially important is knowledge about verbs, which are the primary source of relational information in a sentence—the predicate-argument structure that relates an action or state to its participants (i.e., who did what to whom). In facing the task of automatic acquisition of knowledge about verbs, two basic questions must be addressed:

- What information about verbs and their relational properties needs to be learned?
- What information can in practice be learned through automatic means?

In answering these questions, some approaches to lexical acquisition have focused on learning syntactic information about verbs, by automatically extracting subcategorization frames from a corpus or machine-readable dictionary (Brent 1993; Briscoe and Carroll 1997; Dorr 1997; Lapata 1999; Manning 1993; McCarthy and Korhonen 1998).

* Linguistics Department; University of Geneva; 2 rue de Candolle; 1211 Geneva 4, Switzerland; merlo@lettres.unige.ch
† Department of Computer Science; University of Toronto; 6 King's College Road; Toronto, ON M5S 3H5 Canada; suzanne@cs.toronto.edu

Table 1
Examples of verbs from the three optionally intransitive classes.

Unergative	The horse raced past the barn. The jockey raced the horse past the barn.
Unaccusative	The butter melted in the pan. The cook melted the butter in the pan.
Object-Drop	The boy played. The boy played soccer.

Other work has attempted to learn deeper semantic properties such as selectional restrictions (Resnik 1996; Riloff and Schmelzenbach 1998), verbal aspect (Klavans and Chodorow 1992; Siegel 1999), or lexical-semantic verb classes such as those proposed by Levin (1993) (Aone and McKee 1996; McCarthy 2000; Lapata and Brew 1999; Schulte im Walde 2000). In this paper, we focus on argument structure—the thematic roles assigned by a verb to its arguments—as the way in which the relational semantics of the verb is represented at the syntactic level.

Specifically, our proposal is to automatically classify verbs based on argument structure properties, using statistical corpus-based methods. We address the problem of classification because it provides a means for lexical organization which can effectively capture generalizations over verbs (Palmer 2000). Within the context of classification, the use of argument structure provides a finer discrimination among verbs than that induced by subcategorization frames (as we see below in our example classes, which allow the same subcategorizations but differ in thematic assignment), but a coarser classification than that proposed by Levin (in which classes such as ours are further subdivided according to more detailed semantic properties). This level of classification granularity appears to be appropriate for numerous language engineering tasks. Because knowledge of argument structure captures fundamental participant/event relations, it is crucial in parsing and generation (e.g., Srinivas and Joshi [1999]; Stede [1998]), in machine translation (Dorr 1997), and in information retrieval (Klavans and Kan 1998) and extraction (Riloff and Schmelzenbach 1998). Our use of statistical corpus-based methods to achieve this level of classification is motivated by our hypothesis that class-based differences in argument structure are reflected in statistics over the usages of the component verbs, and that those statistics can be automatically extracted from a large annotated corpus.

The particular classification problem within which we investigate this hypothesis is the task of learning the three major classes of optionally intransitive verbs in English: unergative, unaccusative, and object-drop verbs. (For the unergative/unaccusative distinction, see Perlmutter [1978]; Burzio [1986]; Levin and Rappaport Hovav [1995]). Table 1 shows an example of a verb from each class in its transitive and intransitive usages. These three classes are motivated by theoretical linguistic properties (see discussion and references below, and in Stevenson and Merlo [1997b]; Merlo and Stevenson [2000b]). Furthermore, it appears that the classes capture typological distinctions that are useful for machine translation (for example, causative unergatives are ungrammatical in many languages), as well as processing distinctions that are useful for generating naturally occurring language (for example, reduced relatives with unergative verbs are awkward, but they are acceptable, and in fact often preferred to full relatives for unaccusative and object-drop verbs) (Stevenson and Merlo 1997b; Merlo and Stevenson 1998).

Table 2
Summary of thematic role assignments by class.

	Transitive		Intransitive
Classes	Subject	Object	Subject
Unergative	Agent (of Causation)	Agent	Agent
Unaccusative	Agent (of Causation)	Theme	Theme
Object-Drop	Agent	Theme	Agent

The question then is what underlies these distinctions. We identify the property that precisely distinguishes among these three classes as that of argument structure—i.e., the thematic roles assigned by the verbs. The thematic roles for each class, and their mapping to subject and object positions, are summarized in Table 2. Note that verbs across these three classes allow the same subcategorization frames (taking an NP object or occurring intransitively); thus, classification based on subcategorization alone would not distinguish them. On the other hand, each of the three classes is comprised of multiple Levin classes, because the latter reflect more detailed semantic distinctions among the verbs (Levin 1993); thus, classification based on Levin's labeling would miss generalizations across the three broader classes. By contrast, as shown in Table 2, each class has a unique pattern of thematic assignments, which categorize the verbs precisely into the three classes of interest.

Although the granularity of our classification differs from Levin's, we draw on her hypothesis that semantic properties of verbs are reflected in their syntactic behavior. The behavior that Levin focuses on is the notion of **diathesis alternation**—an alternation in the expression of the arguments of a verb, such as the different mappings between transitive and intransitive that our verbs undergo. Whether a verb participates in a particular diathesis alternation or not is a key factor in Levin's approach to classification. We, like others in a computational framework, have extended this idea by showing that *statistics* over the alternants of a verb effectively capture information about its class (Lapata 1999; McCarthy 2000; Lapata and Brew 1999).

In our specific task, we analyze the pattern of thematic assignments given in Table 2 to develop statistical indicators that are able to determine the class of an optionally intransitive verb by capturing information across its transitive and intransitive alternants. These indicators serve as input to a machine learning algorithm, under a supervised training methodology, which produces an automatic classification system for our three verb classes. Since we rely on patterns of behavior across multiple occurrences of a verb, we begin with the problem of assigning a single class to the entire set of usages of a verb within the corpus. For example, we measure properties across all occurrences of a word, such as *raced*, in order to assign a single classification to the lexical entry for the verb *race*. This contrasts with work classifying individual occurrences of a verb in each local context, which have typically relied on training that includes instances of the verbs to be classified—essentially developing a bias that is used in conjunction with the local context to determine the best classification for new instances of previously seen verbs. By contrast, our method assigns a classification to verbs that have not previously been seen in the training data. Thus, while we do not as yet assign different classes to the instances of a verb, we can assign a single predominant class to new verbs that have never been encountered.

To preview our results, we demonstrate that combining just five numerical indicators, automatically extracted from large text corpora, is sufficient to reduce the error

rate in this classification task by more than 50% over chance. Specifically, we achieve almost 70% accuracy in a task whose baseline (chance) performance is 34%, and whose expert-based upper bound is calculated at 86.5%.

Beyond the interest for the particular classification task at hand, this work addresses more general issues concerning verb class distinctions based in argument structure. We evaluate our hypothesis that such distinctions are reflected in statistics over corpora through a computational experimental methodology in which we investigate as indicated each of the subhypotheses below, in the context of the three verb classes under study:

- Lexical features capture argument structure differences between verb classes.[1]
- The linguistically distinctive features exhibit distributional differences across the verb classes that are apparent within linguistic experience (i.e., they can be collected from text).
- The statistical distributions of (some of) the features contribute to learning the classifications of the verbs.

In the following sections, we show that all three hypotheses above are borne out. In Section 2, we describe the argument structure distinctions of our three verb classes in more detail. In support of the first hypothesis above, we discuss lexical correlates of the underlying differences in thematic assignments that distinguish the three verb classes under investigation. In Section 3, we show how to approximate these features by simple syntactic counts, and how to perform these counts on available corpora. We confirm the second hypothesis above, by showing that the differences in distribution predicted by the underlying argument structures are largely found in the data. In Section 4, in a series of machine learning experiments and a detailed analysis of errors, we confirm the third hypothesis by showing that the differences in the distribution of the extracted features are successfully used for verb classification. Section 5 evaluates the significance of these results by comparing the program's accuracy to an expert-based upper bound. We conclude the paper with a discussion of its contributions, comparison to related work, and suggestions for future extensions.

2. Deriving Classification Features from Argument Structure

Our task is to automatically build a classifier that can distinguish the three major classes of optionally intransitive verbs in English. As described above, these classes are differentiated by their argument structures. In the first subsection below, we elaborate on our description of the thematic role assignments for each of the verb classes under investigation—unergative, unaccusative, and object-drop. This analysis yields a distinctive pattern of thematic assignment for each class. (For more detailed discussion concerning the linguistic properties of these classes, and the behavior of their component verbs, please see Stevenson and Merlo [1997b]; Merlo and Stevenson [2000b].)

Of course, the key to any automatic classification task is to determine a set of useful features for discriminating the items to be classified. In the second subsection below,

1 By lexical we mean features that we think are likely stored in the lexicon, because they are properties of words and not of phrases or sentences. Note, however, that some lexical features may not necessarily be stored with individual words—indeed, the motivation for classifying verbs to capture generalizations within each class suggests otherwise.

we show how the analysis of thematic distinctions enables us to determine lexical properties that we hypothesize will exhibit useful, detectable frequency differences in our corpora, and thus serve as the machine learning features for our classification experiments.

2.1 The Argument Structure Distinctions

The verb classes are exemplified below, in sentences repeated from Table 1 for ease of exposition.

Unergative:	(1a) The horse raced past the barn.
	(1b) The jockey raced the horse past the barn.
Unaccusative:	(2a) The butter melted in the pan.
	(2b) The cook melted the butter in the pan.
Object-Drop:	(3a) The boy played.
	(3b) The boy played soccer.

The example sentences illustrate that all three classes participate in a diathesis alternation that relates a transitive and intransitive form of the verb. However, according to Levin (1993), each class exhibits a different type of diathesis alternation, which is determined by the particular semantic relations of the arguments to the verb. We make these distinctions explicit by drawing on a standard notion of thematic role, as each class has a distinct pattern of thematic assignments (i.e., different argument structures).

We assume here that a thematic role is a label taken from a fixed inventory of grammaticalized semantic relations; for example, an Agent is the doer of an action, and a Theme is the entity undergoing an event (Gruber 1965). While admitting that such notions as Agent and Theme lack formal definitions (in our work and in the literature more widely), the distinctions are clear enough to discriminate our three verb classes. For our purposes, these roles can simply be thought of as semantic labels which are non-decomposable, but there is nothing in our approach that rests on this assumption. Thus, our approach would also be compatible with a feature-based definition of participant roles, as long as the features capture such general distinctions as, for example, the doer of an action and the entity acted upon (Dowty 1991).

Note that in our focus on verb class distinctions we have not considered finer-grained features that rely on more specific semantic features, such as, for example, that the subject of the intransitive *melt* must be something that can change from solid to liquid. While this type of feature may be important for semantic distinctions among individual verbs, it thus far seems irrelevant to the level of verb classification that we adopt, which groups verbs more broadly according to syntactic and (somewhat coarser-grained) semantic properties.

Our analysis of thematic assignment—which was summarized in Table 2, repeated here as Table 3—is elaborated here for each verb class. The sentences in (1) above illustrate the relevant alternants of an unergative verb, *race*. Unergatives are intransitive action verbs whose transitive form, as in (1b), can be the causative counterpart of the intransitive form (1a). The type of causative alternation that unergatives participate in is the "induced action alternation" according to Levin (1993). For our thematic analysis, we note that the subject of an intransitive activity verb is specified to be an Agent. The subject of the transitive form is indicated by the label Agent of Causation, which indicates that the thematic role assigned to the subject is marked as the role which is

Table 3
Summary of thematic assignments.

	Transitive		Intransitive
Classes	Subject	Object	Subject
Unergative	Agent (of Causation)	Agent	Agent
Unaccusative	Agent (of Causation)	Theme	Theme
Object-Drop	Agent	Theme	Agent

introduced with the causing event. In a causative alternation, the semantic argument of the subject of the intransitive surfaces as the object of the transitive (Brousseau and Ritter 1991; Hale and Keyser 1993; Levin 1993; Levin and Rappaport Hovav 1995). For unergatives, this argument is an Agent and thus the alternation yields an object in the transitive form that receives an Agent thematic role (Cruse 1972). These thematic assignments are shown in the first row of Table 3.

The sentences in (2) illustrate the corresponding forms of an unaccusative verb, *melt*. Unaccusatives are intransitive change-of-state verbs, as in (2a); the transitive counterpart for these verbs also exhibits a causative alternation, as in (2b). This is the "causative/inchoative alternation" (Levin, 1993). Like unergatives, the subject of a transitive unaccusative is marked as the Agent of Causation. Unlike unergatives, though, the alternating argument of an unaccusative (the subject of the intransitive form that becomes the object of the transitive) is an entity undergoing a change of state, without active participation, and is therefore a Theme. The resulting pattern of thematic assignments is indicated in the second row of Table 3.

The sentences in (3) use an object-drop verb, *play*. These are activity verbs that exhibit a non-causative diathesis alternation, in which the object is simply optional. This is dubbed "the unexpressed object alternation" (Levin 1993), and has several subtypes that we do not distinguish here. The thematic assignment for these verbs is simply Agent for the subject (in both transitive and intransitive forms), and Theme for the optional object; see the last row of Table 3.

For further details and support of this analysis, please see the discussion in Stevenson and Merlo (1997b) and Merlo and Stevenson (2000b). For our purposes here, the important fact to note is that each of the three classes can be uniquely identified by the pattern of thematic assignments across the two alternants of the verbs.

2.2 Features for Automatic Classification

Our next task then is to derive, from these thematic patterns, useful features for automatically classifying the verbs. In what follows, we refer to the columns of Table 3 to explain how we expect the thematic distinctions to give rise to distributional properties, which, when appropriately approximated through corpus counts, will discriminate across the three classes.

Transitivity Consider the first two columns of thematic roles in Table 3, which illustrate the role assignment in the transitive construction. The Prague school's notion of linguistic markedness (Jakobson 1971; Trubetzkoy 1939) enables us to establish a scale of markedness of these thematic assignments and make a principled prediction about their frequency of occurrence. Typical tests to determine the unmarked element of a pair or scale are **simplicity**—the unmarked element is simpler, **distribution**—the unmarked member is more widely attested across languages, and **frequency**—the un-

marked member is more frequent (Greenberg 1966; Moravcsik and Wirth 1983). The claim of markedness theory is that, once an element has been identified by one test as the unmarked element of a scale, then all other tests will be correlated. The three thematic assignments appear to be ranked on a scale by the simplicity and distribution tests, as we describe below. From this, we can conclude that frequency, as a third correlated test, should also be ranked by the same scale, and we can therefore make predictions about the expected frequencies of the three thematic assignments.

First, we note that the specification of an Agent of Causation for transitive unergatives (such as *race*) and unaccusatives (such as *melt*) indicates a causative construction. Causative constructions relate two events, the causing event and the core event described by the intransitive verb; the Agent of Causation is the Agent of the causing event. This double event structure can be considered as more complex than the single event that is found in a transitive object-drop verb (such as *play*) (Stevenson and Merlo 1997b). The simplicity test thus indicates that the causative unergatives and unaccusatives are marked in comparison to the transitive object-drop verbs.

We further observe that the causative transitive of an unergative verb has an Agent thematic role in object position which is subordinated to the Agent of Causation in subject position, yielding an unusual "double agentive" thematic structure. This lexical causativization of unergatives (in contrast to analytic causativization) is a distributionally rarer phenomenon—found in fewer languages—than lexical causatives of unaccusatives. In asking native speakers about our verbs, we have found that lexical causatives of unergative verbs are not attested in Italian, French, German, Portuguese, Gungbe (Kwa family), and Czech. On the other hand, the lexical causatives are possible for unaccusative verbs (i.e., where the object is a Theme) in all these languages. Vietnamese appears to allow a very restricted form of causativization of unergatives limited to only those cases that have a comitative reading. The typological distribution test thus indicates that unergatives are more marked than unaccusatives in the transitive form.

From these observations, we can conclude that unergatives (such as *race*) have the most marked transitive argument structure, unaccusatives (such as *melt*) have an intermediately marked transitive argument structure, and object-drops (such as *play*) have the least marked transitive argument structure of the three. Under the assumptions of markedness theory outlined above, we then predict that unergatives are the least frequent in the transitive, that unaccusatives have intermediate frequency in the transitive, and that object-drop verbs are the most frequent in the transitive.

Causativity Due to the causative alternation of unergatives and unaccusatives, the thematic role of the *subject* of the intransitive is identical to that of the *object* of the transitive, as shown in the second and third columns of thematic roles in Table 3. Given the identity of thematic role mapped to subject and object positions across the two alternants, we expect to observe the same noun occurring at times as subject of the verb, and at other times as object of the verb. In contrast, for object-drop verbs, the thematic role of the subject of the intransitive is identical to that of the subject of the transitive, not the object of the transitive. We therefore expect that it will be less common for the same noun to occur in subject and object position across instances of the same object-drop verb.

Thus, we hypothesize that this pattern of thematic role assignments will be reflected in a differential amount of usage across the classes of the same nouns as subjects and objects for a given verb. Generally, we would expect that causative verbs (in our case, the unergative and unaccusative verbs) would have a greater degree of overlap of nouns in subject and object position than non-causative transitive verbs (in

our case, the object-drop verbs). However, since the causative is a transitive use, and the transitive use of unergatives is expected to be rare (see above), we do not expect unergatives to exhibit a high degree of detectable overlap in a corpus. Thus, this overlap of subjects and objects should primarily distinguish unaccusatives (predicted to have high overlap of subjects and objects) from the other two classes (each of which is predicted to have low [detectable] overlap of subjects and objects).

Animacy Finally, considering the roles in the first and last columns of thematic assignments in Table 3, we observe that unergative and object-drop verbs assign an agentive role to their subject in both the transitive and intransitive, while unaccusatives assign an agentive role to their subject only in the transitive. Under the assumption that the intransitive use of unaccusatives is not rare, we then expect that unaccusatives will occur less often overall with an agentive subject than will the other two verb classes. (The assumption that unaccusatives are not rare in the intransitive is based on the linguistic complexity of the causative transitive alternant, and is borne out in our corpus analysis.) On the further assumption that Agents tend to be animate entities more so than Themes are, we expect that unaccusatives will occur less frequently with an animate subject compared to unergative and object-drop verbs. Note the importance of our use of frequency distributions: the claim is not that only Agents can be animate, but rather that nouns that receive an Agent role will more often be animate than nouns that receive a Theme role.

Additional Features The above interactions between thematic roles and the syntactic expressions of arguments thus lead to three features whose distributional properties appear promising for distinguishing unergative, unaccusative and object-drop verbs: transitivity, causativity, and animacy of subject. We also investigate two additional syntactic features: the use of the passive or active voice, and the use of the past participle or simple past part-of-speech (POS) tag (VBN or VBD, in the Penn Treebank style). These features are related to the transitive/intransitive alternation, since a passive use implies a transitive use of the verb, as well as to the use of a past participle form of the verb.[2]

Table 4 summarizes the features we derive from the thematic properties, and our expectations concerning their frequency of use. We hypothesize that these five features will exhibit distributional differences in the observed usages of the verbs that can be used for classification. In the next section, we describe the actual corpus counts that we develop to approximate the features we have identified. (Notice that the counts will be imperfect approximations to the thematic knowledge, beyond the inevitable errors due to automatic extraction from large automatically annotated corpora. Even when the counts are precise, they only constitute an approximation to the actual thematic notions, since the features we are using are not logically implied by the knowledge we want to capture, but only statistically correlated.)

3. Data Collection and Analysis

Clearly, some of the features we've proposed are difficult (e.g., the passive use) or impossible (e.g., animate subject use) to automatically extract with high accuracy from a

2 For our sample verbs, the statistical correlation between the transitive and passive features is highly significant ($N = 59$, $R = .44$, $p = .001$), as is the correlation between the transitive and past participle features ($N = 59$, $R = .36$, $p = .005$). (Since, as explained in the next section, our features are expressed as proportions—e.g., percent transitive use out of detected transitive and intransitive use—correlations of intransitivity with passive or past participle use have the same magnitude but are negative.)

Table 4
The features and expected behavior.

Feature	Expected Frequency Pattern	Explanation
Transitivity	Unerg < Unacc < ObjDrop	Unaccusatives and unergatives have a causative transitive, hence lower transitive use. Furthermore, unergatives have an agentive object, hence very low transitive use.
Causativity	Unerg, ObjDrop < Unacc	Object-drop verbs do not have a causal agent, hence low "causative" use. Unergatives are rare in the transitive, hence low causative use.
Animacy	Unacc < Unerg, ObjDrop	Unaccusatives have a Theme subject in the intransitive, hence lower use of animate subjects.
Passive Voice	Unerg < Unacc < ObjDrop	Passive implies transitive use, hence correlated with transitive feature.
VBN Tag	Unerg < Unacc < ObjDrop	Passive implies past participle use (VBN), hence correlated with transitive (and passive).

large corpus, given the current state of annotation. However, we do assume that currently available corpora, such as the Wall Street Journal (WSJ), provide a representative, and large enough, sample of language from which to gather corpus counts that can approximate the distributional patterns of the verb class alternations. Our work draws on two text corpora—one an automatically tagged combined corpus of 65 million words (primarily WSJ), the second an automatically parsed corpus of 29 million words (a subset of the WSJ text from the first corpus). Using these corpora, we develop counting procedures that yield relative frequency distributions for approximations to the five linguistic features we have determined, over a sample of verbs from our three classes.

3.1 Materials and Method

We chose a set of 20 verbs from each class based primarily on the classification in Levin (1993).[3] The complete list of verbs appears in Table 5; the group 1/group 2 designation is explained below in the section on counting. As indicated in the table, unergatives are manner-of-motion verbs (from the "run" class in Levin), unaccusatives are change-of-state verbs (from several of the classes in Levin's change-of-state super-class), while object-drop verbs were taken from a variety of classes in Levin's classification, all of which undergo the unexpressed object alternation. The most frequently used classes are verbs of change of possession, image-creation verbs, and verbs of creation and transformation. The selection of verbs was based partly on our intuitive judgment that the verbs were likely to be used with sufficient frequency in the WSJ. Also, each

3 We used an equal number of verbs from each class in order to have a balanced group of items. One potential disadvantage of this decision is that each verb class is represented equally, even though they may not be equally frequent in the corpora. Although we lose the relative frequency information among the classes that could provide a better bias for assigning a default classification (i.e., the most frequent one), we have the advantage that our classifier will be equally informed (in terms of number of exemplars) about each class.

Note that there are only 19 unaccusative verbs because *ripped*, which was initially counted in the unaccusatives, was then excluded from the analysis as it occurred mostly in a very different usage in the corpus (as verb+particle, in *ripped off*) from the intended optionally intransitive usage.

Table 5
Verbs used in the experiments.

Class Name	Description	Selected Verbs
Unergative	manner of motion	*jumped, rushed, marched, leaped, floated, raced, hurried, wandered, vaulted, paraded* (group 1); *galloped, glided, hiked, hopped, jogged, scooted, scurried, skipped, tiptoed, trotted* (group 2).
Unaccusative	change of state	*opened, exploded, flooded, dissolved, cracked, hardened, boiled, melted, fractured, solidified* (group 1); *collapsed, cooled, folded, widened, changed, cleared, divided, simmered, stabilized* (group 2).
Object-Drop	unexpressed object alternation	*played, painted, kicked, carved, reaped, washed, danced, yelled, typed, knitted* (group 1); *borrowed, inherited, organized, rented, sketched, cleaned, packed, studied, swallowed, called* (group 2).

verb presents the same form in the simple past and in the past participle (the regular "-ed" form). In order to simplify the counting procedure, we included only the "-ed" form of the verb, on the assumption that counts on this single verb form would approximate the distribution of the features across all forms of the verb. Additionally, as far as we were able given the preceding constraints, we selected verbs that could occur in the transitive and in the passive. Finally, we aimed for a frequency cut-off of 10 occurrences or more for each verb, although for unergatives we had to use one verb (*jogged*) that only occurred 8 times in order to have 20 verbs that satisfied the other criteria above.

In performing this kind of corpus analysis, one has to recognize the fact that current corpus annotations do not distinguish verb senses. In these counts, we did not distinguish a core sense of the verb from an extended use of the verb. So, for instance, the sentence *Consumer spending jumped 1.7% in February after a sharp drop the month before* (WSJ 1987) is counted as an occurrence of the manner-of-motion verb *jump* in its intransitive form. This particular sense extension has a transitive alternant, but not a causative transitive (i.e., *Consumer spending jumped the barrier...*, but not *Low taxes jumped consumer spending...*). Thus, while the possible subcategorizations remain the same, rates of transitivity and causativity may be different than for the literal manner-of-motion sense. This is an unavoidable result of using simple, automatic extraction methods given the current state of annotation of corpora.

For each occurrence of each verb, we counted whether it was in a transitive or intransitive use (TRANS), in a passive or active use (PASS), in a past participle or simple past use (VBN), in a causative or non-causative use (CAUS), and with an animate subject or not (ANIM).[4] Note that, except for the VBN feature, for which we simply extract the POS tag from the corpus, all other counts are approximations to the actual linguistic behaviour of the verb, as we describe in detail below.

4 One additional feature was recorded—the log frequency of the verb in the 65 million word corpus—motivated by the conjecture that the frequency of a verb may help in predicting its class. In our machine learning experiments, however, this conjecture was not borne out, as the frequency feature did not improve performance. This is the case for experiments on all of the verbs, as well as for separate experiments on the group 1 verbs (which were matched across the classes for frequency) and the group 2 verbs (which were not). We therefore limit discussion here to the thematically-motivated features.

The first three counts (TRANS, PASS, VBN) were performed on the tagged ACL/DCI corpus available from the Linguistic Data Consortium, which includes the Brown Corpus (of one million words) and years 1987–1989 of the Wall Street Journal, a combined corpus in excess of 65 million words. The counts for these features proceeded as follows:

- TRANS: A number, a pronoun, a determiner, an adjective, or a noun were considered to be indication of a potential object of the verb. A verb occurrence preceded by forms of the verb *be*, or immediately followed by a potential object was counted as transitive; otherwise, the occurrence was counted as intransitive (specifically, if the verb was followed by a punctuation sign—commas, colons, full stops—or by a conjunction, a particle, a date, or a preposition.)
- PASS: A main verb (i.e., tagged VBD) was counted as active. A token with tag VBN was also counted as active if the closest preceding auxiliary was *have*, while it was counted as passive if the closest preceding auxiliary was *be*.
- VBN: The counts for VBN/VBD were simply done based on the POS label within the tagged corpus.

Each of the above three counts was normalized over all occurrences of the "-ed" form of the verb, yielding a single relative frequency measure for each verb for that feature; i.e., percent transitive (versus intransitive) use, percent active (versus passive) use, and percent VBN (versus VBD) use, respectively.

The last two counts (CAUS and ANIM) were performed on a parsed version of the 1988 year of the Wall Street Journal, so that we could extract subjects and objects of the verbs more accurately. This corpus of 29 million words was provided to us by Michael Collins, and was automatically parsed with the parser described in Collins (1997).[5] The counts, and their justification, are described here:

- CAUS: As discussed above, the object of a causative transitive is the same semantic argument of the verb as the subject of the intransitive. The causative feature was approximated by the following steps, intended to capture the degree to which the subject of a verb can also occur as its object. Specifically, for each verb occurrence, the subject and object (if there was one) were extracted from the parsed corpus. The observed subjects across all occurrences of the verb were placed into one multiset of nouns, and the observed objects into a second multiset of nouns. (A multiset, or bag, was used so that our representation indicated the number of times each noun was used as either subject or object.) Then, the proportion of overlap between the two multisets was calculated. We define overlap as the largest multiset of elements belonging to both the

5 Readers might be concerned about the portability of this method to languages for which no large parsed corpus is available. It is possible that using a fully parsed corpus is not necessary. Our results were replicated in English without the need for a fully parsed corpus (Anoop Sarkar, p.c., citing a project report by Wootiporn Tripasai). Our method was applied to 23 million words of the WSJ that were automatically tagged with Ratnaparkhi's maximum entropy tagger (Ratnaparkhi 1996) and chunked with the partial parser CASS (Abney 1996). The results are very similar to ours (best accuracy 66.6%), suggesting that a more accurate tagger than the one used on our corpus might in fact be sufficient to overcome the fact that no full parse is available.

subject and the object multisets; e.g., the overlap between $\{a, a, a, b\}$ and $\{a\}$ is $\{a, a, a\}$. The proportion is the ratio between the cardinality of the overlap multiset, and the sum of the cardinality of the subject and object multisets. For example, for the simple sets of characters above, the ratio would be 3/5, yielding a value of .60 for the CAUS feature.

- ANIM: A problem with a feature like animacy is that it requires either manual determination of the animacy of extracted subjects, or reference to an on-line resource such as WordNet for determining animacy. To approximate animacy with a feature that can be extracted automatically, and without reference to a resource external to the corpus, we take advantage of the well-attested animacy hierarchy, according to which pronouns are the most animate (Silverstein 1976; Dixon 1994). The hypothesis is that the words *I*, *we*, *you*, *she*, *he*, and *they* most often refer to animate entities. This hypothesis was confirmed by extracting 100 occurrences of the pronoun *they*, which can be either animate or inanimate, from our 65 million word corpus. The occurrences immediately preceded a verb. After eliminating repetitions, 94 occurrences were left, which were classified by hand, yielding 71 animate pronouns, 11 inanimate pronouns and 12 unclassified occurrences (for lack of sufficient context to recover the antecedent of the pronoun with certainty). Thus, at least 76% of usages of *they* were animate; we assume the percentage of animate usages of the other pronouns to be even higher. Since the hypothesis was confirmed, we count pronouns (other than *it*) in subject position (Kariaeva [1999]; cf. Aone and McKee [1996]). The values for the feature were determined by automatically extracting all subject/verb tuples including our 59 example verbs from the parsed corpus, and computing the ratio of occurrences of pronoun subjects to all subjects for each verb.

Finally, as indicated in Table 5, the verbs are designated as belonging to "group 1" or "group 2". All the verbs are treated equally in our data analysis and in the machine learning experiments, but this designation does indicate a difference in details of the counting procedures described above. The verbs in group 1 had been used in an earlier study in which it was important to minimize noisy data (Stevenson and Merlo 1997a), so they generally underwent greater manual intervention in the counts. In adding group 2 for the classification experiment, we chose to minimize the intervention in order to demonstrate that the classification process is robust enough to withstand the resulting noise in the data.

For group 2, the transitivity, voice, and VBN counts were done automatically without any manual intervention. For group 1, these three counts were done automatically by regular expression patterns, and then subjected to correction, partly by hand and partly automatically, by one of the authors. For transitivity, the adjustments vary for the individual verbs. Most of the reassignments from a transitive to an intransitive labelling occurred when the following noun was not the direct object but rather a measure phrase or a date. Most of the reassignments from intransitive to transitive occurred when a particle or a preposition following the verb did not introduce a prepositional phrase, but instead indicated a passive form (*by*) or was part of a phrasal verb. Some verbs were mostly used adjectivally, in which case they were excluded from the transitivity counts. For voice, the required adjustments included cases of coordination of the past participle when the verb was preceded by a conjunction, or a comma.

Table 6
Aggregated relative frequency data for the five features. E = unergatives, A = unaccusatives, O = object-drops.

		TRANS		PASS		VBN		CAUS		ANIM	
Class	N	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
E	20	0.23	0.23	0.07	0.12	0.21	0.26	0.00	0.00	0.25	0.24
A	19	0.40	0.24	0.33	0.27	0.65	0.27	0.12	0.14	0.07	0.09
O	20	0.62	0.25	0.31	0.26	0.65	0.23	0.04	0.07	0.15	0.14

These were collected and classified by hand as passive or active based on intuition. Similarly, partial adjustments to the VBN counts were made by hand.

For the causativity feature, subjects and objects were determined by manual inspection of the corpus for verbs belonging to group 1, while they were extracted automatically from the parsed corpus for group 2. The group 1 verbs were sampled in three ways, depending on total frequency. For verbs with less than 150 occurrences, all instances of the verbs were used for subject/object extraction. For verbs whose total frequency was greater than 150, but whose VBD frequency was in the range 100–200, we extracted subjects and objects of the VBD occurrences only. For higher frequency verbs, we used only the first 100 VBD occurrences.[6] The same script for computing the overlap of the extracted subjects and objects was then used on the resulting subject/verb and verb/object tuples for both group 1 and group 2 verbs.

The animacy feature was calculated over subject/verb tuples extracted automatically for both groups of verbs from the parsed corpus.

3.2 Data Analysis

The data collection described above yields the following data points in total: TRANS: 27403; PASS: 20481; VBN: 36297; CAUS: 11307; ANIM: 7542. (Different features yield different totals because they were sampled independently, and the search patterns to extract some features are more imprecise than others.) The aggregate means by class of the normalized frequencies for all verbs are shown in Table 6; item by item distributions are provided in Appendix A, and raw counts are available from the authors. Note that aggregate means are shown for illustration purposes only—all machine learning experiments are performed on the individual normalized frequencies for each verb, as given in Appendix A.

The observed distributions of each feature are indeed roughly as expected according to the description in Section 2. Unergatives show a very low relative frequency of the TRANS feature, followed by unaccusatives, then object-drop verbs. Unaccusative verbs show a high frequency of the CAUS feature and a low frequency of the ANIM feature compared to the other classes. Somewhat unexpectedly, object-drop verbs exhibit a non-zero mean CAUS value (almost half the verbs have a CAUS value greater than zero), leading to a three-way causative distinction among the verb classes. We suspect that the approximation that we used for causative use—the overlap between subjects

6 For this last set of high-frequency verbs (*exploded, jumped, opened, played, rushed*), we used the first 100 occurrences as the simplest way to collect the sample. In response to an anonymous reviewer's concern, we later verified that these counts were not different from counts obtained by random sampling of 100 VBD occurrences. A paired t-test of the two sets of counts (first 100 sampling and random sampling) indicates that the two sets of counts are not statistically different ($t = 1.283$, $DF = 4$, $p = 0.2687$).

Table 7
Manually (Man) and automatically (Aut) calculated features for a random sample of verbs. T = TRANS, P = PASS, V = VBN, C = CAUS, A = ANIM.

	Unergative				Unaccusative				Object-Drop			
	hopped		scurried		folded		stabilized		inherited		swallowed	
	Man	Aut	Man	Aut	Man	Aut	Man	Aut	Man	Aut	Man	Aut
T	0.21	0.21	0.00	0.00	0.71	0.23	0.24	0.18	1.00	0.64	0.96	0.35
P	0.00	0.00	0.00	0.00	0.44	0.33	0.19	0.13	0.39	0.13	0.54	0.44
V	0.03	0.00	0.10	0.00	0.56	0.73	0.71	0.92	0.56	0.60	0.64	0.79
C	0.00	0.00	0.00	0.00	0.54	0.00	0.24	0.35	0.00	0.06	0.00	0.04
A	0.93	1.00	0.90	0.14	0.23	0.00	0.02	0.00	0.58	0.32	0.35	0.22

and objects for a verb—also captures a "reciprocity" effect for some object-drop verbs (such as *call*), in which subjects and objects can be similar types of entities. Finally, although expected to be a redundant indicator of transitivity, PASS and VBN, unlike TRANS, have very similar values for unaccusative and object-drop verbs, indicating that their distributions are sensitive to factors we have not yet investigated.

One issue we must address is how precisely the automatic counts reflect the actual linguistic behaviour of the verbs. That is, we must be assured that the patterns we note in the data in Table 6 are accurate reflections of the differential behaviour of the verb classes, and not an artifact of the way in which we estimate the features, or a result of inaccuracies in the counts. In order to evaluate the accuracy of our feature counts, we selected two verbs from each class, and determined the "true" value of each feature for each of those six verbs through manual counting. The six verbs were randomly selected from the group 2 subset of the verbs, since counts for group 2 verbs (as explained above) had not undergone manual correction. This allows us to determine the accuracy of the fully automatic counting procedures. The selected verbs (and their frequencies) are: *hopped* (29), *scurried* (21), *folded* (189), *stabilized* (286), *inherited* (357), *swallowed* (152). For verbs that had a frequency of over 100 in the "-ed" form, we performed the manual counts on the first 100 occurrences.

Table 7 shows the results of the manual counts, reported as proportions to facilitate comparison to the normalized automatic counts, shown in adjoining columns. We observe first that, overall, most errors in the automatic counts occur in the unaccusative and object-drop verbs. While tagging errors affect the VBN feature for all of the verbs somewhat, we note that TRANS and PASS are consistently underestimated for unaccusative and object-drop verbs. These errors make the unaccusative and object-drop feature values more similar to each other, and therefore potentially harder to distinguish. Furthermore, because the TRANS and PASS values are underestimated by the automatic counts, and therefore lower in value, they are also closer to the values for the unergative verbs. For the CAUS feature, we predict the highest values for the unaccusative verbs, and while that prediction is confirmed, the automatic counts for that class also show the most errors. Finally, although the general pattern of higher values for the ANIM feature of unergatives and object-drop verbs is preserved in the automatic counts, the feature is underestimated for almost all the verbs, again making the values for that feature closer across the classes than they are in reality.

We conclude that, although there are inaccuracies in all the counts, the general patterns expected based on our analysis of the verb classes hold in both the manual and automatic counts. Errors in the estimating and counting procedures are therefore

not likely to be responsible for the pattern of data in Table 6 above, which generally matches our predictions. Furthermore, the errors, at least for this random sample of verbs, occur in a direction that makes our task of distinguishing the classes more difficult, and indicates that developing more accurate search patterns may possibly sharpen the class distinctions, and improve the classification performance.

4. Experiments in Classification

In this section, we turn to our computational experiments that investigate whether the statistical indicators of thematic properties that we have developed can in fact be used to classify verbs. Recall that the task we have set ourselves is that of automatically learning the best class for a set of usages of a verb, as opposed to classifying individual occurrences of the verb. The frequency distributions of our features yield a vector for each verb that represents the estimated values for the verb on each dimension across the entire corpus:

Vector template: [verb-name, TRANS, PASS, VBN, CAUS, ANIM, class]

Example: [opened, .69, .09, .21, .16, .36, unacc]

The resulting set of 59 vectors constitutes the data for our machine learning experiments. We use this data to train an automatic classifier to determine, given the feature values for a new verb (not from the training set), which of the three major classes of English optionally intransitive verbs it belongs to.

4.1 Experimental Methodology

In pilot experiments on a subset of the features, we investigated a number of supervised machine learning methods that produce automatic classifiers (decision tree induction, rule learning, and two types of neural networks), as well as hierarchical clustering; see Stevenson et al. (1999) for more detail. Because we achieved approximately the same level of performance in all cases, we narrowed our further experimentation to the publicly available version of the C5.0 machine learning system (http://www.rulequest.com), a newer version of C4.5 (Quinlan 1992), due to its ease of use and wide availability. The C5.0 system generates both decision trees and corresponding rule sets from a training set of known classifications. In our experiments, we found little to no difference in performance between the trees and rule sets, and report only the rule set results.

In the experiments below, we follow two methodologies in training and testing, each of which tests a subset of cases held out from the training data. Thus, in all cases, the results we report are on test data that was never seen in training.[7]

The first training and testing methodology we follow is 10-fold cross-validation. In this approach, the system randomly divides the data into ten parts, and runs ten times on a different 90%-training-data/10%-test-data split, yielding an average accuracy and standard error across the ten test sets. This training methodology is very useful for

7 One anonymous reviewer raised the concern that we do not test on verbs that were unseen by the authors prior to finalizing the specific features to count. However, this does not reduce the generality of our results. The features we use are motivated by linguistic theory, and derived from the set of thematic properties that discriminate the verb classes. It is therefore very unlikely that they are skewed to the particular verbs we have chosen. Furthermore, our cross-validation experiments, described in the next subsection, show that our results hold across a very large number of randomly selected subsets of this sample of verbs.

our application, as it yields performance measures across a large number of training data/test data sets, avoiding the problems of outliers in a single random selection from a relatively small data set such as ours.

The second methodology is a single hold-out training and testing approach. Here, the system is run N times, where N is the size of the data set (i.e., the 59 verbs in our case), each time holding out a single data vector as the test case and using the remaining N-1 vectors as the training set. The single hold-out methodology yields an overall accuracy rate (when the results are averaged across all N trials), but also—unlike cross-validation—gives us classification results on each individual data vector. This property enables us to analyze differential performance on the individual verbs and across the different verb classes.

Under both training and testing methodologies, the baseline (chance) performance in this task—a three-way classification—is 33.9%. In the single hold-out methodology, there are 59 test cases, with 20, 19, and 20 verbs each from the unergative, unaccusative, and object-drop classes, respectively. Chance performance of picking a single class label as a default and assigning it to all cases would yield at most 20 out of the 59 cases correct, or 33.9%. For the cross-validation methodology, the determination of a baseline is slightly more complex, as we are testing on a random selection of 10% of the full data set in each run. The 33.9% figure represents the expected relative proportion of a test set that would be labelled correctly by assignment of a default class label to the entire test set. Although the precise make-up of the test cases vary, on average the test set will represent the class membership proportions of the entire set of verbs. Thus, as with the single hold-out approach, chance accuracy corresponds to a maximum of 20/59, or 33.9%, of the test set being labelled correctly.

The theoretical maximum accuracy for the task is, of course, 100%, although in Section 5 we discuss some classification results from human experts that indicate that a more realistic expectation is much lower (around 87%).

4.2 Results Using 10-Fold Cross-Validation

We first report the results of experiments using a training methodology of 10-fold cross-validation repeated 50 times. This means that the 10-fold cross-validation procedure is repeated for 50 different random divisions of the data. The numbers reported are the averages of the results over all the trials. That is, the average accuracy and standard error from each random division of the data (a single cross-validation run including 10 training and test sets) are averaged across the 50 different random divisions. This large number of experimental trials gives us a very tight bound on the mean accuracy reported, enabling us to determine with high confidence the statistical significance of differences in results.

Table 8 shows that performance of classification using individual features varies greatly, from little above the baseline to almost 22% above the baseline, or a reduction of a third of the error rate, a very good result for a single feature. (All reported accuracies in Table 8 are statistically distinct, at the $p < .01$ level, using an ANOVA [$df = 249$, $F = 334.72$], with a Tukey-Kramer post test.)

The first line of Table 9 shows that the combination of all features achieves an accuracy of 69.8%, which is 35.9% over the baseline, for a reduction in the error rate of 54%. This is a rather considerable result, given the very low baseline (33.9%). Moreover, recall that our training and testing sets are always disjoint (cf., Lapata and Brew [1999]; Siegel [1999]); in other words, we are predicting the classification of verbs that were never seen in the training corpus, the hardest situation for a classification algorithm.

The second through sixth lines of Table 9 show the accuracy achieved on each subset of features that results from removing a single feature. This allows us to evaluate

Table 8
Percent accuracy and standard error of the verb classification task using each feature individually, under a training methodology of 10-fold cross-validation repeated 50 times.

Feature	%Accuracy	%SE
CAUS	55.7	.1
VBN	52.5	.5
PASS	50.2	.5
TRANS	47.1	.4
ANIM	35.3	.5

Table 9
Percent accuracy and standard error of the verb classification task using features in combination, under a training methodology of 10-fold cross-validation repeated 50 times.

Features Used	Feature Not Used	%Accuracy	%SE
1. TRANS PASS VBN CAUS ANIM		69.8	.5
2. TRANS VBN CAUS ANIM	PASS	69.8	.5
3. TRANS PASS VBN ANIM	CAUS	67.3	.6
4. TRANS PASS CAUS ANIM	VBN	66.5	.5
5. TRANS PASS VBN CAUS	ANIM	63.2	.6
6. PASS VBN CAUS ANIM	TRANS	61.6	.6

the contribution of each feature to the performance of the classification process, by comparing the performance of the subset without it, to the performance using the full set of features. We see that the removal of PASS (second line) has no effect on the results, while removal of the remaining features yields a 2–8% decrease in performance. (In Table 9, the differences between all reported accuracies are statistically significant, at the $p < .05$ level, *except* for between lines 1 and 2, lines 3 and 4, and lines 5 and 6, using an ANOVA [$df = 299$, $F = 37.52$], with a Tukey-Kramer post test.) We observe that the behavior of the features in combination cannot be predicted by the individual feature behavior. For example, CAUS, which is the best individually, does not greatly affect accuracy when combined with the other features (compare line 3 to line 1). Conversely, ANIM and TRANS, which do not classify verbs accurately when used alone, are the most relevant in a combination of features (compare lines 5 and 6 to line 1). We conclude that experimentation with combinations of features is required to determine the relevance of individual features to the classification task.

The general behaviour in classification based on individual features and on size 4 and size 5 subsets of features is confirmed for all subsets. Appendix B reports the results for all subsets of feature combinations, in order of decreasing performance. Table 10 summarizes this information. In the first data column, the table illustrates the average accuracy across all subsets of each size. The second through sixth data columns report the average accuracy of all the size n subsets in which each feature occurs. For example, the second data cell in the second row (54.9) indicates the average accuracy of all subsets of size 2 that contain the feature VBN. The last row of the table indicates the average accuracy for each feature of all subsets containing that feature.

Table 10
Average percent accuracy of feature subsets, by subset size and by sets of each size including each feature.

Subset Size	Mean Accuracy by Subset Size	Mean Accuracy of Subsets that Include Each Feature				
		VBN	PASS	TRANS	ANIM	CAUS
1	48.2	52.5	50.2	47.1	35.3	55.7
2	55.1	54.9	52.8	56.4	58.0	57.6
3	60.5	60.1	58.5	62.3	61.1	60.5
4	65.7	65.5	64.7	66.7	66.3	65.3
5	69.8	69.8	69.8	69.8	69.8	69.8
Mean Acc/Feature:		60.6	59.2	60.5	58.1	61.8

The first observation—that more features perform better—is confirmed overall, in all subsets. Looking at the first data column of Table 10, we can observe that, on average, larger sets of features perform better than smaller sets. Furthermore, as can be seen in the following individual feature columns, individual features perform better in a bigger set than in a smaller set, without exception. The second observation—that the performance of individual features is not always a predictor of their performance in combination—is confirmed by comparing the average performance of each feature in subsets of different sizes to the average across all subsets of each size. We can observe, for instance, that the feature CAUS, which performs very well alone, is average in feature combinations of size 3 or 4. By contrast, the feature ANIM, which is the worst if used alone, is very effective in combination, with above average performance for all subsets of size 2 or greater.

4.3 Results Using Single Hold-Out Methodology

One of the disadvantages of the cross-validation training methodology, which averages performance across a large number of random test sets, is that we do not have performance data for each verb, nor for each class of verbs. In another set of experiments, we used the same C5.0 system, but employed a single hold-out training and testing methodology. In this approach, we hold out a single verb vector as the test case, and train the system on the remaining 58 cases. We then test the resulting classifier on the single hold-out case, and record the assigned class for that verb. This procedure is repeated for each of the 59 verbs. As noted above, the single hold-out methodology has the benefit of yielding both classification results on each individual verb, and an overall accuracy rate (the average results across all 59 trials). Moreover, the results on individual verbs provide the data necessary for determining accuracy for each verb class. This allows us to determine the contribution of individual features as above, but with reference to their effect on the performance of individual classes. This is important, as it enables us to evaluate our hypotheses concerning the relation between the thematic features and verb class distinctions, which we turn to in Section 4.4.

We performed single hold-out experiments on the full set of features, as well as on each subset of features with a single feature removed. The first line of Table 11 shows that the overall accuracy for all five features is almost exactly the same as that achieved with the 10-fold cross-validation methodology (69.5% versus 69.8%). As with the cross-validation results, the removal of PASS does not degrade performance—in fact, here its removal appears to improve performance (see line 2 of Table 11). However, it should be noted that this increase in performance results from one additional verb being

Table 11
Percent accuracy of the verb classification task using features in combination, under a single hold-out training methodology.

Features Used	Feature Not Used	%Accuracy on All Verbs
1. TRANS PASS VBN CAUS ANIM		69.5
2. TRANS VBN CAUS ANIM	PASS	71.2
3. TRANS PASS VBN ANIM	CAUS	62.7
4. TRANS PASS CAUS ANIM	VBN	61.0
5. TRANS PASS VBN CAUS	ANIM	61.0
6. PASS VBN CAUS ANIM	TRANS	64.4

Table 12
F score of classification within each class, under a single hold-out training methodology.

Features Used	Feature Not Used	F score (%) for Unergs	F score (%) for Unaccs	F score (%) for Objdrops
1. TRANS PASS VBN CAUS ANIM		73.9	68.6	64.9
2. TRANS VBN CAUS ANIM	PASS	76.2	75.7	61.6
3. TRANS PASS VBN ANIM	CAUS	65.1	60.0	62.8
4. TRANS PASS CAUS ANIM	VBN	66.7	65.0	51.3
5. TRANS PASS VBN CAUS	ANIM	72.7	47.0	60.0
6. PASS VBN CAUS ANIM	TRANS	78.1	51.5	61.9

classified correctly. The remaining lines of Table 11 show that the removal of any other feature has a 5–8% negative effect on performance, again similar to the cross-validation results. (Although note that the precise accuracy achieved is not the same in each case as with 10-fold cross-validation, indicating that there is some sensitivity to the precise make-up of the training set when using a subset of the features.)

Table 12 presents the results of the single hold-out experiments in terms of performance within each class, using an F measure with balanced precision and recall.[8] The first line of the table shows clearly that, using all five features, the unergatives are classified with greater accuracy ($F = 73.9\%$) than the unaccusative and object-drop verbs (F scores of 68.6% and 64.9%, respectively). The features appear to be better at distinguishing unergatives than the other two verb classes. The remaining lines of Table 12 show that this pattern holds for all of the subsets of features as well. Clearly, future work on our verb classification task will need to focus on determining features that better discriminate unaccusative and object-drop verbs.

One potential explanation that we can exclude is that the pattern of results is due simply to the frequencies of the verbs—that is, that more frequent verbs are more accurately classified. We examined the relation between classification accuracy and log

8 For all previous results, we reported an accuracy measure (the percentage of correct classifications out of all classifications). Using the terminology of true or false positives/negatives, this is the same as truePositives/(truePositives + falseNegatives). In the earlier results, there are no falsePositives or trueNegatives, since we are only considering for each verb whether it is correctly classified (truePositive) or not (falseNegative). However, when we turn to analyzing the data for each class, the possibility arises of having falsePositives and trueNegatives *for that class*. Hence, here we use the balanced F score, which calculates an overall measure of performance as $2PR/(P + R)$, in which P (precision) is truePositives/(truePositives + falsePositives), and R (recall) is truePositives/(truePositives + falseNegatives).

frequencies of the verbs, both by class and individually. By class, unergatives have the lowest average log frequency (1.8), but are the best classified, while unaccusatives and object-drops are comparable (average log frequency = 2.4). If we group individual verbs by frequency, the proportion of errors to the total number of verbs is not linearly related to frequency (log frequency < 2: 7 errors/24 verbs, or 29% error; log frequency between 2 and 3: 7 errors/25 verbs, or 28% error; log frequency > 3: 4 errors/10 verbs, or 40% error). Moreover, it seems that the highest-frequency verbs pose the most problems to the program. In addition, the only verb of log frequency < 1 is correctly classified, while the only one with log frequency > 4 is not. In conclusion, we do not find that there is a simple mapping from frequency to accuracy. In particular, it is not the case that more frequent classes or verbs are more accurately classified.

One factor possibly contributing to the poorer performance on unaccusatives and object-drops is the greater degree of error in the automatic counting procedures for these verbs, which we discussed in Section 3.2. In addition to exploration of other linguistic features, another area of future work is to develop better search patterns, for transitivity and passive in particular. Unfortunately, one limiting factor in automatic counting is that we inherit the inevitable errors in POS tags in an automatically tagged corpus. For example, while the unergative verbs are classified highly accurately, we note that two of the three errors in misclassifying unergatives (*galloped* and *paraded*) are due to a high degree of error in tagging.[9] The verb *galloped* is incorrectly tagged VBN instead of VBD in all 12 of its uses in the corpus, and the verb *paraded* is incorrectly tagged VBN instead of VBD in 13 of its 33 uses in the corpus. After correcting only the VBN feature of these two verbs to reflect the actual part of speech, overall accuracy in classification increases by almost 10%, illustrating the importance of both accurate counts and accurate annotation of the corpora.

4.4 Contribution of the Features to Classification

We can further use the single hold-out results to determine the contribution of each feature to accuracy within each class. We do this by comparing the class labels assigned using the full set of five features (TRANS, PASS, VBN, CAUS, ANIM) with the class labels assigned using each size 4 subset of features. The difference in classifications between each four-feature subset and the full set of features indicates the changes in class labels that we can attribute to the added feature in going from the four-feature to five-feature set. Thus, we can see whether the features indeed contribute to discriminating the classes in the manner predicted in Section 2.2, and summarized here in Table 13.

We illustrate the data with a set of confusion matrices, in Tables 14 and 15, which show the pattern of errors according to class label for each set of features. In each confusion matrix, the rows indicate the actual class of incorrectly classified verbs, and the columns indicate the assigned class. For example, the first row of the first panel of Table 14 shows that one unergative was incorrectly labelled as unaccusative, and two unergatives as object-drop. To determine the confusability of any two classes (the

9 The third error in classification of unergatives is the verb *floated*, which we conjecture is due not to counting errors, but to the linguistic properties of the verb itself. The verb is unusual for a manner-of-motion verb in that the action is inherently "uncontrolled", and thus the subject of the intransitive/object of the transitive is a more passive entity than with the other unergatives (perhaps indicating that the inventory of thematic roles should be refined to distinguish activity verbs with less agentive subjects). We think that this property relates to the notion of internal and external causation that is an important factor in distinguishing unergative and unaccusative verbs. We refer the interested reader to Stevenson and Merlo (1997b), which discusses the latter issue in more detail.

Table 13
Expected class discriminations for each feature.

Feature	Expected Frequency Pattern
Transitivity	Unerg < Unacc < ObjDrop
Causativity	Unerg, ObjDrop < Unacc
Animacy	Unacc < Unerg, ObjDrop
Passive Voice	Unerg < Unacc < ObjDrop
VBN Tag	Unerg < Unacc < ObjDrop

Table 14
Confusion matrix indicating number of errors in classification by verb class, for the full set of five features, compared to two of the four-feature sets. E = unergatives, A = unaccusatives, O = object-drops.

		Assigned Class								
		All features			w/o CAUS			w/o ANIM		
		E	A	O	E	A	O	E	A	O
Actual	E		1	2		4	2		2	2
Class	A	4		3	5		2	5		6
	O	5	3		4	5		3	5	

Table 15
Confusion matrix indicating number of errors in classification by verb class, for the full set of five features and for three of the four-feature sets. *E* = unergatives, *A* = unaccusatives, *O* = object-drops.

		Assigned Class											
		All features			w/o TRANS			w/o PASS			w/o VBN		
		E	A	O	E	A	O	E	A	O	E	A	O
Actual	E		1	2		2	2		1	3		1	5
Class	A	4		3	3		7	1		4	2		4
	O	5	3		2	5		5	3		4	6	

opposite of discriminability), we look at two cells in the matrix: the one in which verbs of the first class were assigned the label of the second class, and the one in which verbs of the second class were assigned the label of the first class. (These pairs of cells are those opposite the diagonal of the confusion matrix.) By examining the decrease (or increase) in confusability of each pair of classes in going from a four-feature experiment to the five-feature experiment, we gain insight into how well (or how poorly) the added feature helps to discriminate each pair of classes.

An analysis of the confusion matrices reveals that the behavior of the features largely conforms to our linguistic predictions, leading us to conclude that the features

we counted worked largely for the reasons we had hypothesized. We expected CAUS and ANIM to be particularly helpful in identifying unaccusatives, and these predictions are confirmed. Compare the second to the first panel of Table 14 (the errors without the CAUS feature compared to the errors with the CAUS feature added to the set). We see that, without the CAUS feature, the confusability between unaccusatives and unergatives, and between unaccusatives and object-drops, is 9 and 7 errors, respectively; but when CAUS is added to the set of features, the confusability between these pairs of classes drops substantially, to 5 and 6 errors, respectively. On the other hand, the confusability between unergatives and object-drops becomes slightly worse (errors increasing from 6 to 7). The latter indicates that the improvement in unaccusatives is not simply due to an across-the-board improvement in accuracy as a result of having more features. We see a similar pattern with the ANIM feature. Comparing the third to the first panel of Table 14 (the errors without the ANIM feature compared to the errors with the ANIM feature added to the set), we see an even larger improvement in discriminability of unaccusatives when the ANIM feature is added. The confusability of unaccusatives and unergatives drops from 7 errors to 5 errors, and of unaccusatives and object-drops from 11 errors to 6 errors. Again, confusability of unergatives and object-drops is worse, with an increase in errors of 5 to 7.

We had predicted that the TRANS feature would make a three-way distinction among the verb classes, based on its predicted linear relationship between the classes (see the inequalities in Table 13). We had further expected that PASS and VBN would behave similarly, since these features are correlated to TRANS. To make a three-way distinction among the verb classes, we would expect confusability between all three pairs of verb classes to decrease (i.e., discriminability would improve) with the addition of TRANS, PASS, or VBN. We find that these predictions are confirmed in part.

First consider the TRANS feature. Comparing the second to the first panel of Table 15, we find that unergatives are already accurately classified, and the addition of TRANS to the set does indeed greatly reduce the confusability of unaccusatives and object-drops, with the number of errors dropping from 12 to 6. However, we also observe that the confusability of unergatives and unaccusatives is not improved, and the confusability of unergatives and object-drops is worsened with the addition of the TRANS feature, with errors in the latter case increasing from 4 to 7. We conclude that the expected three-way discriminability of TRANS is most apparent in the reduced confusion of unaccusative and object-drop verbs.

Our initial prediction was that PASS and VBN would behave similarly to TRANS—that is, also making a three-way distinction among the classes—although the aggregate data revealed little difference in these feature values between unaccusatives and object-drops. Comparing the third to the first panel of Table 15, we observe that the addition of the PASS feature hinders the discriminability of unergatives and unaccusatives (increasing errors from 2 to 5); it does help in discriminating the other pairs of classes, but only slightly (reducing the number of errors by 1 in each case). The VBN feature shows a similar pattern, but is much more helpful at distinguishing unergatives from object-drops, and object-drops from unaccusatives. In comparing the fourth to the first panel of Table 15, we find that the confusability of unergatives and object-drops is reduced from 9 errors to 7, and of unaccusatives and object-drops from 10 errors to 6. The latter result is somewhat surprising, since the aggregate VBN data for the unaccusative and object-drop classes are virtually identical. We conclude that contribution of a feature to classification is not predictable from the apparent discriminability of its numeric values across the classes. This observation emphasizes the importance of an experimental method to evaluating our approach to verb classification.

Table 16
Percent agreement (%Agr) and pair-wise agreement (K) of three experts (E1, E2, E3) and the program compared to each other and to a gold standard (Levin).

	PROGRAM		E1		E2		E3	
	%Agr	K	%Agr	K	%Agr	K	%Agr	K
E1	59%	.36						
E2	68%	.50	75%	.59				
E3	66%	.49	70%	.53	77%	.66		
LEVIN	69.5%	.54	71%	.56	86.5%	.80	83%	.74

5. Establishing the Upper Bound for the Task

In order to evaluate the performance of the algorithm in practice, we need to compare it to the accuracy of classification performed by an expert, which gives a realistic upper bound for the task. The lively theoretical debate on class membership of verbs, and the complex nature of the linguistic information necessary to accomplish this task, led us to believe that the task is difficult and not likely to be performed at 100% accuracy even by experts, and is also likely to show differences in classification between experts. We report here the results of two experiments which measure expert accuracy in classifying our verbs (compared to Levin's classification as the gold standard), as well as inter-expert agreement. (See also Merlo and Stevenson [2000a] for more details.) To enable comparison of responses, we performed a closed-form questionnaire study, where the number and types of the target classes are defined in advance, for which we prepared a forced-choice and a non-forced-choice variant. The forced-choice study provides data for a maximally restricted experimental situation, which corresponds most closely to the automatic verb classification task. However, we are also interested in slightly more natural results—provided by the non-forced-choice task—where the experts can assign the verbs to an "others" category.

We asked three experts in lexical semantics (all native speakers of English) to complete the forced-choice electronic questionnaire study. Neither author was among the three experts, who were all professionals in computational or theoretical linguistics with a specialty in lexical semantics. Materials consisted of individually randomized lists of the same 59 verbs used for the machine learning experiments, using Levin's (1993) electronic index, available from Chicago University Press. The verbs were to be classified into the three target classes—unergative, unaccusative, and object-drop—which were described in the instructions.[10] (All materials and instructions are available at URL http://www.latl.unige.ch/latl/personal/paola.html.)

Table 16 shows an analysis of the results, reporting both percent agreement and pairwise agreement (according to the Kappa statistic) among the experts and the program.[11] Assessing the percentage of verbs on which the experts agree gives us

10 The definitions of the classes were as follows. Unergative: A verb that assigns an agent theta role to the subject in the intransitive. If it is able to occur transitively, it can have a causative meaning. Unaccusative: A verb that assigns a patient/theme theta role to the subject in the intransitive. When it occurs transitively, it has a causative meaning. Object-Drop: A verb that assigns an agent role to the subject and patient/theme role to the object, which is optional. When it occurs transitively, it does not have a causative meaning.

11 In the comparison of the program to the experts, we use the results of the classifier under single hold-out training—which yields an accuracy of 69.5%—because those results provide the classification for each of the individual verbs.

an intuitive measure. However, this measure does not take into account how much the experts agree *over* the expected agreement by chance. The latter is provided by the Kappa statistic, which we calculated following Klauer (1987, 55–57) (using the z distribution to determine significance; $p < 0.001$ for all reported results). The Kappa value measures the experts', and our classifier's, degree of agreement over chance, with the gold standard and with each other. Expected chance agreement varies with the number and the relative proportions of categories used by the experts. This means that two given pairs of experts might reach the same percent agreement on a given task, but not have the same expected chance agreement, if they assigned verbs to classes in different proportions. The Kappa statistic ranges from 0, for no agreement above chance, to 1, for perfect agreement. The interpretation of the scale of agreement depends on the domain, like all correlations. Carletta (1996) cites the convention from the domain of content analysis indicating that $.67 < K < .8$ indicates marginal agreement, while $K > .8$ is an indication of good agreement. We can observe that only one of our agreement figures comes close to reaching what would be considered "good" under this interpretation. Given the very high level of expertise of our human experts, we suspect then that this is too stringent a scale for our task, which is qualitatively quite different from content analysis.

Evaluating the experts' performance summarized in Table 16, we can remark two things, which confirm our expectations. First, the task is difficult—i.e., not performed at 100% (or close) even by trained experts, when compared to the gold standard, with the highest percent agreement with Levin at 86.5%. Second, with respect to comparison of the experts among themselves, the rate of agreement is never very high, and the variability in agreement is considerable, ranging from .53 to .66. This evaluation is also supported by a 3-way agreement measure (Siegel and Castellan 1988). Applying this calculation, we find that the percentage of verbs to which the three experts gave the same classification (60%, $K = 0.6$) is smaller than any of the pairwise agreements, indicating that the experts do not all agree on the same subset of verbs.

The observation that the experts often disagree on this difficult task suggests that a combination of expert judgments might increase the upper bound. We tried the simplest combination, by creating a new classification using a majority vote: each verb was assigned the label given by at least two experts. Only three cases did not have any majority label; in these cases we used the classification of the most accurate expert. This new classification does not improve the upper bound, reaching only 86.4% ($K = .80$) compared to the gold standard.

The evaluation is also informative with respect to the performance of the program. On the one hand, we observe that if we take the best performance achieved by an expert in this task—86.5%—as the maximum achievable accuracy in classification, our algorithm then reduces the error rate over chance by approximately 68%, a very respectable result. In fact, the accuracy of 69.5% achieved by the program is only 1.5% less than one of the human experts in comparison to the gold standard. On the other hand, the algorithm still does not perform at expert level, as indicated by the fact that, for all experts, the lowest agreement score is with the program.

One interesting question is whether experts and program disagree on the same verbs, and show similar patterns of errors. The program makes 18 errors, in total, compared to the gold standard. However, in 9 cases, at least one expert agrees with the classification given by the program. The program makes fewer errors on unergatives (3) and comparably many on unaccusatives and object-drops (7 and 8 respectively), indicating that members of the latter two classes are quite difficult to classify. This differs from the pattern of average agreement between the experts and Levin, who agree on 17.7 (of 20) unergatives, 16.7 (of 19) unaccusatives, and 11.3 (of 20) object-drops. This

clearly indicates that the object-drop class is the most difficult for the human experts to define. This class is the most heterogeneous in our verb list, consisting of verbs from several subclasses of the "unexpressed object alternation" class in (Levin, 1993). We conclude that the verb classification task is likely easier for very homogeneous classes, and more difficult for more broadly defined classes, even when the exemplars share the critical syntactic behaviors.

On the other hand, frequency does not appear to be a simple factor in explaining patterns of agreement between experts, or increases in accuracy. As in Section 4.3, we again analyze the relation between log frequency of the verbs and classification performance, here considering the performance of the experts. We grouped verbs in three log frequency classes: verbs with log frequency less than 2 (i.e., frequency less than 100), those with log frequency between 2 and 3 (i.e., frequency between 100 and 1000), and those with log frequency over 3 (i.e., frequency over 1000). The low-frequency group had 24 verbs (14 unergatives, 5 unaccusatives, and 5 object-drop), the intermediate-frequency group had 25 verbs (5 unergatives, 9 unaccusatives, and 11 object-drops), and the high-frequency group had 10 verbs (1 unergative, 5 unaccusatives, and 4 object-drops). We found that verbs with high and low frequency yield better accuracy and agreement among the experts than the verbs with mid frequency. Neither the accuracy of the majority classification, nor the accuracy of the expert that had the best agreement with Levin, were linearly affected by frequency. For the majority vote, verbs with frequency less than 100 yield an accuracy of 92%, $K = .84$; verbs with frequency between 100 and 1000, accuracy 80%, $K = .69$; and for verbs with frequency over 1000, accuracy 90%, $K = .82$. For the "best" expert, the pattern is similar: verbs with frequency less than 100 yield an accuracy of 87.5%, $K = .74$; verbs with frequency between 100 and 1000, accuracy 84%, $K = .76$; and verbs with frequency over 1000, accuracy 90%, $K = .82$.

We can see here that different frequency groups yield different classification behavior. However, the relation is not simple, and it is clearly affected by the composition of the frequency group: the middle group contains mostly unaccusative and object-drop verbs, which are the verbs with which our experts have the most difficulty. This confirms that the class of the verb is the predominant factor in their pattern of errors. Note also that the pattern of accuracy across frequency groupings is not the same as that of the program (see Section 4.3, which revealed the most errors by the program on the highest frequency verbs), again indicating qualitative differences in performance between the program and the experts.

Finally, one possible shortcoming of the above analysis is that the forced-choice task, while maximally comparable to our computational experiments, may not be a natural one for human experts. To explore this issue, we asked two different experts in lexical semantics (one native speaker of English and one bilingual) to complete the non-forced-choice electronic questionnaire study; again, neither author served as one of the experts. In this task, in addition to the three verb classes of interest, an answer of "other" was allowed. Materials consisted of individually randomized lists of 119 target and filler verbs taken from Levin's (1993) electronic index, as above. The targets were again the same 59 verbs used for the machine learning experiments. To avoid unwanted priming of target items, the 60 fillers were automatically selected from the set of verbs that do not share any class with any of the senses of the 59 target verbs in Levin's index. In this task, if we take only the target items into account, the experts agreed 74.6% of the time ($K = 0.64$) with each other, and 86% ($K = 0.80$) and 69% ($K = 0.57$) with the gold standard. (If we take all the verbs into consideration, they agreed in 67% of the cases [$K = 0.56$] with each other, and 68% [$K = 0.55$] and 60.5% [$K = 0.46$] with the gold standard, respectively.) These results show that the forced-

choice and non-forced-choice task are comparable in accuracy of classification and inter-judge agreement on the target classes, giving us confidence that the forced-choice results provide a reasonably stable upper bound for computational experiments.

6. Discussion

The work presented here contributes to some central issues in computational linguistics, by providing novel insights, data, and methodology in some cases, and by reinforcing some previously established results in others. Our research stems from three main hypotheses:

1. Argument structure is the relevant level of representation for verb classification.
2. Argument structure is manifested distributionally in syntactic alternations, giving rise to differences in subcategorization frames or the distributions of their usage, or in the properties of the NP arguments to a verb.
3. This information is detectable in a corpus and can be learned automatically.

We discuss the relevant debate on each of these hypotheses, and the contribution of our results to each, in the following subsections.

6.1 Argument Structure and Verb Classification

Argument structure has previously been recognized as one of the most promising candidates for accurate classification. For example, Basili, Pazienza, and Velardi (1996) argue that relational properties of verbs—their argument structure—are more informative for classification than their definitional properties (e.g., the fact that a verb describes a manner of motion or a way of cooking). Their arguments rest on linguistic and psycholinguistic results on classification and language acquisition (in particular, Pinker, [1989]; Rosch [1978]).

Our results confirm the primary role of argument structure in verb classification. Our experimental focus is particularly clear in this regard because we deal with verbs that are "minimal pairs" with respect to argument structure. By classifying verbs that show the same subcategorizations (transitive and intransitive) into different classes, we are able to eliminate one of the confounds in classification work created by the fact that subcategorization and argument structure are often co-variant. We can infer that the accuracy in our classification is due to argument structure information, as subcategorization is the same for all verbs. Thus, we observe that the *content* of the thematic roles assigned by a verb is crucial for classification.

6.2 Argument Structure and Distributional Statistics

Our results further support the assumption that thematic differences across verb classes are apparent not only in differences in subcategorization frames, but also in differences in their frequencies. This connection relies heavily on the hypothesis that lexical semantics and lexical syntax are correlated, following Levin (1985; 1993). However, this position has been challenged by Basili, Pazienza, and Velardi (1996) and Boguraev and Briscoe (1989), among others. For example, in an attempt to assess the actual completeness and usefulness of the Longman Dictionary of Contemporary English (LDOCE) entries, Boguraev and Briscoe (1989) found that people assigned a

"change of possession" meaning both to verbs that had dative-related subcategorization frames (as indicated in the LDOCE) and to verbs that did not. Conversely, they also found that both verbs that have a change-of-possession component in their meaning and those that do not could have a dative code. They conclude that the thesis put forth by Levin (1985) is only partially supported. Basili, Pazienza, and Velardi (1996) show further isolated examples meant to illustrate that lexical syntax and semantics are not in a one-to-one relation.

Many recent results, however, seem to converge in supporting the view that the relation between lexical syntax and semantics can be usefully exploited (Aone and McKee 1996; Dorr 1997; Dorr, Garman, and Weinberg 1995; Dorr and Jones 1996; Lapata and Brew 1999; Schulte im Walde 2000; Siegel 1998; Siegel 1999). Our work in particular underscores the relation between the syntactic manifestations of argument structure, and lexical semantic class. In light of these recent successes, the conclusions in Boguraev and Briscoe (1989) are clearly too pessimistic. In fact, their results do not contradict the more recent ones. First of all, it is not the case that if an implication holds from argument structure to subcategorization (change of possession implies dative shift), the converse also holds. It comes as no surprise that verbs that do not have any change-of-possession component in their meaning may also show dative shift syntactically. Secondly, as Boguraev and Briscoe themselves note, Levin's statement should be interpreted as a statistical trend, and as such, Boguraev and Briscoe's results also confirm it. They claim however, that in adopting a statistical point of view, predictive power is lost. Our work shows that this conclusion is not appropriate either: the correlation is strong enough to be useful to predict semantic classification, at least for the argument structures that have been investigated.

6.3 Detection of Argument Structure in Corpora

Given the manifestation of argument structure in statistical distributions, we view corpora, especially if annotated with currently available tools, as repositories of implicit grammars, which can be exploited in automatic verb-classification tasks. Besides establishing a relationship between syntactic alternations and underlying semantic properties of verbs, our approach extends existing corpus-based learning techniques to the detection and automatic acquisition of argument structure. To date, most work in this area has focused on learning of subcategorization from unannotated or syntactically annotated text (e.g., Brent [1993]; Sanfilippo and Poznanski [1992]; Manning [1993]; Collins [1997]). Others have tackled the problem of lexical semantic classification, but using only subcategorization frequencies as input data (Lapata and Brew 1999; Schulte im Walde 2000). Specifically, these researchers have not explicitly addressed the definition of features to tap directly into thematic role differences that are not reflected in subcategorization distinctions. On the other hand, when learning of thematic role assignment has been the explicit goal, the text has been semantically annotated (Webster and Marcus 1989), or external semantic resources have been consulted (Aone and McKee 1996; McCarthy 2000). We extend these results by showing that thematic information can be induced from linguistically-guided counts in a corpus, without the use of thematic role tagging or external resources such as WordNet.

Finally, our results converge with the increasing agreement that corpus-based techniques are fruitful in the automatic construction of computational lexicons, providing machine readable dictionaries with complementary, reusable resources, such as frequencies of argument structures. Moreover, these techniques produce data that is easily updated, as the information contained in corpora changes all the time, allowing for adaptability to new domains or usage patterns. This dynamic aspect could be exploited if techniques such as the one presented here are developed, which can work

on a rough collection of texts, and do not require a carefully balanced corpus or time-consuming semantic tagging.

7. Related Work

We conclude from the discussion above that our own work and work of others support our hypotheses concerning the importance of the relation between classes of verbs and the syntactic expression of argument structure in corpora. In light of this, it is instructive to evaluate our results in the context of other work that shares this view. Some related work requires either exact exemplars for acquisition, or external pre-compiled resources. For example, Dorr (1997) summarizes a number of automatic classification experiments based on encoding Levin's alternations directly, as symbolic properties of a verb (Dorr, Garman, and Weinberg 1995; Dorr and Jones 1996). Each verb is represented as the binary settings of a vector of possible alternations, acquired through a large corpus analysis yielding exemplars of the alternation. To cope with sparse data, the corpus information is supplemented by syntactic information obtained from the LDOCE and semantic information obtained from WordNet. This procedure classifies 95 unknown verbs with 61% accuracy. Dorr also remarks that this result could be improved to 83% if missing LDOCE codes were added. While Dorr's work requires finding exact exemplars of the alternation, Oishi and Matsumoto (1997) present a method that, like ours, uses surface indicators to approximate underlying properties. From a dictionary of dependency relations, they extract case-marking particles as indicators of the grammatical function properties of the verbs (which they call thematic properties), such as subject and object. Adverbials indicate aspectual properties. The combination of these two orthogonal dimensions gives rise to a classification of Japanese verbs.

Other work has sought to combine corpus-based extraction of verbal properties with statistical methods for classifying verbs. Siegel's work on automatic aspectual classification (1998, 1999) also reveals a close relationship between verb-related syntactic and semantic information. In this work, experiments to learn aspectual classification from linguistically-based numerical indicators are reported. Using combinations of seven statistical indicators (some morphological and some reflecting syntactic co-occurrences), it is possible to learn the distinction between events and states for 739 verb tokens with an improvement of 10% over the baseline (error rate reduction of 74%), and to learn the distinction between culminated and non-culminated events for 308 verb tokens with an improvement of 11% (error rate reduction of 29%) (Siegel 1999).

In work on lexical semantic verb classification, Lapata and Brew (1999) further support the thesis of a predictive correlation between syntax and semantics in a statistical framework, showing that the frequency distributions of subcategorization frames within and across classes can disambiguate the usages of a verb with more than one known lexical semantic class. On 306 verbs that are disambiguated by subcategorization frame, they achieve 91.8% accuracy on a task with a 65.7% baseline, for a 76% reduction in error rate. On 31 verbs that can take the same subcategorization(s) in different classes—more similar to our situation in that subcategorization alone cannot distinguish the classes—they achieve 83.9% accuracy compared to a 61.3% baseline, for a 58% reduction in error. Aone and McKee (1996), working with a much coarser-grained classification of verbs, present a technique for predicate-argument extraction from multi-lingual texts. Like ours, their work goes beyond statistics over subcategorizations to include counts over the more directly semantic feature of animacy. No numerical evaluation of their results is provided.

Schulte im Walde (2000) applies two clustering methods to two types of frequency data for 153 verbs from 30 Levin (1993) classes. One set of experiments uses verb subcategorization frequencies, and the other uses subcategorization frequencies plus selectional preferences (a numerical measure based on an adaptation of the relative entropy method of Resnik [1996]). The best results achieved are a correct classification of 58 verbs out of 153, with a precision of 61% and recall of 36%, obtained using only subcategorization frequencies. We calculate that this corresponds to an F-score of 45% with balanced precision and recall.[12] The use of selectional preference information decreases classification performance under either clustering algorithm. The results are somewhat difficult to evaluate further, as there is no description of the classes included. Also, the method of counting correctness entails that some "correct" classes may be split across distant clusters (this level of detail is not reported), so it is unclear how coherent the class behaviour actually is.

McCarthy (2000) proposes a method to identify diathesis alternations. After learning subcategorization frames, based on a parsed corpus, selectional preferences are acquired for slots of the subcategorization frames, using probability distributions over Wordnet classes. Alternations are detected by testing the hypothesis that, given any verb, the selectional preferences for arguments occurring in alternating slots will be more similar to each other than those for slots that do not alternate. For instance, given a verb participating in the causative alternation, its selectional preferences for the subject in an intransitive use, and for the object in a transitive use, will be more similar to each other than the selectional preferences for these two slots of a verb that does not participate in the causative alternation. This method achieves the best accuracy for the causative and the conative alternations (73% and 83%, respectively), despite sparseness of data. McCarthy reports that a simpler measure of selectional preferences based simply on head words yields a lower 63% accuracy. Since this latter measure is very similar to our CAUS feature, we think that our results would also improve by adopting a similar method of abstracting from head words to classes.

Our work extends each of these approaches in some dimension, thereby providing additional support for the hypothesis that syntax and semantics are correlated in a systematic and predictive way. We extend Dorr's alternation-based automatic classification to a statistical setting. By using distributional approximations of indicators of alternations, we solve the sparse data problem without recourse to external sources of knowledge, such as the LDOCE, and in addition, we are able to learn argument structure alternations using exclusively positive examples. We improve on the approach of Oishi and Matsumoto (1997) by learning argument structure properties, which, unlike grammatical functions, are not marked morphologically, and by not relying on external sources of knowledge. Furthermore, in contrast to Siegel (1998) and Lapata and Brew (1999) our method applies successfully to previously unseen words—i.e., test cases that were not represented in the training set.[13] This is a very important property of lexical acquisition algorithms to be used for lexicon organization, as their main interest lies in being applied to unknown words.

On the other hand, our approach is similar to the approaches of Siegel, and Lapata and Brew (1999), in attempting to learn semantic notions from distributions of

12 A baseline of 5% is reported, based on a closest-neighbor pairing of verbs, but it is not straightforward to compare this task to the proposed clustering algorithm. Determining a meaningful baseline for unsupervised clustering is clearly a challenge, but this gives an indication that the clustering task is indeed difficult.

13 Siegel (1998) reports two experiments over verb types with disjoint training and test sets, but the results were not significantly different from the baseline.

indicators that can be gleaned from a text. In our case, we are trying to learn argument structure, a finer-grained classification than the dichotomic distinctions studied by Siegel. Like Lapata and Brew, three of our indicators—TRANS, VBN, PASS—are based on the assumption that distributional differences in subcategorization frames are related to underlying verb class distinctions. However, we also show that other syntactic indicators—CAUS and ANIM—can be devised that tap directly into the argument structure of a verb. Unlike Schulte im Walde (2000), we find the use of these semantic features helpful in classification—using only TRANS and its related features, VBN and PASS, we achieve only 55% accuracy, in comparison to 69.8% using the full set of features. This can perhaps be seen as support for our hypothesis that argument structure is the right level of representation for verb class distinctions, since it appears that our features that capture thematic differences are useful in classification, while Schulte im Walde's selectional restriction features were not.

Aone and McKee (1996) also use features that are intended to tap into both subcategorization and thematic role distinctions—frequencies of the transitive use and animate subject use. In our task, we show that subject animacy can be profitably approximated solely with pronoun counts, avoiding the need for reference to external sources of semantic information used by Aone and McKee. In addition, our work extends theirs in investigating much finer-grained verb classes, and in classifying verbs that have multiple argument structures. While Aone and McKee *define* each of their classes according to a single argument structure, we demonstrate the usefulness of syntactic features that capture relations across different argument structures of a single verb. Furthermore, while Aone and McKee, and others, look at relative frequency of subcategorization frames (as with our TRANS feature), or relative frequency of a property of NPs within a particular grammatical function (as with our ANIM feature), we also look at the paradigmatic relations across a text between thematic arguments in different alternations (with our CAUS feature).

McCarthy (2000) shows that a method very similar to ours can be used for *identifying* alternations. Her qualitative results confirm, however, what was argued in Section 2 above: counts that tap directly into the thematic assignments are necessary to fully identify a diathesis alternation. In fact, on close inspection, McCarthy's method does not distinguish between the induced-action alternation (which the unergatives exhibit) and the causative/inchoative alternation (which the unaccusatives exhibit); thus, her method does not discriminate two of our classes. It is likely that a combination of our method, which makes the necessary thematic distinctions, and her more sophisticated method of detecting alternations would give very good results.

8. Limitations and Future Work

The classification results show that our method is powerful, and suited to the classification of unknown verbs. However, we have not yet addressed the problem of verbs that can have multiple classifications. We think that many cases of ambiguous classification of the lexical entry for a verb can be addressed with the notion of intersective sets introduced by Dang et al. (1998). This is an important concept, which proposes that "regular" ambiguity in classification—i.e., sets of verbs that have the same multi-way classifications according to Levin (1993)—can be captured with a finer-grained notion of lexical semantic classes. Thus, subsets of verbs that occur in the intersection of two or more Levin classes form in themselves a coherent semantic (sub)class. Extending our work to exploit this idea requires only defining the classes appropriately; the basic approach will remain the same. Given the current demonstration of our method on fine-grained classes that share subcategoriza-

tion alternations, we are optimistic regarding its future performance on intersective sets.

Because we assume that thematic properties are reflected in alternations of argument structure, our features require searching for relations across occurrences of each verb. This motivated our initial experimental focus on verb types. However, when we turn to consider ambiguity, we must also address the problem that individual instances of verbs may come from different classes, and we may (like Lapata and Brew [1999]) want to classify the individual tokens of a verb. In future research we plan to extend our method to the case of ambiguous tokens, by experimenting with the combination of several sources of information: the classification of each instance will be a function of a bias for the verb type (using the cross-corpus statistics we collect), but also of features of the usage of the instance being classified (cf., Lapata and Brew [1999]; Siegel [1998]).

Finally, corpus-based learning techniques collect statistical information related to language use, and are a good starting point for studying human linguistic performance. This opens the way to investigating the relation of linguistic data in text to people's linguistic behaviour and use. For example, Merlo and Stevenson (1998) show that, contrary to the naive assumption, speakers' preferences in syntactic disambiguation are not simply directly related to frequency (i.e., a speaker's preference for one construction over another is not simply modelled by the frequency of the construction, or of the words in the construction). Thus, the kind of corpus investigation we are advocating—founded on in-depth linguistic analysis—holds promise for building more natural NLP systems which go beyond the simplest assumptions, and tie together statistical computational linguistic results with experimental psycholinguistic data.

9. Conclusions

In this paper, we have presented an in-depth case study, in which we investigate machine learning techniques for automatically classifying a set of verbs into classes determined by their argument structures. We focus on the three major classes of optionally intransitive verbs in English, which cannot be discriminated by their subcategorizations, and therefore require distinctive features that are sensitive to the thematic properties of the verbs. We develop such features and automatically extract them from very large, syntactically annotated corpora. Results show that a small number of linguistically motivated lexical features are sufficient to achieve a 69.8% accuracy rate in a three-way classification task with a baseline (chance) performance of 33.9%, for which the best performance achieved by a human expert is 86.5%.

Returning to our original questions of what can and need be learned about the relational properties of verbs, we conclude that argument structure is both a highly useful and learnable aspect of verb knowledge. We observe that relevant semantic properties of verb classes (such as causativity, or animacy of subject) may be successfully approximated through countable syntactic features. In spite of noisy data (arising from diverse sources such as tagging errors, or limitations of our extraction patterns), the lexical properties of interest are reflected in the corpora robustly enough to positively contribute to classification.

We remark, however, that deep linguistic analysis cannot be eliminated—in our approach it is embedded in the selection of the features to count. Specifically, our features are derived through a detailed analysis of the differences in thematic role assignments across the verb classes under investigation. Thus, an important contribution of the work is the proposed mapping between the thematic assignment properties of

the verb classes, and the statistical distributions of their surface syntactic properties. We think that using such linguistically motivated features makes the approach very effective and easily scalable: we report a 54% reduction in error rate (a 68% reduction, when the human expert-based upper bound is considered), using only five features that are readily extractable from automatically annotated corpora.

Acknowledgments

We gratefully acknowledge the financial support of the following organizations: the Swiss NSF (fellowship 8210-46569 to PM); the United States NSF (grants #9702331 and #9818322 to SS); the Canadian NSERC (grant to SS); the University of Toronto; and the Information Sciences Council of Rutgers University. Much of this research was carried out while PM was a visiting scientist at IRCS, University of Pennsylvania, and while SS was a faculty member at Rutgers University, both of whose generous and supportive environments were of great benefit to us. We thank Martha Palmer, Michael Collins, Natalia Kariaeva, Kamin Whitehouse, Julie Boland, Kiva Dickinson, and three anonymous reviewers, for their helpful comments and suggestions, and for their contributions to this research. We also greatly thank our experts for the gracious contribution of their time in answering our electronic questionnaire.

References

Abney, Steven. 1996. Partial parsing via finite-state cascades. In John Carroll, editor, *Proceedings of the Workshop on Robust Parsing at the Eighth Summer School on Logic, Language and Information*, number 435 in CSRP, pages 8–15. University of Sussex, Brighton.

Aone, Chinatsu and Douglas McKee. 1996. Acquiring predicate-argument mapping information in multilingual texts. In Branimir Boguraev and James Pustejovsky, editors, *Corpus Processing for Lexical Acquisition*. MIT Press, pages 191–202.

Basili, Roberto, Maria-Teresa Pazienza, and Paola Velardi. 1996. A context-driven conceptual clustering method for verb classification. In Branimir Boguraev and James Pustejovsky, editors, *Corpus Processing for Lexical Acquisition*. MIT Press, pages 117–142.

Boguraev, Branimir and Ted Briscoe. 1989. Utilising the LDOCE grammar codes. In Branimir Boguraev and Ted Briscoe, editors, *Computational Lexicography for Natural Language Processing*. Longman, London, pages 85–116.

Boguraev, Branimir and James Pustejovsky. 1996. Issues in text-based lexicon acquisition. In Branimir Boguraev and James Pustejovsky, editors, *Corpus Processing for Lexical Acquisition*. MIT Press, pages 3–20.

Brent, Michael. 1993. From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Linguistics*, 19(2):243–262.

Briscoe, Ted and John Carroll. 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of the Fifth Applied Natural Language Processing Conference*, pages 356–363.

Brousseau, Anne-Marie and Elizabeth Ritter. 1991. A non-unified analysis of agentive verbs. In *West Coast Conference on Formal Linguistics*, number 20, pages 53–64.

Burzio, Luigi. 1986. *Italian Syntax: A Government-Binding Approach*. Reidel: Dordrecht.

Carletta, Jean. 1996. Assessing agreement on classification tasks: the Kappa statistics. *Computational Linguistics*, 22(2):249–254.

Collins, Michael John. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the ACL*, pages 16–23, Madrid, Spain.

Cruse, D. A. 1972. A note on English causatives. *Linguistic Inquiry*, 3(4):520–528.

Dang, Hoa Trang, Karin Kipper, Martha Palmer, and Joseph Rosenzweig. 1998. Investigating regular sense extensions based on intersective Levin classes. In *Proceedings of the 36th Annual Meeting of the ACL and the 17th International Conference on Computational Linguistics (COLING-ACL '98)*, pages 293–299, Montreal. Université de Montreal.

Dixon, Robert M. W. 1994. *Ergativity*. Cambridge University Press, Cambridge.

Dorr, Bonnie. 1997. Large-scale dictionary construction for foreign language tutoring and interlingual machine translation. *Machine Translation*, 12(4):1–55.

Dorr, Bonnie, Joe Garman, and Amy Weinberg. 1995. From syntactic encodings to thematic roles: Building lexical entries for interlingual MT. *Journal of Machine Translation*, 9(3):71–100.

Dorr, Bonnie and Doug Jones. 1996. Role of word sense disambiguation in lexical acquisition: Predicting semantics from syntactic cues. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 322–327, Copenhagen.

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

Greenberg, Joseph H. 1966. *Language Universals*. Mouton, The Hague, Paris.

Gruber, Jeffrey. 1965. *Studies in Lexical Relation*. MIT Press, Cambridge, MA.

Hale, Ken and Jay Keyser. 1993. On argument structure and the lexical representation of syntactic relations. In K. Hale and J. Keyser, editors, *The View from Building 20*. MIT Press, pages 53–110.

Jakobson, Roman. 1971. Signe Zéro. In *Selected Writings*, volume 2, 2d ed. Mouton, The Hague, pages 211–219.

Kariaeva, Natalia. 1999. Discriminating between unaccusative and object-drop verbs: Animacy factor. Ms., Rutgers University. New Brunswick, NJ.

Klavans, Judith and Martin Chodorow. 1992. Degrees of stativity: The lexical representation of verb aspect. In *Proceedings of the Fourteenth International Conference on Computational Linguistics (COLING '92)*, pages 1126–1131, Nantes, France.

Klavans, Judith and Min-Yen Kan. 1998. Role of verbs in document analysis. In *Proceedings of the 36th Annual Meeting of the ACL and the 17th International Conference on Computational Linguistics (COLING-ACL '98)*, pages 680–686, Montreal. Université de Montreal.

Lapata, Maria. 1999. Acquiring lexical generalizations from corpora: A case study for diathesis alternations. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 397–404, College Park, MD.

Lapata, Maria and Chris Brew. 1999. Using subcategorization to resolve verb class ambiguity. In *Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 266–274, College Park, MD.

Levin, Beth. 1985. Introduction. In Beth Levin, editor, *Lexical Semantics in Review*, number 1 in Lexicon Project Working Papers. Centre for Cognitive Science, MIT, Cambridge, MA, pages 1–62.

Levin, Beth. 1993. *English Verb Classes and Alternations*. University of Chicago Press, Chicago, IL.

Levin, Beth and Malka Rappaport Hovav. 1995. *Unaccusativity*. MIT Press, Cambridge, MA.

Manning, Christopher D. 1993. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 235–242. Ohio State University.

McCarthy, Diana. 2000. Using semantic preferences to identify verbal participation in role switching alternations. In *Proceedings of ANLP-NAACL 2000*, pages 256–263, Seattle, WA.

McCarthy, Diana and Anna Korhonen. 1998. Detecting verbal participation in diathesis alternations. In *Proceedings of the 36th Annual Meeting of the ACL and the 17th International Conference on Computational Linguistics (COLING-ACL '98)*, pages 1493–1495, Montreal, Université de Montreal.

Merlo, Paola and Suzanne Stevenson. 1998. What grammars tell us about corpora: the case of reduced relative clauses. In *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 134–142, Montreal.

Merlo, Paola and Suzanne Stevenson. 2000a. Establishing the upper-bound and inter-judge agreement in a verb classification task. In *Second International Conference on Language Resources and Evaluation (LREC-2000)*, volume 3, pages 1659–1664.

Merlo, Paola and Suzanne Stevenson. 2000b. Lexical syntax and parsing architecture. In Matthew Crocker, Martin Pickering, and Charles Clifton, editors, *Architectures and Mechanisms for Language Processing*. Cambridge University Press, Cambridge, pages 161–188.

Moravcsik, Edith and Jessica Wirth. 1983. Markedness—an Overview. In Fred Eckman, Edith Moravcsik, and Jessica Wirth, editors, *Markedness*. Plenum Press, New York, NY, pages 1–13.

Oishi, Akira and Yuji Matsumoto. 1997. Detecting the organization of semantic subclasses of Japanese verbs. *International Journal of Corpus Linguistics*, 2(1):65–89.

Palmer, Martha. 2000. Consistent criteria for sense distinctions. *Special Issue of Computers and the Humanities, SENSEVAL98: Evaluating Word Sense Disambiguation Systems*, 34(1–2):217–222.

Perlmutter, David. 1978. Impersonal passives and the unaccusative hypothesis. In *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, volume 4, pages 157–189.

Pinker, Steven. 1989. *Learnability and Cognition: the Acquisition of Argument Structure*. MIT Press, Cambridge, MA.

Quinlan, J. Ross. 1992. *C4.5: Programs for Machine Learning*. Series in Machine Learning. Morgan Kaufmann, San Mateo, CA.

Ratnaparkhi, Adwait. 1996. A maximum entropy part-of-speech tagger. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*, pages 133–142, Philadelphia, PA.

Resnik, Philip. 1996. Selectional constraints: an information-theoretic model and its computational realization. *Cognition*, 61(1–2):127–160.

Riloff, Ellen and Mark Schmelzenbach. 1998. An empirical approach to conceptual case frame acquisition. In *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 49–56.

Rosch, Eleanor. 1978. Principles of categorization. In *Cognition and Categorization*. Lawrence Erlbaum Assoc, Hillsdale, NJ.

Sanfilippo, Antonio and Victor Poznanski. 1992. The acquisition of lexical knowledge from combined machine-readable dictionary sources. In *Proceedings of the Third Applied Natural Language Processing Conference*, pages 80–87, Trento, Italy.

Schulte im Walde, Sabine. 2000. Clustering verbs semantically according to their alternation behaviour. In *Proceedings of COLING 2000*, pages 747–753, Saarbruecken, Germany.

Siegel, Eric 1998. *Linguistic Indicators for Language Understanding: Using machine learning methods to combine corpus-based indicators for aspectual classification of clauses*. Ph.D. thesis, Dept. of Computer Science, Columbia University.

Siegel, Eric. 1999. Corpus-based linguistic indicators for aspectual classification. In *Proceedings of ACL'99*, pages 112–119, College Park, MD. University of Maryland.

Siegel, Sidney and John Castellan. 1988. *Nonparametric statistics for the Behavioral Sciences*. McGraw-Hill, New York.

Silverstein, Michael. 1976. Hierarchy of features and ergativity. In Robert Dixon, editor, *Grammatical Categories in Australian Languages*. Australian Institute of Aboriginal Studies, Canberra, pages 112–171.

Srinivas, Bangalore and Aravind K. Joshi. 1999. Supertagging: An approach to almost parsing. *Computational Linguistics*, 25(2):237–265.

Stede, Manfred. 1998. A generative perspective on verb alternations. *Computational Linguistics*, 24(3):401–430.

Stevenson, Suzanne and Paola Merlo. 1997a. Architecture and experience in sentence processing. In *Proceedings of the 19th Annual Conference of the Cognitive Science Society*, pages 715–720.

Stevenson, Suzanne and Paola Merlo. 1997b. Lexical structure and processing complexity. *Language and Cognitive Processes*, 12(1–2):349–399.

Stevenson, Suzanne, Paola Merlo, Natalia Kariaeva, and Kamin Whitehouse. 1999. Supervised learning of lexical semantic verb classes using frequency distributions. In *Proceedings of SigLex99: Standardizing Lexical Resources (SigLex'99)*, pages 15–21, College Park, MD.

Trubetzkoy, Nicolaj S. 1939. *Grundzüge der Phonologie*. Travaux du Cercle Linguistique de Prague, Prague.

Webster, Mort and Mitch Marcus. 1989. Automatic acquisition of the lexical semantics of verbs from sentence frames. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 177–184, Vancouver, Canada.

Appendix A

The following three tables contain the overall frequency and the normalized feature values for each of the 59 verbs in our experimental set.

	Unergative Verbs					
	Freq	VBN	PASS	TRANS	CAUS	ANIM
Min Value	8	0.00	0.00	0.00	0.00	0.00
Max Value	4088	1.00	0.39	0.74	0.00	1.00
floated	176	0.43	0.26	0.74	0.00	0.17
hurried	86	0.40	0.31	0.50	0.00	0.37
jumped	4088	0.09	0.00	0.03	0.00	0.20
leaped	225	0.09	0.00	0.05	0.00	0.13
marched	238	0.10	0.01	0.09	0.00	0.12
paraded	33	0.73	0.39	0.46	0.00	0.50
raced	123	0.01	0.00	0.06	0.00	0.15
rushed	467	0.22	0.12	0.20	0.00	0.10
vaulted	54	0.00	0.00	0.41	0.00	0.03
wandered	67	0.02	0.00	0.03	0.00	0.32
galloped	12	1.00	0.00	0.00	0.00	0.00
glided	14	0.00	0.00	0.08	0.00	0.50
hiked	25	0.28	0.12	0.29	0.00	0.40
hopped	29	0.00	0.00	0.21	0.00	1.00
jogged	8	0.29	0.00	0.29	0.00	0.33
scooted	10	0.00	0.00	0.43	0.00	0.00
scurried	21	0.00	0.00	0.00	0.00	0.14
skipped	82	0.22	0.02	0.64	0.00	0.16
tiptoed	12	0.17	0.00	0.00	0.00	0.00
trotted	37	0.19	0.17	0.07	0.00	0.18

	Unaccusative Verbs					
	Freq	VBN	PASS	TRANS	CAUS	ANIM
Min Value	13	0.16	0.00	0.02	0.00	0.00
Max Value	5543	0.95	0.80	0.76	0.41	0.36
boiled	58	0.92	0.70	0.42	0.00	0.00
cracked	175	0.61	0.19	0.76	0.02	0.14
dissolved	226	0.51	0.58	0.71	0.05	0.11
exploded	409	0.34	0.02	0.66	0.37	0.04
flooded	235	0.47	0.57	0.44	0.04	0.03
fractured	55	0.95	0.76	0.51	0.00	0.00
hardened	123	0.92	0.55	0.56	0.12	0.00
melted	70	0.80	0.44	0.02	0.00	0.19
opened	3412	0.21	0.09	0.69	0.16	0.36
solidified	34	0.65	0.21	0.68	0.00	0.12
collapsed	950	0.16	0.00	0.16	0.01	0.02
cooled	232	0.85	0.21	0.29	0.13	0.11
folded	189	0.73	0.33	0.23	0.00	0.00
widened	1155	0.18	0.02	0.13	0.41	0.01
changed	5543	0.73	0.23	0.47	0.22	0.08
cleared	1145	0.58	0.40	0.50	0.31	0.06
divided	1539	0.93	0.80	0.17	0.10	0.05
simmered	13	0.83	0.00	0.09	0.00	0.00
stabilized	286	0.92	0.13	0.18	0.35	0.00

	Object-Drop Verbs					
	Freq	VBN	PASS	TRANS	CAUS	ANIM
Min Value	39	0.10	0.04	0.21	0.00	0.00
Max Value	15063	0.95	0.99	1.00	0.24	0.42
carved	185	0.85	0.66	0.98	0.00	0.00
danced	88	0.22	0.14	0.37	0.00	0.00
kicked	308	0.30	0.18	0.97	0.00	0.33
knitted	39	0.95	0.99	0.93	0.00	0.00
painted	506	0.72	0.18	0.71	0.00	0.38
played	2689	0.38	0.16	0.24	0.00	0.00
reaped	172	0.56	0.05	0.90	0.00	0.22
typed	57	0.81	0.74	0.81	0.00	0.00
washed	137	0.79	0.60	1.00	0.00	0.00
yelled	74	0.10	0.04	0.38	0.00	0.00
borrowed	1188	0.77	0.15	0.60	0.13	0.19
inherited	357	0.60	0.13	0.64	0.06	0.32
organized	1504	0.85	0.38	0.65	0.18	0.07
rented	232	0.72	0.22	0.61	0.00	0.42
sketched	44	0.67	0.17	0.44	0.00	0.20
cleaned	160	0.83	0.47	0.21	0.05	0.21
packed	376	0.84	0.12	0.40	0.05	0.19
studied	901	0.66	0.17	0.57	0.05	0.11
swallowed	152	0.79	0.44	0.35	0.04	0.22
called	15063	0.56	0.22	0.72	0.24	0.16

Appendix B

Performance of all the subsets of features, in order of decreasing accuracy. To determine whether the difference between any two results is statistically significant, the 95% confidence interval can be calculated for each of the two results, and the two ranges checked to see whether they overlap. To do this, take each accuracy plus and minus 2.01 times its associated standard error to get the 95% confidence range ($df = 49$, $t = 2.01$). If the two ranges overlap, then the difference in accuracy is not significant at the $p < .05$ level.

Accuracy	SE	Features	Accuracy	SE	Features
69.8	0.5	TRANS PASS VBN CAUS ANIM	57.3	0.5	TRANS CAUS
69.8	0.5	TRANS VBN CAUS ANIM	57.3	0.5	PASS VBN ANIM
67.3	0.6	TRANS PASS VBN ANIM	56.7	0.5	PASS CAUS ANIM
66.7	0.5	TRANS VBN ANIM	55.7	0.5	VBN CAUS
66.5	0.5	TRANS PASS CAUS ANIM	55.7	0.1	CAUS
64.4	0.5	TRANS VBN CAUS	55.4	0.4	PASS CAUS
63.2	0.6	TRANS PASS VBN CAUS	55.0	0.6	TRANS PASS VBN
63.0	0.5	TRANS PASS ANIM	54.7	0.4	TRANS PASS
62.9	0.4	TRANS CAUS ANIM	54.2	0.5	TRANS VBN
62.1	0.5	CAUS ANIM	52.5	0.5	VBN
61.7	0.5	TRANS PASS CAUS	50.9	0.5	PASS ANIM
61.6	0.6	PASS VBN CAUS ANIM	50.2	0.6	PASS VBN
60.1	0.4	VBN CAUS ANIM	50.2	0.5	PASS
59.5	0.6	TRANS ANIM	47.1	0.4	TRANS
59.4	0.5	VBN ANIM	35.3	0.5	ANIM
57.4	0.6	PASS VBN CAUS			