Proceedings of the 9th Conference of the European ACL (EACL 99), 45-52.

Automatic Verb Classification Using Distributions of Grammatical Features

Suzanne Stevenson
Dept of Computer Science
and Center for Cognitive Science (RuCCS)
Rutgers University
CoRE Building, Busch Campus
New Brunswick, NJ 08903
U.S.A.
suzanne@ruccs.rutgers.edu

Paola Merlo
LATL-Department of Linguistics
University of Geneva
2 rue de Candolle
1211 Genève 4
SWITZERLAND
merlo@lettres.unige.ch

Abstract

We apply machine learning techniques to classify automatically a set of verbs into lexical semantic classes, based on distributional approximations of diatheses, extracted from a very large annotated corpus. Distributions of four grammatical features are sufficient to reduce error rate by 50% 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). 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 automatically classify verbs by using statistical approximations to verb diatheses, to train an automatic classifier. We use verb diatheses, following Levin and Dorr, for two reasons. First, verb diatheses are syntactic cues to semantic classes,

[1] We are aware that a distributional approach rests on one strong assumption on the nature of the representations under study: semantic notions and syntactic notions are correlated, at least in part. This assumption is not uncontroversial (Briscoe and Copestake, 1995; Levin, 1993; Dorr and Jones, 1996; Dorr, 1997). We adopt it here as a working hypothesis without further discussion.

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.

We follow a computational experimental methodology by investigating as indicated each of the hypotheses below:

H1: Linguistically and psychologically motivated features for distinguishing the verb classes are apparent within linguistic experience.

> We analyze the three classes based on properties of the verbs that have been shown to be relevant for linguistic classification (Levin 93), or for disambiguation in syntactic processing (MacDonald94, Trueswell96) to determine potentially relevant distinctive features. We then count those features (or approximations to them) in a very large corpus.

H2: The distributional patterns of (some of) those features contribute to learning the classifications of the verbs.

> We apply machine learning techniques to determine whether the features support the learning of the classifications.

H3: Non-overlapping features are the most effective in learning the classifications of the verbs.

> We analyze the contribution of different features to the classification process.

To preview, we find that, related to (H1), linguistically motivated features (related to diatheses) that distinguish the verb classes can be extracted from an annotated, and in one case parsed, corpus. In relation to (H2), a subset of these features is sufficient to halve the error rate compared to chance in automatic verb classification, suggesting that distributional data provides useful knowledge to the classification of verbs. Furthermore, in relation to (H3) we find that features that are distributionally predictable, because they are highly correlated to other features, contribute little to classification performance. We conclude that the usefulness of distributional features to the learner is determined by their informativeness.

2 Determining the Features

In this section, we present motivation for the features that we investigate 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.

<u>Unergative</u>:
(1a) The horse raced past the barn.
(1b) The jockey raced the horse past the barn.

<u>Unaccusative</u>:
(2a) The butter melted in the pan.
(2b) The cook melted the butter in the pan.

<u>Object-drop</u>:
(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 Features of the MV/RR Alternatives

Not only do the verbs under study 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 (?).

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

<u>Main Verb</u>: The horse raced past the barn quickly.

<u>Reduced Relative</u>: 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). Although these properties are redundant with the intransitive/transitive distinction, recent work in machine learning (Ratnaparkhi, 1997; Ratnaparkhi, 1998) has shown that using overlapping features can be beneficial for learning in a maximum entropy framework, and we want to explore it in this setting to test **H3** above.[2] In the next section,

[2] These properties are redundant with the intransitive/transitive distinction, as passive implies transitive use, and necessarily entails the use of a past participle. We performed a correlation analysis that

we describe how we compile the corpus counts for each of the four properties, in order to approximate the distributional information of these alternations.

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 linguistic features (VBD/VBN, active/passive, intransitive/transitive, causative/non-causative) over a sample of verbs from the three lexical semantic classes.

3.1 Materials

We chose a set of 20 verbs from each class – divided into two groups each, as will be explained below – based primarily on the classification of verbs in (Levin, 1993).

The unergatives are manner of motion verbs: *jumped*, *rushed*, *marched*, *leaped*, *floated*, *raced*, *hurried*, *wandered*, *vaulted*, *paraded* (group 1); *galloped*, *glided*, *hiked*, *hopped*, *jogged*, *scooted*, *scurried*, *skipped*, *tiptoed*, *trotted* (group 2).

The unaccusatives are verbs of 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).

The object-drop verbs are unspecified object alternation verbs: *played*, *painted*, *kicked*, *carved*, *reaped*, *washed*, *danced*, *yelled*, *typed*, *knitted* (group 1); *borrowed*, *inherited*, *organised*, *rented*, *sketched*, *cleaned*, *packed*, *studied*, *swallowed*, *called* (group 2).

The verbs were selected from Levin's classes on the basis of our intuitive judgment that they are likely to be used with sufficient frequency to be found in the corpus we had available. 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 group 2 of unaccusatives, was then excluded from the analysis as it occurred mostly in a different usage in the corpus; ie, as a verb plus particle.)

yielded highly significant R=.44 between intransitive and active use, and R=.36 between intransitive and main verb (VBD) use. We discuss the effects of feature overlap in the experimental section.

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, entailing that we can extract both active and passive occurrences by searching on a single token. 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).[3] More precisely, the following occurrences were counted in the corpus.

INTR: the closest nominal group following the verb token was considered to be a potential object of the verb. A verb occurrence immmediately followed by a potential object was counted as transitive. If no object followed, the occurrence was counted as intransitive.

ACT: main verb (ie, those tagged VBD) were counted as active. Tokens with tag VBN were also counted as active if the closest preceding auxiliary was *have*, while they were counted as passive if the closest preceding auxiliary was *be*.

VBD: A part-of-speech tagged corpus was used, hence the counts for VBD/VBN were simply done based on the POS label according to the tagged corpus.

CAUS: The causative feature was approximated by the following steps. First, for each verb occurrence subjects and objects were extracted from a parsed corpus (Collins 1997). Then the propor-

[3] In performing this kind of corpus analysis, one has to take into account the fact that current corpus annotations do not distinguish verb senses. However, 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 kind of extension of meaning does not modify subcategorization distributions (Roland and Jurafsky, 1998), although it might modify the rate of causativity, but this is an unavoidable limitation at the current state of annotation of corpora.

tion of overlap between the two multisets of nouns was calculated, meant to capture the property of the causative construction that 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.

The verbs in group 1 had been used in an earlier study, in which it was important to minimize noisy data, 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 transitivity and voice, the method of count depended on the group. For group 1, the counts were done automatically by regular expression patterns, and then corrected, partly by hand and partly automatically. For group 2, the counts were done automatically without any manual intervention. For causativity, the same counting scripts were used for both groups of verbs, but the input to the counting programs was determined by manual inspection of the corpus for verbs belonging to group 1, while it was extracted automatically from a parsed corpus for group 2 (WSJ 1988, parsed with the parser from (Collins, 1997).

Each count was normalized over all occurrences of the verb, yielding a total of four relative frequency features: VBD (%VBD tag), ACT (%active use), INTR (%intransitive use), CAUS (%causative use).[4]

4 Experiments in Clustering and Classification

Our goal was to determine whether statistical indicators can be automatically combined to determine the class of a verb from its distributional properties. We experimented both with self-aggregating and supervised methods. 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.

Vector template: [verb, VBD, ACT, INTR, CAUS]

Example: [opened, .793, .910, .308, .158]

[4]All raw and normalized corpus data are available from the authors.

Table 1: Accuracy of the Verb Clustering Task.

Features	Accuracy
1. VBD ACT INTR CAUS	52%
2. VBD ACT CAUS	54%
3. VBD ACT INTR	45%
4. ACT INTR CAUS	47%
5. VBD INTR CAUS	66%

We must now determine which of the distributions actually contribute to learning the verb classifications. First we describe computational experiments in unsupervised learning, using hierarchical clustering, then we turn to supervised classification.

4.1 Unsupervised Learning

Other work in automatic lexical semantic classification has taken an approach in which clustering over statistical features is used in the automatic formation of classes (Pereira et al., 1993; Pereira et al., 1997; Resnik, 1992). We used the hierarchical clustering algorithm available in SPlus5.0, imposing a cut point that produced three clusters, to correspond to the three verb classes. Table 1 shows the accuracy achieved using the four features described above (row 1), and all three-feature subsets of those four features (rows 2–5). Note that chance performance in this task (a three-way classification) is 33% correct.

The highest accuracy in clustering, of 66%—or half the error rate compared to chance—is obtained only by the triple of features in row 5 in the table: VBD, INTR, and CAUS. All other subsets of features yield a much lower accuracy, of 45-54%. We can conclude that some of the features contribute useful information to guide clustering, but the inclusion of ACT actually degrades performance. Clearly, having fewer but more relevant features is important to accuracy in verb classification. We will return to the issue in detail of which features contribute most to learning in our discussion of supervised learning below.

A problem with analyzing the clustering performance is that it is not always clear what counts as a misclassification. We cannot actually know what the identity of the verb class is for each cluster. In the above results, we imposed a classification based on the class of the majority of verbs in a cluster, but often there was a tie between classes within a cluster, and/or the same class was the majority class in more than one cluster. To evaluate better the effects of the features in learning, we therefore turned to a supervised learning method,

Table 2: Accuracy of the Verb Classification Task.

Features	Decision Trees Accuracy	Decision Trees Standard Error	Rule Sets Accuracy	Rule Sets Standard Error
1. VBD ACT INTR CAUS	64.2%	1.7%	64.9%	1.6%
2. VBD ACT CAUS	55.4%	1.5%	55.7%	1.4%
3. VBD ACT INTR	54.4%	1.4%	56.7%	1.5%
4. ACT INTR CAUS	59.8%	1.2%	58.9%	0.9%
5. VBD INTR CAUS	60.9%	1.2%	62.3%	1.2%

where the classification of each verb in a test set is unambiguous.

4.2 Supervised learning

For our supervised learning experiments, we used the publicly available version of the C5.0 machine learning algorithm,[5] a newer version of C4.5 (Quinlan, 1992), which generates decision trees from a set of known classifications. We also had the system extract rule sets automatically from the decision trees. For all reported experiments, we ran a 10-fold cross-validation repeated ten times, and the numbers reported are averages over all the runs.[6]

Table 2 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. As seen in the table, classification based on the four features performs at 64–65%, or 31% over chance. (Recall that this is a 3-way decision, hence baseline is 33%).

Given the resources needed to extract the features from the corpus and to annotate the corpus itself, we need to understand the relative contribution of each feature to the results – one or more of the features may make little or no contribution to the successful classification behavior. Observe that when either the INTR or CAUS feature is removed (rows 2 and 3, respectively, of Table 2), performance degrades considerably, with a decrease in accuracy of 8–10% from the maximum achieved with the four features (row 1). However, when the VBD feature is removed (row 4), there is a smaller decrease in accuracy, of 4–6%. When the ACT feature is removed (row 5), there is an even smaller decrease, of 2–4%. In fact, the accuracy here is very close to the accuracy of the four-feature results when the standard error is taken into account. We conclude then that INTR and CAUS contribute the most to the accuracy of the classification, while ACT seems to contribute little. (Compare the clustering results, in which the best performance was achieved with the subset of features excluding ACT.) This shows that not all the linguistically relevant features are equally useful in learning.

We think that this pattern of results is related to the combination of the feature distributions: some distributions are highly correlated, while others are not. According to our calculations, CAUS is not significantly correlated with any other feature; of the features that are significantly correlated, VBD is more highly correlated with ACT than with INTR (R=.67 and R=.36 respectively), while INTR is more highly correlated with ACT than with VBD (R=.44 and R=.36 respectively).

We expect combinations of features that are not correlated to yield better classification accuracy. If we compare the accuracy of the 3-feature combinations in Table 2 (rows 2–5), this hypothesis is confirmed. The three combinations that contain the feature CAUS (rows 2, 4 and 5)—the uncorrelated feature—have better performance than the combination that does not (row 3), as expected. Now consider the subsets of three features that include CAUS with a pair of the other correlated features. The combination containing VBD and INTR (row 5)—the least correlated pair of the features VBD, INTR, and ACT—has the best accuracy, while the combination containing the highly correlated VBD and ACT (row 2) has the worst accuracy. The accuracy of the subset {VBD, INTR, CAUS} (row 5) is also better than the accuracy of the subset {ACT, INTR, CAUS} (row 4), because INTR overlaps with VBD less than with ACT.[7]

[5] Available for a number of platforms from `http://www.rulequest.com/`.

[6] A 10-fold cross-validation means that 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. This procedure is then repeated for 10 different random divisions of the data, and accuracy and standard error are again averaged across the ten runs.

[7] We suspect that another factor comes into play, namely how noisy the feature is. The similarity in performance using INTR or CAUS in combination with

5 Conclusions

In this paper, we have presented an in-depth case study, in which we apply 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 halve the error rate over chance. 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) may be usefully approximated through countable 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 50% reduction in error rate, with only 4 features that are relatively straightforward to count.

Acknowledgements

This research was partly sponsored by the Swiss National Science Foundation, under fellowship 8210-46569 to P. Merlo, and by the US National Science Foundation, under grants #9702331 and #9818322 to S. Stevenson. We thank Martha Palmer for getting us started on this work and Michael Collins for giving us acces to the output of his parser.

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.

VBD and ACT (rows 2 and 3) might be due to the fact that the counts for CAUS are a more noisy approximation of the actual feature distribution than the counts for INTR. We reserve defining a precise model of noise, and its interaction with the other features, for future research.

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.

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.

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.

Adwait Ratnaparkhi. 1997. A linear observed time statistical parser based on maximum entropy models. In *2nd Conf. on Empirical Methods in NLP*, pages 1–10, Providence, RI.

Adwait Ratnaparkhi. 1998. Statistical models for unsupervised prepositional phrase attachment. In *Proc. of the 36th Annual Meeting of the ACL*, Montreal, 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.

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