In Proceedings of SigLex99: Standardizing Lexical Resources, 15-22.

Supervised Learning of Lexical Semantic Verb Classes Using Frequency Distributions

Suzanne Stevenson
Rutgers University
suzanne@cs.rutgers.edu

Paola Merlo
University of Geneva
merlo@lettres.unige.ch

Natalia Kariaeva
Rutgers University
kariaeva@rci.rutgers.edu

Kamin Whitehouse
Rutgers University
kaminw@rci.rutgers.edu

Abstract

We report a number of computational experiments in supervised learning whose goal is to automatically classify a set of verbs into lexical semantic classes, based on frequency distribution approximations of grammatical features extracted from a very large annotated corpus. Distributions of five syntactic features that approximate transitivity alternations and thematic role assignments are sufficient to reduce error rate by 56% over chance. We conclude that corpus data is a usable repository of verb class information, and that corpus-driven extraction of grammatical features is a promising methodology for automatic lexical acquisition.

1 Introduction

Recent years have witnessed a shift in grammar development methodology, from crafting large grammars, to annotation of corpora. Correspondingly, there has been a change from developing rule-based parsers to developing statistical methods for inducing grammatical knowledge from annotated corpus data. The shift has mostly occurred because building wide-coverage grammars is time-consuming, error prone, and difficult. The same can be said for crafting the rich lexical representations that are a central component of linguistic knowledge, and research in automatic lexical acquisition has sought to address this ((Dorr and Jones, 1996; Dorr, 1997), among others). Yet there have been few attempts to learn fine-grained lexical classifications from the statistical analysis of distributional data, analogously to the induction of syntactic knowledge (though see, e.g., (Brent, 1993; Klavans and Chodorow, 1992; Resnik, 1992)). In this paper, we propose such an approach for the automatic classification of verbs into lexical semantic classes.[1]

We can express the issues raised by this approach as follows.

1. Which linguistic distinctions among lexical classes can we expect to find in a corpus?
2. How easily can we extract the frequency distributions that approximate the relevant linguistic properties?
3. Which frequency distributions work best to distinguish the verb classes?

In exploring these questions, we focus on verb classification for several reasons. Verbs are very important sources of knowledge in many language engineering tasks, and the relationships among verbs appear to play a major role in the organization and use of this knowledge. Knowledge about verb classes is crucial for lexical acquisition in support of language generation and machine translation (Dorr, 1997) and document classification (Klavans and Kan, 1998), yet manual classification of large numbers of verbs is a difficult and resource intensive task (Levin, 1993; Miller et al., 1990; Dang et al., 1998).

To address these issues, we suggest that one can train an automatic classifier for verbs on the basis of statistical approximations to verb diatheses. We use diatheses—alternations in the expression of the arguments of the verb—following Levin and Dorr, for two reasons. First, verb diatheses are syntactic cues

[1]We are aware that a distributional approach rests on one strong assumption regarding the nature of the representations under study: semantic notions and syntactic notions are correlated, at least in part. This assumption is under debate (Briscoe and Copestake, 1995; Levin, 1993; Dorr and Jones, 1996; Dorr, 1997), but we adopt it here without further discussion.

to semantic classes, hence they can be more easily captured by corpus-based techniques. Second, using verb diatheses reduces noise. There is a certain consensus (Briscoe and Copestake, 1995; Pustejovsky, 1995; Palmer, 1999) that verb diatheses are regular sense extensions. Hence focussing on this type of classification allows one to abstract from the problem of word sense disambiguation and treat residual differences in word senses as noise in the classification task.

We present an in-depth case study, in which we apply machine learning techniques to automatically classify a set of verbs based on distributions of grammatical indicators of diatheses, extracted from a very large corpus. We look at three very interesting classes of verbs: unergatives, unaccusatives, and object-drop verbs (Levin, 1993). These are interesting classes because they all participate in the transitivity alternation, and they are minimal pairs – that is, a small number of well-defined distinctions differentiate their transitive/intransitive behavior. Thus, we expect the differences in their distributions to be small, entailing a fine-grained discrimination task that provides a challenging testbed for automatic classification.

The specific theoretical question we investigate is whether the factors underlying the verb class distinctions are reflected in the statistical distributions of lexical features related to diatheses presented by the individual verbs in the corpus. In doing this, we address the questions above by determining what are the lexical features that could distinguish the behavior of the classes of verbs with respect to the relevant diatheses, which of those features can be gleaned from the corpus, and which of those, once the statistical distributions are available, can be used successfully by an automatic classifier.

In initial work (Stevenson and Merlo, 1999), we found that linguistically motivated features that distinguish the verb classes can be extracted from an annotated, and in one case parsed, corpus. These features are sufficient to almost halve the error rate compared to chance (45% reduction) in automatic verb classification, suggesting that distributional data provides knowledge useful to the classification of verbs. The focus of our original study was the demonstration in principle of learning verb classes from frequency distributions of syntactic features, and an analysis of the relative contribution of the various features to learning. This paper turns to the important next steps of replicating our findings using other training methods and learning algorithms, and analyzing the performance on each of the three classes of verbs. This more detailed analysis of accuracy within each class in turn leads to the development of a new distributional feature intended to improve discriminability among two of the classes. The addition of the new feature successfully reduces the error rate of our initial results in classification by 19%, for a 56% overall reduction in error rate compared to chance.

2 Determining the Features

In this section, we present motivation for the initial features that we investigated in terms of their role in learning the verb classes. We first present the linguistically derived features, then turn to evidence from experimental psycholinguistics to extend the set of potentially relevant features.

2.1 Features of the Verb Classes

The three verb classes under investigation – unergatives, unaccusatives, and object-drop – differ in the properties of their transitive/intransitive alternations, which are exemplified below.

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 washed the hall.
(3b) The boy washed.

The sentences in (1) use an unergative verb, *raced*. Unergatives are intransitive action verbs whose transitive form is the causative counterpart of the intransitive form. Thus, the subject of the intransitive (1a) becomes the object of the transitive (1b) (Brousseau and Ritter, 1991; Hale and Keyser, 1993; Levin and Rappaport Hovav, 1995). The sentences in (2) use an unaccusative verb, *melted*. Unaccusatives are intransitive change of state verbs (2a); like unergatives, the transitive counterpart for these verbs is also causative (2b). The sentences in (3) use an object-drop verb, *washed*; these verbs have a non-causative transitive/intransitive alternation, in which the object is simply optional.

Both unergatives and unaccusatives have a causative transitive form, but differ in the semantic roles that they assign to the participants in the event described. In an intransitive unergative, the subject is an Agent (the doer of the event), and in an intransitive unaccusative, the subject is a Theme (something affected by the event). The role assignments to the corresponding semantic arguments of the transitive forms—i.e., the direct objects—are the same, with the addition of a Causal Agent (the causer of the event) as subject in both cases. Object-drop verbs simply assign Agent to the subject and Theme to the optional object.

We expect the differing semantic role assignments of the verb classes to be reflected in their syntactic behavior, and consequently in the distributional data we collect from a corpus. The three classes can be characterized by their occurrence in two alternations: the transitive/intransitive alternation and the causative alternation. Unergatives are distinguished from the other classes in being rare in the transitive form (see (Stevenson and Merlo, 1997) for an explanation of this fact). Both unergatives and unaccusatives are distinguished from object-drop in being causative in their transitive form, and similarly we expect this to be reflected in amount of detectable causative use. Furthermore, since the causative is a transitive use, and the transitive use of unergatives is expected to be rare, causativity should primarily distinguish unaccusatives from object-drops. In conclusion, we expect the defining features of the verb classes—the intransitive/transitive and causative alternations—to lead to distributional differences in the observed usages of the verbs in these alternations.

2.2 Psycholinguistically Relevant Features

The verbs under study not only differ in their thematic properties, they also differ in their processing properties. Because these verbs can occur both in a transitive and an intransitive form, they have been particularly studied in the context of the main verb/reduced relative (MV/RR) ambiguity illustrated below (Bever, 1970):

The horse raced past the barn fell.

The verb *raced* can be interpreted as either a past tense main verb, or as a past participle within a reduced relative clause (i.e., *the horse* [that was] *raced past the barn*). Because *fell* is the main verb, the reduced relative interpretation of *raced* is required for a coherent analysis of the complete sentence. But the main verb interpretation of *raced* is so strongly preferred that people experience great difficulty at the verb *fell*, unable to integrate it with the interpretation that has been developed to that point.

However, the reduced relative interpretation is not difficult for all verbs, as in the following example:

The boy washed in the tub was angry.

The difference in ease of interpreting the resolutions of this ambiguity has been shown to be sensitive to both frequency differentials (MacDonald, 1994; Trueswell, 1996) and to verb class distinctions (Stevenson and Merlo, 1997; Filip et al., 1999).

Consider the features that distinguish the two resolutions of the MV/RR ambiguity:

MV: The horse raced past the barn quickly.

RR: The horse raced past the barn fell.

In the main verb resolution, the ambiguous verb *raced* is used in its intransitive form, while in the reduced relative, it is used in its transitive, causative form. These features correspond directly to the defining alternations of the three verb classes under study (intransitive/transitive, causative). Additionally, we see that other related features to these usages serve to distinguish the two resolutions of the ambiguity. The main verb form is active and a main verb part-of-speech (labeled as VBD by automatic POS taggers); by contrast, the reduced relative form is passive and a past participle (tagged as VBN). Since these features (active/passive and VBD/VBN) are related to the intransitive/transitive alternation, we expect them to also exhibit distributional differences among the verb classes. Specifically, we expect the unergatives to yield a higher proportion of active and VBD usage, since, as noted above, the transitive use of unergatives is rare.

3 Frequency Distributions of the Features

We assume that currently available large corpora are a reasonable approximation to language (Pullum, 1996). Using a combined corpus of 65-million words, we measured the relative frequency distributions of the four linguistic features (VBD/VBN, active/passive, intransitive/transitive, causative/noncausative) over a sample of verbs from the three lexical semantic classes.

3.1 Materials

We chose a set of 20 verbs from each class based primarily on the classification of verbs in (Levin, 1993) (see Appendix A). The unergatives are manner of motion verbs. The unaccusatives are verbs of change of state. The object-drop verbs are unspecified object alternation verbs. The verbs were selected from Levin's classes based on their absolute frequency. Furthermore, they do not generally show massive departures from the intended verb sense in the corpus. (Though note that there are only 19 unaccusatives because *ripped*, which was initially counted in the unaccusatives, was then excluded from the analysis as it occurred mostly in a different usage in the corpus, as a verb plus particle.) Most of the verbs can occur in the transitive and in the passive. Each verb presents the same form in the simple past and in the past participle. In order to simplify the counting procedure, we made the assumption that counts on this single verb form would approximate the distribution of the features across all forms of the verb.

Most counts were performed on the tagged version of the Brown Corpus and on the portion of the Wall Street Journal distributed by the ACL/DCI (years 1987, 1988, 1989), a combined corpus in excess of 65 million words, with the exception of causativity which was counted only for the 1988 year of the WSJ, a corpus of 29 million words.

3.2 Method

We counted the occurrences of each verb token in a transitive or intransitive use (INTR), in an active or passive use (ACT), in a past participle or simple past use (VBD), and in a causative or non-causative use (CAUS). More precisely, features were counted as follows:

INTR: a verb occurrence was counted as transitive if immediately followed by a nominal group, else it was counted as intransitive.

ACT: main verbs (tagged VBD) were counted as active; participles (tagged VBN) counted as active if the closest preceding auxiliary was *have*, as passive if the closest preceding auxiliary was *be*.

VBD: occurrences tagged VBD were simple past, VBN were past participle.

(Each of the above three counts was normalized over all occurrences of the verb, yielding a single relative frequency measure for each verb for that feature.)

CAUS: The causative feature was approximated by the following steps. First, for each verb, all cooccurring subjects and objects were extracted from a parsed corpus (Collins, 1997). Then the proportion of overlap between the two multisets of nouns was calculated, meant to capture the causative alternation, where the subject of the intransitive can occur as the object of the transitive. We define overlap as the largest multiset of elements belonging to both the subjects and the object multisets, e.g. $\{a, a, a, b\} \cap \{a\} = \{a, a, a\}$. The proportion is the ratio between the overlap and the sum of the subject and object multisets. (For example, for the simple sets above, the ratio would be 3/5 or .60.)

All raw and normalized corpus data are available from the authors, and more detail concerning data collection can be found in (Stevenson and Merlo, 1999).

4 Experiments in Verb Classification

The frequency distributions of the verb alternation features yield a vector for each verb that represents the relative frequency values for the verb on each dimension; the set of 59 vectors constitute the data for our machine learning experiments.

Template: [verb, VBD, ACT, INTR, CAUS, class]

Example: [opened, .79, .91, .31, .16, unacc]

Our goal was to determine whether automatic classification techniques could determine the class of a verb from the distributional properties represented in this vector.

In related work (Stevenson and Merlo, 1999), we describe initial unsupervised and supervised learning experiments on this data, and discuss the contribution of the four different features (the frequency distributions) to accuracy in verb classification. In this paper, we extend the work in several ways. First, we report further analysis of replications of our initial supervised learning results. Next, we demonstrate similar performance using different training methods and learning algorithms, indicating that the performance is independent of the particular learning approach. Furthermore, these additional experiments allow us to evaluate the performance separately on each of the three verb classes. Finally, based on this evaluation, we suggest a new feature to better distinguish the thematic properties of the classes, and present experimental results showing that its use improves our original accuracy rate.

4.1 Initial Experiments

Initial experiments were carried out using a decision tree induction algorithm, the C5.0 system available from http://www.rulequest.com/ (Quinlan, 1992), to automatically create a classification program from a training set of verb vectors with known classification.[2] In our earlier experiments, we ran 10-fold cross-validations repeated 10 times; here we repeat the cross-validations 50 times, and the numbers reported are averages over all the runs.[3]

Table 1 shows the results of our experiments on the four features we counted in the corpora (VBD, ACT, INTR, CAUS), as well as all three-feature subsets of those four. The baseline (chance) performance in this task is 33.8%, since there are 59 vectors and

[2] The system generates both decision trees and rule sets for use in classification. Since the difference in performance between the two is never significant, we report here only the results using the extracted rules. The rules provide a confidence level for each classification, which is unavailable with the decision tree data structure.

[3] A 10-fold cross-validation means that the system randomly divides the data into 10 parts, and runs 10 times on a different 90%-training-data/10%-test-data split, yielding an average accuracy and standard error. This procedure is then repeated for 50 different random divisions of the data, and accuracy and standard error are again averaged across the 50 runs.

Features	Acc%	SE%
VBD ACT INTR CAUS	63.7	0.6
VBD INTR CAUS	62.7	0.6
ACT INTR CAUS	59.9	0.5
VBD ACT CAUS	56.8	0.5
VBD ACT INTR	54.5	0.5

Table 1: Percentage Accuracy (Acc%) and Standard Error (SE%) of C5.0 (33.8% baseline).

3 possible classes. (That is, assigning one of the two most common classes—of 20 verbs each—to all cases would yield 20 out of 59 correct, or 33.8%.) As seen in the table, classification based on the four features performs at 63.7%, or 30% over chance. The true mean of the sample cross-validations lies within plus or minus two standard errors of the reported mean (df=49, t=2.01, p<.05). In all cases, the range is plus or minus 1.0 or 1.2, yielding a very narrow predicted accuracy range. Furthermore, we performed t-tests comparing the results of the 50 cross-validations for each of the different feature subsets. All pairs were significantly different (p<.05) except for the results using all four features (first row in the table) and those excluding ACT (second row in the table). We conclude that all features except ACT contribute positively to classification performance, and that ACT does not degrade performance. In our replications, then, we focus on all four features.

4.2 Replication with Different Training and Learning Methods

There are conceptual and practical reasons for investigating the performance of other training approaches and learning algorithms applied to our verb distribution data. Conceptually, it is desirable to know whether a particular learning algorithm or training technique affects the level of performance. Practically, different methods enable us to evaluate more easily the performance of the classification method within each verb class. (When we run repeated cross-validations with the C5.0 system, we don't have access to the accuracy rate for each class; the system only outputs an overall mean error rate.) To preview, we find that the different training and learning methods we tried all gave similar performance to our original results, and in addition allowed us to evaluate the accuracy within each verb class.

In one set of experiments, we used the same C5.0 system, but employed a training and testing methodology that used a single hold-out case. We held out a single verb vector, trained on the remaining 58 cases, then tested the resulting classifier on the single hold-out case, and recorded the correct and assigned classes for that verb. This was then repeated for each of the 59 verbs. This approach yields both an overall accuracy rate (when the results are averaged across all 59 trials), as well as providing the data necessary for determining accuracy for each verb class (because we have the classification of each verb when it is the test case). The results are presented in Table 2. The overall accuracy is a little less than that achieved with the 10-fold cross-validation methodology (61.0% versus 63.7%). However, we can see clearly now that the unergative verbs are classified with much greater accuracy (75%), while the unaccusative and object-drop verbs are classified with much lower accuracy (57.9% and 50% respectively). The distributional features we have appear to be much better at distinguishing unergatives than unaccusative or object-drop verbs.

Classes	Percent Accuracy
All Classes	61.0
Unergative	75.0
Unaccusative	57.9
ObjectDrop	50.0

Table 2: Percentage Accuracy of C5.0 With Single Hold-Out Training.

To test this directly under our original training assumptions, we ran two different experiments, using 10-fold cross-validation repeated 10 times. The first experiment tested the ability of the classifier to distinguish between unergatives and the other two verb types, without having to distinguish between the latter two. The data included the 20 unergative verbs and a random sample of 10 unaccusative and 10 object-drop verbs; 10 different random samples were selected to form 10 such data sets. In these data sets, the verbs were labeled as unergative or "other". The baseline (chance) classification accuracy for this data is 50%; the mean accuracy achieved across all data sets was 78.5% (standard error 0.8%), a sizable improvement over chance. The second experiment was intended to determine how well the classifier can distinguish unaccusative from object-drop verbs. The data consisted of one set that included all the unaccusative and object-drop verbs, with no unergatives. Because there are only 19 unaccusative verbs, the baseline accuracy rate is 51% (20/39); here the classifier achieved an accuracy only slightly above chance, at 58.3% (standard error 1.8%). These results, summarized in Table 3, clearly confirm the higher accuracy of classifying unergative verbs with the current feature set.

This pattern of results was repeated under a very

Classes	Acc%	SE%
Unergative vs. Other	78.5	0.8
Unaccusative vs. ObjectDrop	58.3	1.8

Table 3: Percentage Accuracy (Acc%) and Standard Error (SE%) of C5.0 (50-51% baseline).

Classes	PCA%	FMP%
All Classes	65.0	63.9
Unergative	85.0	71.7
Unaccusative	60.0	55.0
ObjectDrop	50.0	65.0

Table 4: Percentage Accuracy of PCA (PCA%) and Feature Map (FMP%) Neural Networks.

different type of learning algorithm as well. We performed a set of neural network experiments, using NeuroSolutions 3.0 (see http://www.nd.com), and report here on the networks that achieve the best performance on our data. These are principal components analysis and automatic feature map networks, which are essentially feed-forward perceptrons with pre-processing units that transform the existing features into a more useful format. In our tests, both methods performed best overall when there were no hidden layer units, and the networks were trained for 1000 epochs. The mean accuracy rates of 10-fold cross-validations with these parameter settings are summarized in Table 4. Again, the overall percentage accuracy is in the low sixties, with better performance on the unergatives than on the other two verb classes; the difference was particularly striking with the PCA networks. This overall pattern doesn't change with further training; in fact, training up to 10,000 epochs resulted in very low accuracy (of 45%) for either unaccusatives, object-drops, or both.

To summarize, following a different training approach with C5.0 (the single hold-out method), and applying very different learning approaches (two kinds of neural networks), resulted in similar overall performance to our original C5.0 results. This indicates that the accuracy achieved is at least somewhat independent of specific learning or training techniques. Moreover, these different methods, along with experiments directly testing unergative versus unaccusative/object-drop classification, allow us to examine more closely where the resulting classifiers have the most serious problems. In all cases, the accuracy is best for unergatives, and the accuracy of unaccusatives, object-drops, or both, is degraded. If this performance is indeed a reliable indication of the inherent discriminability of the distributional data, then we must examine more closely the properties of the data itself to understand (and potentially improve) the performance.

Classes	VBD	ACT	INTR	CAUS
Unerg vs. Unacc	***	***	*	***
Unerg vs. ObjDrop	***	***	***	*
Unacc vs. ObjDrop	ns	ns	**	*

*** $p \leq .001$
** $p < .01$
* $p \leq .05$
ns non-significant

Table 5: Significance Levels of T-Tests Comparing Feature Values Between Verb Classes.

4.3 Discriminating Unaccusative and Object-Drop Verbs

To understand why the data discriminates unergatives reasonably well, but not unaccusatives and object-drops, we need to directly test the discriminability of the features across the classes. We do so by using t-tests to compare the values of the different features—VBD, ACT, INTR, CAUS—for unergative and unaccusative verbs, unergative and object-drop verbs, and unaccusative and object-drop verbs. In each case, the t-test is giving the likelihood that the two sets of values—e.g., the VBD feature values for unergatives and for unaccusatives—are drawn from different populations. Table 5 shows that all sets of features are significantly different for unergative and unaccusative verbs, and for unergative and object-drop verbs. However, only INTR and CAUS are significantly different for unaccusative and object-drop verbs, indicating that we need additional features that have different values across these two classes.

In Section 2.1, we noted the differing semantic role assignments for the verb classes, and hypothesized that these differences would affect the expression of syntactic features that are countable in a corpus. For example, the CAUS feature approximates semantic role information by encoding the overlap between nouns that can occur in the subject and object positions of a causative verb. Here we suggest another feature, that of animacy of subject, that is intended to distinguish nouns that receive an Agent role from those that receive a Theme role. Recall that object-drop verbs assign Agent to their subject in both the transitive and intransitive alternations, while unaccusatives assign Agent to their subject only in the transitive, and Theme in the intransitive. We expect then that object-drop verbs will occur more often with an animate subject. Note again that we are

Features	Acc%	SE%
VBD ACT INTR CAUS	63.7	0.6
VBD ACT INTR CAUS PRO	70.7	0.4

Table 6: Percentage Accuracy (Acc%) and Standard Error (SE%) of C5.0, With and Without New PRO Feature, All Verb Classes (33.8% baseline).

making use of frequency distributions—the claim is not that only Agents can be animate, but rather that nouns that receive the Agent role will more often be animate than nouns that receive the Theme role.

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 instead count pronouns (other than *it*) in subject position. The assumption is that the words *I*, *we*, *you*, *she*, *he*, and *they* most often refer to animate entities. The values for the new feature, PRO, were determined by automatically extracting all subject/verb tuples including our 59 examples verbs (from the WSJ88 parsed corpus), and computing the ratio of occurrences of pronouns to all subjects.

We again apply t-tests to our new data to determine whether the sets of PRO values differ across the verb classes. Interestingly, we find that the PRO values for unaccusative verbs (the only class to assign Theme role to the subject in one of its alternations) are significantly different from those for both unergative and object-drop verbs ($p<.05$). Moreover, the PRO values for unergative and object-drop verbs (whose subjects are Agents in both alternations) are not significantly different. This pattern confirms the ability of the feature to capture the thematic distinction between unaccusative verbs and the other two classes.

Table 6 shows the result of applying C5.0 (10-fold cross-validation repeated 50 times) to the three-way classification task using the PRO feature in conjunction with the four previous features. Accuracy improves to over 70%, a reduction in the error rate of almost 20% due to this single new feature. Moreover, classifying the unaccusative and object-drop verbs using the new feature in conjunction with the previous four leads to accuracy of over 68% (compared to 58% without PRO). We conclude that this feature is important in distinguishing unaccusative and object-drop verbs, and likely contributes to the improvement in the three-way classification because of this. Future work will examine the performance within the verb classes of this new set of features, to see whether accuracy has also improved for unergative verbs.

5 Conclusions

In this paper, we have presented an in-depth case study, in which we investigate various machine learning techniques to automatically classify a set of verbs, based on distributional features extracted from a very large corpus. Results show that a small number of linguistically motivated grammatical features are sufficient to reduce the error rate by more than 50% over chance, achieving a 70% accuracy rate in a three-way classification task. This leads us to conclude that corpus data is a usable repository of verb class information. On one hand, we observe that semantic properties of verb classes (such as causativity, or animacy of subject) may be usefully approximated through countable syntactic features. Even with some noise, lexical properties are reflected in the corpus robustly enough to positively contribute in classification. On the other hand, however, we remark that deep linguistic analysis cannot be eliminated—in our approach, it is embedded in the selection of the features to count. We also think that using linguistically motivated features makes the approach very effective and easily scalable: we report a 56% reduction in error rate, with only five features that are relatively straightforward to count.

Acknowledgements

This research was partly sponsored by the Swiss National Science Foundation, under fellowship 8210-46569 to Paola Merlo; by the US National Science Foundation, under grants #9702331 and #9818322 to Suzanne Stevenson; and by the Information Sciences Council of Rutgers University. We thank Martha Palmer for getting us started on this work, and Michael Collins for giving us access to the output of his parser. We gratefully acknowledge the help of Kiva Dickinson, who calculated normalizations of the corpus data.

Appendix A

The unergatives are manner of motion verbs: *jumped*, *rushed*, *marched*, *leaped*, *floated*, *raced*, *hurried*, *wandered*, *vaulted*, *paraded*, *galloped*, *glided*, *hiked*, *hopped*, *jogged*, *scooted*, *scurried*, *skipped*, *tiptoed*, *trotted*.

The unaccusatives are verbs of change of state: *opened*, *exploded*, *flooded*, *dissolved*, *cracked*, *hardened*, *boiled*, *melted*, *fractured*, *solidified*, *collapsed*, *cooled*, *folded*, *widened*, *changed*, *cleared*, *divided*, *simmered*, *stabilized*.

The object-drop verbs are unspecified object alternation verbs: *played*, *painted*, *kicked*, *carved*, *reaped*, *washed*, *danced*, *yelled*, *typed*, *knitted*, *borrowed*, *inher-*

ited, *organized*, *rented*, *sketched*, *cleaned*, *packed*, *studied*, *swallowed*, *called*.

References

Thomas G. Bever. 1970. The cognitive basis for linguistic structure. In J. R. Hayes, editor, *Cognition and the Development of Language*. John Wiley, New York.

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

Edward Briscoe and Ann Copestake. 1995. Lexical rules in the TDFS framework. Technical report, Acquilex-II Working Papers.

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

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

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

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

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

Hana Filip, Michael Tanenhaus, Greg Carlson, Paul Allopenna, and Joshua Blatt. 1999. Reduced relatives judged hard require constraint-based analyses. In P. Merlo and S. Stevenson, editors, *Sentence Processing and the Lexicon: Formal, Computational, and Experimental Perspectives*, John Benjamins, Holland.

Ken Hale 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*, pages 53–110. MIT Press.

Judith L. Klavans and Martin Chodorow. 1992. Degrees of stativity: The lexical representation of verb aspect. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*.

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

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

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

Maryellen C. MacDonald. 1994. Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes*, 9(2):157–201.

Paola Merlo 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, CA.

George Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. 1990. Five papers on Wordnet. Technical report, Cognitive Science Lab, Princeton University.

Martha Palmer. 1999. Consistent criteria for sense distinctions. *Computing for the Humanities*.

Fernando Pereira, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of english words. In *Proc. of the 31th Annual Meeting of the ACL*, pages 183–190.

Fernando Pereira, Ido Dagan, and Lillian Lee. 1997. Similarity-based methods for word sense disambiguation. In *Proc. of the 35th Annual Meeting of the ACL and the 8th Conf. of the EACL (ACL/EACL'97)*, pages 56 –63.

Geoffrey K. Pullum. 1996. Learnability, hyperlearning, and the poverty of the stimulus. In Jan Johnson, Matthew L. Juge, and Jeri L. Moxley, editors, *22nd Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on the Role of Learnability in Grammatical Theory*, pages 498–513, Berkeley, California. Berkeley Linguistics Society.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press.

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

Philip Resnik. 1992. Wordnet and distributional analysis: a class-based approach to lexical discovery. In *AAAI Workshop in Statistically-based NLP Techniques*, pages 56–64.

Doug Roland and Dan Jurafsky. 1998. How verb subcategorization frequencies are affected by corpus choice. In *Proc. of the 36th Annual Meeting of the ACL*, Montreal, CA.

Suzanne Stevenson and Paola Merlo. 1997. Lexical structure and parsing complexity. *Language and Cognitive Processes*, 12(2/3):349–399.

Suzanne Stevenson and Paola Merlo. 1999. Automatic verb classification using distributions of grammatical features. In *Proc. of the 9th Conference of the European Chapter of the ACL*, Bergen, Norway, pages 45–52.

John Trueswell. 1996. The role of lexical frequency in syntactic ambiguity resolution. *J. of Memory and Language*, 35:566–585.