Corpus-Based Linguistic Indicators for Aspectual Classification

Eric V. Siegel
Department of Computer Science
Columbia University
New York, NY 10027

Abstract

Fourteen indicators that measure the frequency of lexico-syntactic phenomena linguistically related to aspectual class are applied to aspectual classification. This group of indicators is shown to improve classification performance for two aspectual distinctions, stativity and completedness (i.e., telicity), over unrestricted sets of verbs from two corpora. Several of these indicators have not previously been discovered to correlate with aspect.

1 Introduction

Aspectual classification maps clauses to a small set of primitive categories in order to reason about time. For example, *events* such as, *"You called your father,"* are distinguished from *states* such as, *"You resemble your father."* These two high-level categories correspond to primitive distinctions in many domains, e.g., the distinction between *procedure* and *diagnosis* in the medical domain.

Aspectual classification further distinguishes events according to *completedness* (i.e., telicity), which determines whether an event reaches a culmination point in time at which a new state is introduced. For example, *"I made a fire"* is culminated, since a new state is introduced – something is made, whereas, *"I gazed at the sunset"* is non-culminated.

Aspectual classification is necessary for interpreting temporal modifiers and assessing temporal entailments (Vendler, 1967; Dowty, 1979; Moens and Steedman, 1988; Dorr, 1992), and is therefore a necessary component for applications that perform certain natural language interpretation, natural language generation, summarization, information retrieval, and machine translation tasks.

Aspect introduces a large-scale, domain-dependent lexical classification problem. Although an aspectual lexicon of verbs would suffice to classify many clauses by their main verb only, a verb's primary class is often domain-dependent (Siegel, 1998b). Therefore, it is necessary to produce a specialized lexicon for each domain.

Most approaches to automatically categorizing words measure co-occurrences between open-class lexical items (Schűtze, 1992; Hatzivassiloglou and McKeown, 1993; Pereira et al., 1993). This approach is limited since co-occurrences between open-class lexical items is sparse, and is not specialized for particular semantic distinctions such as aspect.

In this paper, we describe an expandable framework to classify verbs with linguistically-specialized numerical indicators. Each linguistic indicator measures the frequency of a lexico-syntactic *marker*, e.g. the *perfect* tense. These markers are linguistically related to aspect, so the indicators are specialized for aspectual classification in particular. We perform an evaluation of fourteen linguistic indicators over unrestricted sets of verbs from two corpora. When used in combination, this group of indicators is shown to improve classification performance for two aspectual distinctions: stativity and completedness. Moreover, our analysis reveals a predictive value for several indicators that have not previously been discovered to correlate with aspect in the linguistics literature.

The following section further describes aspect, and introduces linguistic insights that are exploited by linguistic indicators. The next section describes the set of linguistic indicators evaluated in this paper. Then, our experimental method and results are given, followed by a discussion and conclusions.

Table 1: Aspectual classes. This table comes from Moens and Steedman (Moens and Steedman, 1988).

	EVENTS		STATES
	punctual	extended	
Culm	CULM recognize	CULM PROCESS build a house	understand
Non-Culm	POINT hiccup	PROCESS run, swim	

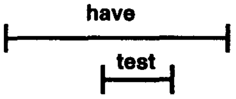
2 Aspect in Natural Language

Table 1 summarizes the three aspectual distinctions, which compose five aspectual categories. In addition to the two distinctions described in the previous section, *atomicity* distinguishes events according to whether they have a time duration (*punctual* versus *extended*). Therefore, four classes of events are derived: *culmination*, *culminated process*, *process*, and *point*. These aspectual distinctions are defined formally by Dowty (1979).

Several researchers have developed models that incorporate aspectual class to assess temporal constraints between clauses (Passonneau, 1988; Dorr, 1992). For example, stativity must be identified to detect temporal constraints between clauses connected with *when*, e.g., in interpreting (1),

(1) She *had* good strength when objectively *tested.*

the following temporal relationship holds:



However, in interpreting (2),

(2) Phototherapy was *discontinued* when the bilirubin *came* down to 13.

the temporal relationship is different:

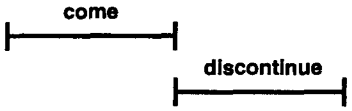


Table 2: Several aspectual markers and associated constraints on aspectual class, primarily from Klavans' summary (1994).

If a clause can occur:	**then it is:**
with a *temporal* adverb (e.g., *then*)	**Event**
in *progressive*	**Extended Event**
with a duration *in*-PP (e.g., *in an hour*)	**Culm Event**
in the *perfect* tense	**Culm Event** or **State**

These aspectual distinctions are motivated by a series of entailment *constraints.* In particular, certain lexico-syntactic features of a clause, such as temporal adjuncts and tense, are constrained by and contribute to the aspectual class of the clause (Vendler, 1967; Dowty, 1979). Tables 2 illustrates an array of linguistic constraints. Each entry in this table describes an aspectual *marker* and the constraints on the aspectual category of any clause that appears with that marker. For example, a clause must be an extended event to appear in the progressive tense, e.g.,

(3) He was prospering in India. (extended),

which contrasts with,

(4) *You were noticing it. (punctual).

and,

(5) *She was seeming sad. (state).

As a second example, an event must be *culminated* to appear in the perfect tense, for example,

(6) She had made an attempt. (culm.),

which contrasts with,

(7) *He has cowered down. (non-culm.)

3 Linguistic Indicators

The best way to exploit aspectual markers is not obvious, since, while the presence of a marker in a particular clause indicates a constraint on the aspectual class of the clause, the absence thereof does not place any constraint. Therefore, as with most statistical methods for natural language, the linguistic constraints associated with markers are best exploited by a system that measures co-occurrence frequencies. For example, a verb that appears more frequently in the progressive is more likely to describe an event. Klavans and Chodorow (1992) pioneered the application of statistical corpus analysis to aspectual classification by ranking verbs according to the frequencies with which they occur with certain aspectual markers.

Table 3 lists the linguistic indicators evaluated for aspectual classification. Each indica-

Ling Indicator	Example Clause
`frequency`	`(not applicable)`
"not" or *"never"*	*She can* **not** *explain why.*
temporal adverb	*I saw to it* **then**.
no subject	*He was admitted.*
past/pres partic	*...blood pressure going up.*
duration *in*-PP	*She built it* **in an hour**.
perfect	*They* **have** *landed.*
present tense	*I am happy.*
progressive	*I am behaving myself.*
manner adverb	*She studied* **diligently**.
evaluation adverb	*They performed* **horribly**.
past tense	*I was happy.*
duration *for*-PP	*I sang* **for ten minutes**.
continuous adverb	*She will live* **indefinitely**.

Table 3: Fourteen linguistic indicators evaluated for aspectual classification.

tor has a unique value for each verb. The first indicator, `frequency`, is simply the frequency with which each verb occurs over the entire corpus. The remaining 13 indicators measure how frequently each verb occurs in a clause with the named linguistic marker. For example, the next three indicators listed measure the frequency with which verbs 1) are modified by *not* or *never*, 2) are modified by a *temporal* adverb such as *then* or *frequently*, and 3) have no deep subject (e.g., passive phrases such as, *"She was admitted to the hospital"*). Further details regarding these indicators and their linguistic motivation is given by Siegel (1998b).

There are several reasons to expect superior classification performance when employing multiple linguistic indicators in combination rather than using them individually. While individual indicators have predictive value, they are predictively incomplete. This incompleteness has been illustrated empirically by showing that some indicators help for only a subset of verbs (Siegel, 1998b). Such incompleteness is due in part to sparsity and noise of data when computing indicator values over a corpus with limited size and some parsing errors. However, this incompleteness is also a consequence of the linguistic characteristics of various indicators. For example:

- Aspectual *coercion* such as *iteration* compromises indicator measurements (Moens and Steedman, 1988). For example, a punctual event appears with the progressive in, *"She was sneezing for a week."* (point → process → culminated process) In this example, *for a week* can only modify an extended event, requiring the first coercion. In addition, this *for*-PP also makes an event culminated, causing the second transformation.
- Some aspectual markers such as the pseudo-cleft and *manner* adverbs test for *intentional* events, and therefore are not compatible with all events, e.g., "**I died* **diligently**."
- The *progressive* indicator's predictiveness for stativity is compromised by the fact that many `location` verbs can appear with the progressive, even in their stative sense, e.g. *"The book was lying on the shelf."* (Dowty, 1979)
- Several indicators measure phenomena that are not linguistically constrained by any aspectual category, e.g., the *present tense*, `frequency` and *not/never* indicators.

4 Method and Results

In this section, we evaluate the set of fourteen linguistic indicators for two aspectual distinctions: stativity and completedness. Evaluation is over corpora of medical reports and novels, respectively. This data is summarized in Table 4 (available at `www.cs.columbia.edu/~evs/VerbData`).

First, linguistic indicators are each evaluated individually. A training set is used to select indicator value thresholds for classification. Then, we report the classification performance achieved by combining multiple indicators. In this case, the training set is used to optimize a model for combining indicators. In both cases, evaluation is performed over a separate test set of clauses.

The combination of indicators is performed by four standard supervised learning algorithms: decision tree induction (Quinlan, 1986), CART (Friedman, 1977), log-linear regression (Santner and Duffy, 1989) and genetic programming (GP) (Cramer, 1985; Koza, 1992).

A pilot study showed no further improvement in accuracy or recall tradeoff by additional learning algorithms: Naive Bayes (Duda and

	stativity	**completedness**
corpus:	3,224 med reports	10 novels
size:	1,159,891	846,913
parsed clauses:	97,973	75,289
training:	739 (634 events)	307 (196 culm)
testing:	739 (619 events)	308 (195 culm)
verbs in test set:	222	204
clauses excluded:	*be* and *have*	stative

Table 4: Two classification problems on different data sets.

Hart, 1973), Ripper (Cohen, 1995), ID3 (Quinlan, 1986), C4.5 (Quinlan, 1993), and metalearning to combine learning methods (Chan and Stolfo, 1993).

4.1 Stativity

Our experiments are performed across a corpus of 3,224 medical discharge summaries. A medical discharge summary describes the symptoms, history, diagnosis, treatment and outcome of a patient's visit to the hospital. These reports were parsed with the English Slot Grammar (ESG) (McCord, 1990), resulting in 97,973 clauses that were parsed fully with no self-diagnostic errors (ESG produced error messages on 12,877 of this corpus' 51,079 complex sentences).

Be and *have*, the two most popular verbs, covering 31.9% of the clauses in this corpus, are handled separately from all other verbs. Clauses with *be* as their main verb, comprising 23.9% of the corpus, always denote a state. Clauses with *have* as their main verb, composing 8.0% of the corpus, are highly ambiguous, and have been addressed separately by considering the direct object of such clauses (Siegel, 1998a).

4.1.1 Manual Marking

1,851 clauses from the parsed corpus were manually marked according to stativity. As a linguistic test for marking, each clause was tested for readability with "*What happened was...*"[1] A comparison between human markers for this test performed over a different corpus is reported below in Section 4.2.1. Of these, 373 clauses were rejected because of parsing problems. This left 1,478 clauses, divided equally into training and testing sets.

Linguistic Indicator	Stative Mean	Event Mean	T-test P-value
frequency	932.89	667.57	0.0000
"not" or *"never"*	4.44%	1.56%	0.0000
temporal adverb	1.00%	2.70%	0.0000
no subject	36.05%	57.56%	0.0000
past/pres partic	20.98%	15.37%	0.0005
duration *in*-PP	0.16%	0.60%	0.0018
perfect	2.27%	3.44%	0.0054
present tense	11.19%	8.94%	0.0901
progressive	1.79%	2.69%	0.0903
manner adverb	0.00%	0.03%	0.1681
evaluation adverb	0.69%	1.19%	0.1766
past tense	62.85%	65.69%	0.2314
duration *for*-PP	0.59%	0.61%	0.8402
continuous adverb	0.04%	0.03%	0.8438

Table 5: Indicators discriminate between states and events.

83.8% of clauses with main verbs other than *be* and *have* are events, which thus provides a baseline method of 83.8% for comparison. Since our approach examines only the main verb of a clause, classification accuracy over the test cases has a maximum of 97.4% due to the presence of verbs with multiple classes.

4.1.2 Individual Indicators

The values of the indicators listed in Table 5 were computed, for each verb, across the 97,973 parsed clauses from our corpus of medical discharge summaries.

The second and third columns of Table 5 show the average value for each indicator over stative and event clauses, as measured over the training examples. For example, 4.44% of stative clauses are modified by either *not* or *never*, but only 1.56% of event clauses were so modified.

The fourth column shows the results of T-tests that compare indicator values over stative training cases to those over event cases for each indicator. As shown, the differences in stative and event means are statistically significant ($p < .01$) for the first seven indicators.

Each indicator was tested individually for classification accuracy by establishing a classification threshold over the training data, and validating performance over the testing data using the same threshold. Only the frequency indicator succeeded in significantly improving clas-

[1] Manual labeling followed a strict set of linguistically-motivated guidelines, e.g., negations were ignored (Siegel, 1998b).

		States		Events	
	acc	recall	prec	recall	prec
dt	93.9%	74.2%	86.4%	97.7%	95.1%
GP	91.2%	47.4%	97.3%	99.7%	90.7%
llr	86.7%	34.2%	68.3%	96.9%	88.4%
bl	83.8%	0.0%	100.0%	100.0%	83.8%
bl2	94.5%	69.2%	95.4%	99.4%	94.3%

Table 6: Comparison of three learning methods and two performance baselines, distinguishing states from events.

sification accuracy by itself, achieving an accuracy of 88.0%. This improvement in accuracy was achieved simply by discriminating the popular verb *show* as a state, but classifying all other verbs as events. Although many domains may primarily use *show* as an event, its appearances in medical discharge summaries, such as, *"His lumbar puncture* **showed** *evidence of white cells,"* primarily utilize *show* to denote a state.

4.1.3 Indicators in Combination

Three machine learning methods successfully combined indicator values, improving classification accuracy over the baseline measure. As shown in Table 6, the decision tree attained the highest accuracy, 93.9%. Binomial tests showed this to be a significant improvement over the 88.0% accuracy achieved by the `frequency` indicator alone, as well as over the other two learning methods. No further improvement in classification performance was achieved by CART.

The increase in the number of stative clauses correctly classified, i.e. stative recall, illustrates an even greater improvement over the baseline. As shown in Table 6, the three learning methods achieved stative recalls of 74.2%, 47.4% and 34.2%, as compared to the 0.0% stative recall achieved by the baseline, while only a small loss in recall over event clauses was suffered. The baseline does not classify any stative clauses correctly because it classifies all clauses as events.

Classification performance is equally competitive without the `frequency` indicator, although this indicator appears to dominate over others. When decision tree induction was employed to combine only the 13 indicators other than `frequency`, the resulting decision tree achieved 92.4% accuracy and 77.5% stative recall.

4.2 Completedness

In medical discharge summaries, non-culminated event clauses are rare. Therefore, our experiments for classification according to completedness are performed across a corpus of ten novels comprising 846,913 words. These novels were parsed with ESG, resulting in 75,289 fully-parsed clauses (22,505 of 59,816 sentences produced errors).

4.2.1 Manual Marking

884 clauses from the parsed corpus were manually marked according to completedness. Of these, 109 were rejected because of parsing problems, and 160 rejected because they described states. The remaining 615 clauses were divided into training and test sets such that the distribution of classes was equal. The baseline method in this case achieves 63.3% accuracy.

The linguistic test was selected for this task by Passonneau (1988): If a clause in the past progressive necessarily entails the past tense reading, the clause describes a non-culminated event. For example, *We were talking just like men* (non-culm.) entails that *We talked just like men*, but *The woman was building a house* (culm.) does not necessarily entail that *The woman built a house.* Cross-checking between linguists shows high agreement. In particular, in a pilot study manually annotating 89 clauses from this corpus according to stativity, two linguists agreed 81 times. Of 57 clauses agreed to be events, 46 had agreement with respect to completedness.

The verb *say* (point), which occurs nine times in the test set, was initially marked incorrectly as culminated, since points are non-extended and therefore cannot be placed in the progressive. After some initial experimentation, we corrected the class of each occurrence of *say* in the data.

4.2.2 Individual Indicators

Table 7 is analogous to Table 5 for completeness. The differences in culminated and non-culminated means are statistically significant (p < .05) for the first six indicators. However, for completeness, no indicator was shown to significantly improve classification accuracy over the baseline.

Linguistic Indicator	Culm Mean	Non-Culm Mean	T-test P-value
perfect	7.87%	2.88%	0.0000
temporal adverb	5.60%	3.41%	0.0000
manner adverb	0.19%	0.61%	0.0008
progressive	3.02%	5.03%	0.0031
past/pres partic	14.03%	17.98%	0.0080
no subject	30.77%	26.55%	0.0241
duration *in*-PP	0.27%	0.06%	0.0626
present tense	17.18%	14.29%	0.0757
duration *for*-PP	0.34%	0.49%	0.1756
continuous adverb	0.10%	0.49%	0.2563
frequency	345.86	286.55	0.5652
"not" or *"never"*	3.41%	3.15%	0.6164
evaluation adverb	0.46%	0.39%	0.7063
past tense	53.62%	54.36%	0.7132

Table 7: Indicators discriminate between culminated and non-culminated events.

		Culminated		Non-Culm	
	acc	recall	prec	recall	prec
CART	74.0%	86.2%	76.0%	53.1%	69.0%
llr	70.5%	83.1%	73.6%	48.7%	62.5%
llr2	67.2%	81.5%	71.0%	42.5%	57.1%
GP	68.6%	77.3%	74.2%	53.6%	57.8%
dt	68.5%	86.2%	70.6%	38.1%	61.4%
bl	63.3%	100.0%	63.3%	0.0%	100.0%
bl2	70.8%	94.9%	69.8%	29.2%	76.7%

Table 8: Comparison of four learning methods and two performance baselines, distinguishing culminated from non-culminated events.

4.2.3 Indicators in Combination

As shown in Table 8, the highest accuracy, 74.0%, was attained by CART. A binomial test shows this is a significant improvement over the 63.3% baseline.

The increase in non-culminated recall illustrates a greater improvement over the baseline. As shown in Table 8, non-culminated recalls of up to 53.6% were achieved by the learning methods, compared to 0.0%, achieved by the baseline.

Additionally, a non-culminated F-measure of 61.9 was achieved by GP, when optimizing for F-Measure, improving over 53.7 attained by the optimal uninformed baseline. F-measure computes a tradeoff between recall and precision (Van Rijsbergen, 1979). In this work, we weigh recall and precision equally, in which case,

$$F - measure = \frac{recall * precision}{(recall + precision)/2}$$

Automatic methods highly prioritized the *perfect* indicator. The induced decision tree uses the *perfect* indicator as its first discriminator, log-linear regression ranked the *perfect* indicator as fourth out of fourteen, function trees created by GP include the *perfect* indicator as one of five indicators used together to increase classification performance, and the *perfect* indicator tied as most highly correlated with completedness (cf. Table 7).

5 Discussion

Since certain verbs are aspectually ambiguous, and, in this work, clauses are classified by their main verb only, a second baseline approach would be to simply memorize the majority aspect of each verb in the training set, and classify verbs in the test set accordingly. In this case, test verbs that did not appear in the training set would be classified according to majority class. However, classifying verbs and clauses according to numerical indicators has several important advantages over this baseline:

- **Handles rare or unlabeled verbs.** The results we have shown serve to estimate classification performance over "unseen" verbs that were not included in the supervised training sample. Once the system has been trained to distinguish by indicator values, it can automatically classify any verb that appears in unlabeled corpora, since measuring linguistic indicators for a verb is fully automatic. This also applies to verbs that are underrepresented in the training set. For example, one node of the resulting decision tree trained to distinguish according to stativity identifies 19 stative test cases without misclassifying any of 27 event test cases with verbs that occur only one time each in the training set.
- **Success when training doesn't include test verbs.** To test this, all test verbs were eliminated from the training set, and log-linear regression was trained over this smaller set to distinguish according to completedness. The result is shown in Table 8 ("llr2"). Accuracy remained higher than the baseline "bl" (bl2 not applicable), and the recall tradeoff is felicitous.
- **Improved performance.** Memorizing majority aspect does not achieve as high an accuracy as the linguistic indicators for

completedness, nor does it achieve as wide a recall tradeff for both stativity and completedness. These results are indicated as the second baselines ("bl2") in tables 6 and 8, respectively.

- **Scalar values assigned to each verb** allow the tradeoff between recall and precision to be selected for particular applications by selecting the classification threshold. For example, in a separate study, optimizing for F-measure resulted in a more dramatic tradeoff in recall values as compared to those attained when optimizing for accuracy (Siegel, 1998b). Moreover, such scalar values can provide input to systems that perform reasoning on fuzzy or uncertainty knowledge.
- **This framework is expandable** since additional indicators can be introduced by measuring the frequencies of additional aspectual markers. Furthermore, indicators measured over multiple clausal constituents, e.g., main verb-object pairs, alleviate verb ambiguity and sparsity and improve classification performance (Siegel, 1998b).

6 Conclusions

We have developed a full-scale system for aspectual classification with multiple linguistic indicators. Once trained, this system can automatically classify all verbs appearing in a corpus, including "unseen" verbs that were not included in the supervised training sample. This framework is expandable, since additional lexicosyntactic markers may also correlate with aspectual class. Future work will extend this approach to other semantic distinctions in natural language.

Linguistic indicators successfully exploit linguistic insights to provide a much-needed method for aspectual classification. When combined with a decision tree to classify according to stativity, the indicators achieve an accuracy of 93.9% and stative recall of 74.2%. When combined with CART to classify according to completedness, indicators achieved 74.0% accuracy and 53.1% non-culminated recall.

A favorable tradeoff in recall presents an advantage for applications that weigh the identification of non-dominant classes more heavily (Cardie and Howe, 1997). For example, correctly identifying occurrences of *for* that denote event durations relies on positively identifying non-culminated events. A system that summarizes the duration of events which incorrectly classifies *"She ran (for a minute)"* as culminated will not detect that *"for a minute"* describes the duration of the *run* event. This is because durative *for*-PPs that modify culminated events denote the duration of the ensuing state, e.g., *I left the room for a minute.* (Vendler, 1967)

Our analysis has revealed several insights regarding individual indicators. For example, both duration *in*-PP and *manner* adverb are particularly valuable for multiple aspectual distinctions – they were ranked in the top two positions by log-linear modeling for both stativity and completedness.

We have discovered several new linguistic indicators that are not traditionally linked to aspectual class. In particular, verb frequency with no deep subject was positively correlated with both stativity and completedness. Moreover, four other indicators are newly linked to stativity: (1) Verb `frequency`, (2) occurrences modified by *"not"* or *"never"*, (3) occurrences in the past or present participle, and (4) occurrences in the *perfect* tense. Additionally, another three were newly linked to completedness: (1) occurrences modified by a *manner* adverb, (2) occurrences in the past or present participle, and (3) occurrences in the *progressive*.

These new correlations can be understood in pragmatic terms. For example, since points (non-culminated, punctual events, e.g., *hiccup*) are rare, punctual events are likely to be culminated. Therefore, an indicator that discriminates events according to extendedness, e.g., the *progressive*, past/present participle, and duration *for*-PP, is likely to also discriminate between culminated and non-culminated events.

As a second example, the *not/never* indicator correlates with stativity in medical reports because diagnoses (i.e., states) are often ruled out in medical discharge summaries, e.g., *"The patient was* **not** *hypertensive,"* but procedures (i.e., events) that were not done are not usually mentioned, e.g., *"?An examination was* **not** *performed."*

Acknowledgements

Kathleen R. McKeown was extremely helpful regarding the formulation of this work and Judith L. Klavans regarding linguistic techniques, and they, along with Min-Yen Kan and Dragomir R. Radev provided useful feedback on an earlier draft of this paper.

This research was supported in part by the Columbia University Center for Advanced Technology in High Performance Computing and Communications in Healthcare (funded by the New York State Science and Technology Foundation), the Office of Naval Research under contract N00014-95-1-0745 and by the National Science Foundation under contract GER-90-24069.

References

C. Cardie and N. Howe. 1997. Improving minority class prediction using case-specific feature weights. In D. Fisher, editor, *Proceedings of the Fourteenth International Conference on Machine Learning.* Morgan Kaufmann.

P.K. Chan and S.J. Stolfo. 1993. Toward multistrategy parallel and distributed learning in sequence analysis. In *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology.*

W. Cohen. 1995. Fast effective rule induction. In *Proc. 12th Intl. Conf. Machine Learning*, pages 115–123.

N. Cramer. 1985. A representation for the adaptive generation of simple sequential programs. In J. Grefenstette, editor, *Proceedings of the [First] International Conference on Genetic Algorithms.* Lawrence Erlbaum.

B.J. Dorr. 1992. A two-level knowledge representation for machine translation: lexical semantics and tense/aspect. In James Pustejovsky and Sabine Bergler, editors, *Lexical Semantics and Knowledge Representation.* Springer Verlag, Berlin.

D.R. Dowty. 1979. *Word Meaning and Montague Grammar.* D. Reidel, Dordrecht, W. Germany.

R. O. Duda and P.E. Hart. 1973. *Pattern Classification and Scene Analysis.* Wiley, New York.

J.H. Friedman. 1977. A recursive partitioning decision rule for non-parametric classification. *IEEE Transactions on Computers.*

V. Hatzivassiloglou and K. McKeown. 1993. Towards the automatic identification of adjectival scales: clustering adjectives according to meaning. In *Proceedings of the 31st Annual Meeting of the ACL*, Columbus, Ohio, June. Association for Computational Linguistics.

J.L. Klavans and M. Chodorow. 1992. Degrees of stativity: the lexical representation of verb aspect. In *Proceedings of the 14th International Conference on Computation Linguistics.*

J.L. Klavans. 1994. Linguistic tests over large corpora: aspectual classes in the lexicon. Technical report, Columbia University Dept. of Computer Science. unpublished manuscript.

J.R. Koza. 1992. *Genetic Programming: On the programming of computers by means of natural selection.* MIT Press, Cambridge, MA.

M.C. McCord. 1990. SLOT GRAMMAR. In R. Studer, editor, *International Symposium on Natural Language and Logic.* Springer Verlag.

M. Moens and M. Steedman. 1988. Temporal ontology and temporal reference. *Computational Linguistics*, 14(2).

R.J. Passonneau. 1988. A computational model of the semantics of tense and aspect. *Computational Linguistics*, 14(2).

F. Pereira, N. Tishby, and L. Lee. 1993. Distributional clustering of english words. In *Proceedings of the 31st Conference of the ACL*, Columbus, Ohio. Association for Computational Linguistics.

J.R. Quinlan. 1986. Induction of decision trees. *Machine Learning*, 1(1):81–106.

J.R. Quinlan. 1993. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, San Mateo, CA.

T.J. Santner and D.E. Duffy. 1989. *The Statistical Analysis of Discrete Data.* Springer-Verlag, New York.

H. Schütze. 1992. Dimensions of meaning. In *Proceedings of Supercomputing.*

E.V. Siegel and K.R. McKeown. 1996. Gathering statistics to aspectually classify sentences with a genetic algorithm. In K. Oflazer and H. Somers, editors, *Proceedings of the Second International Conference on New Methods in Language Processing*, Ankara, Turkey, Sept. Bilkent University.

E.V. Siegel. 1997. Learning methods for combining linguistic indicators to classify verbs. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, Providence, RI, August.

E.V. Siegel. 1998a. Disambiguating verbs with the wordnet category of the direct object. In *Procedings of the Usage of WordNet in Natural Language Processing Systems Workshop*, Montreal, Canada.

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

C.J. Van Rijsbergen. 1979. *Information Retrieval.* Butterwoths, London.

Z. Vendler. 1967. Verbs and times. In *Linguistics in Philosophy.* Cornell University Press, Ithaca, NY.