Linguistic Indicators for Language Understanding: Using machine learning methods to combine corpus-based indicators for aspectual classification of clauses

-Ph.D. Dissertation-

Eric V. Siegel
Department of Computer Science
Columbia University
New York, NY 10027
evs@cs.columbia.edu

May 20, 1998

Abstract

Linguistics as a field has provided enormous insights that describe how the thoughts behind language are reflected by the structure of sentences. For example, one writes a paper *in* one week, but rides a bicycle *for* one hour. This illustrates how prepositions (*in* and *for*) correspond to the type of event. Specifically, *in* modifies a completed process, while *for* modifies an ongoing process. The area explored by this thesis is, how can we best put our understanding of linguistics to use in order to tap into the vast knowledge encoded in texts?

The ability to distinguish *stative* clauses, e.g., *"She resembles her mother,"* from *event* clauses, e.g., *"She ran down the street,"* is a fundamental component of natural language understanding. These two high-level categories correspond to primitive distinctions in many domains, including, for example, the distinctions between *diagnosis* and *procedure* in the medical domain. *Stativity* is the first of three high-level distinctions that compose the *aspectual class* of a clause. These distinctions in meaning have been well motivated by work in linguistics and natural language understanding.

Aspectual classification is a necessary component for applications that perform certain natural language interpretation, natural language generation, summarization, information retrieval, and machine translation tasks. This is because each of these applications requires the ability to reason about time.

In this thesis, I develop a system to perform aspectual classification with linguistically-based, numerical indicators. These *linguistic indicators* make use of an array of aspectual *markers*, each of which has an associated constraint on aspectual class. For example, only clauses that describe an *event* can appear with the *progressive* marker, e.g., *"I was eating breakfast."* Therefore, the category of a verb or phrase is reflected by a numerical indicator that measures how often it occurs in the progressive. The values for such linguistic indicators are computed automatically across corpora of text. We develop and evaluate fourteen indicators over unrestricted sets of verbs occurring across two corpora. Our analysis reveals a predictive value for several indicators that have not previously been conjectured to correlate with aspect in the linguistics literature.

Then, machine learning is used to combine multiple indicators in order to improve classification performance. The models automatically derived by learning are manually examined, revealing several linguistic insights regarding the indicators and their interactions. Three machine learning techniques are compared for this task: decision tree induction, a genetic algorithm, and log-linear regression.

We conclude that linguistic indicators successfully exploit linguistic insights to provide a much-needed method for aspectual classification. Future work will extend this approach to other semantic distinctions in natural language.

Acknowledgements

I have thoroughly enjoyed and benefited from my career as a graduate student due to the strategic guidance of my advisor, Kathleen R. McKeown. Kathy encouraged me to explore various areas, and ultimately provided priceless support in my chosen research topic. I am fortunate to have discovered the natural language problem that interests me most, applying machine learning methods, my other primary field of interest, to boot.

The other four members of my thesis committee also logged many hours over the last few years, providing valuable insights and feedback for my work. Judith L. Klavans was my primary linguistic consultant and a great source of enthusiasm. Conversations about the linguistic side of this work with Philip Resnik were also extremely stimulating. The two of them kept me on target by providing many linguistic observations to further support my work, as well as teaching me the standard linguistic protocols for testing and evaluation. John R. Kender, in general, served as a priceless guide through my six years of graduate school teaching and research. Moreover, John helped me to more formally assess and defend semantic distinctions set forth by linguists; the design of Section 2.2 resulted from discussions with John. Finally, Salvatore J. Stolfo was my primary machine learning consultant. He was like the Judith and Philip of learning; only due to his insights is the machine learning side of this work meaningful from the vantage of that field.

What can I say about Jacques Robin, my graduate officemate of 5 years? Jacques became a close friend and a great basketball coach. He showed me how to balance the untamed ambitions of a new graduate student with the pragmatics of actual execution, and he showed me how to balance a research life with a night life. Jacques is an inspired, hilarious spark of positive action. He gave me pep talks and companionship. He's better than a pet dog.

The remaining members of Columbia's natural language processing group were incredibly important for feedback on this work. Vasileios Hatzivassiloglou was my primary statistical consultant; he is an encyclopedia of mathematical models, and was a great source of perspective on my work. He helped formalize one interpretation of my results, discussed in Section 2.2.4. Also, Vasileios designed and implemented the clustering of verbs described in Section 6.2.4. James Shaw first pointed out to me that *have* is statively ambiguous, resulting in the work reported in Section 4.3. Other helpful members of Columbia's NLP group include or included Michael Elhadad, Eleazar Eskin, Pascale Fung, Hongyan Jing, Min-Yen Kan, Diane Litman, Rebecca Passennau, Dragomir Radev, Ruth Reeves, Nina Wacholder, and Robert Weida.

Many other people in my life also lended a helping hand. Three professors from my undergraduate days at Brandeis University were extremely formative in my development as a researcher, and invested a lot of time to help me grow: Harry Mairson, Richard Alterman, and James Pustejovsky. Harry showed me the humor behind computer science and helped me first understand advanced research material. Rick was a great undergraduate thesis advisor (language generation), and inspired me to ride my bike to work every day. James first got me interested in linguistics, and taught me to indulge in the deep questions of artificial intelligence. Alexander Day Chaffee, my friend from childhood days of hacking, helped me enormously with the interpretation of numerical results, and even had a couple helpful linguistic insights. Alex knows what is interesting about life and what is fun about computers; he will always be the Dalang of my Gamelan. Other important contributors include: Peter Angeline, Ken Church, Ivan A. Derzhanski, Larry Eshelman, Jussi Karlgren, John Koza, Nelson Minar, Andreas L. Prodromidis, Conor Ryan, Astro Teller, Andy Singleton, David Schaffer, and Dakai Wu.

Thank you, my parents, Lisa A. Schamberg and Andrew Siegel, who are the ultimate best you could want. I appreciate their diligent struggle to find the optimal frequency with which to ask, “Is it done yet?” My sister, Rachel A. Siegel, is the center of the universe. Also, I’d like to thank my issues for remaining repressed (Alexander D. Chaffee, personal communication).

This research is 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.

Contents

List of Figures

Chapter 1

Introduction

"To do is to be." - Descartes
"To be is to do." - Alexander D. Chaffee
"Do-be-do-be-do." - Frank Sinatra

Many challenges for natural language processing require the *classification* of words or phrases as one of a small number of categories. For example, *word sense disambiguation* is the process of finding the meaning of an ambiguous word from its context, e.g., *river bank* versus *Federal bank*. A second example, *aspectual* classification, is the problem of mapping a clause (e.g., a simple sentence) to one of 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."*

Aspectual classification is necessary for interpreting even the most simple narratives in natural language. This is because, in general, the sequential order of clauses are not enough to determine the underlying chronological order. For example, consider:

"*John entered the room* (**event**). *Mary stood up* (**event**)."

In this case, the first sentence describes an event that takes place before the event described by the second sentence. However, in,

"*John entered the room* (**event**). *Mary was seated behind the desk* (**state**)."

the second sentence describes a **state**, which begins before the event described by the first sentence. Aspectual classification is a necessary step towards automatically identifying relationships in time between sentences.

The ability to distinguish **stative** clauses, e.g., *"She resembles her mother,"* from **event** clauses, e.g., *"She ran down the street,"* is a fundamental component of natural language understanding. These two high-level categories correspond to primitive distinctions in many domains, including, for example, the distinctions between *diagnoses* and *procedure* in the medical domain, and between *analyses* and *activity* in the financial domain.

Stativity is the first of three high-level distinctions that compose the *aspectual class* of a clause. Events are further distinguished along two other dimensions. First, *completedness* determines whether an event reaches a culmination or completion 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**. Second, *atomicity*, distinguishes *atomic* (instantaneous) events, such as, *"She noticed the picture on the wall,"* from *extended* events, such as, *"She ran to the store."* By dividing events along the second and third dimensions we derive four classes of events

There is an array of semantic entailments related to aspectual category that linguistically motivates each of these three particular semantic distinctions. For example, one such entailment pertains to prepositional phrases that denote the duration of a state or event. *"For an hour"* can denote the duration of a **non-culminated event**, as in,

"I gazed at the sunset **for an hour***."*

In this case, *an hour* is the duration of the *gazing* event. However, when applied to a **culminated event**, it denotes the duration of the resulting state, as in,

"I left the room **for an hour***."*

In this case, *an hour* is not the duration of the *leaving* event, but, rather, the duration of what resulted from *leaving*, i.e., being gone.

Such *aspectual entailments* also illustrate the value of automatically classifying clauses according to aspect; once the category of a clause has been identified, they support an array of inferences pertaining to time. These inferences are crucial for natural language understanding and generation applications such as machine translation, processing medical reports, summarization, and augmenting lexical databases.

In this thesis, I develop a system to perform aspectual classification with linguistically-based, numerical indicators. The values for indicators are computed automatically across corpora of text. The indicators are then evaluated over unrestricted sets of verbs occurring across two corpora. Our analysis reveals a predictive value for several indicators that have not previously been conjectured to correlate with aspect in the linguistics literature. Then, machine learning is used to combine multiple indicators in order to improve classification performance and to ascertain linguistic insights about indicators and their interactions. Three machine learning techniques are compared for this task: decision tree induction, a genetic algorithm, and log-linear regression.

1.1 Why Aspectual Classification?

Aspectual classification is a necessary component for applications that perform certain natural language interpretation, natural language generation, summarization, information retrieval, and machine translation tasks. This is because each of these applications requires the ability to reason about time, i.e., *temporal reasoning.*

Assessing temporal relationships is a prerequisite for many language interpretation tasks, e.g., detecting upward or downward trends in the stock market domain, or inferring sequences of medical procedures in medical domains. Many applications that process medical reports require aspectual classification because a patient's medical progress and history are established as a series of states and events that are temporally related. For example, one task is to automatically complete a database entry for a patient by processing a medical discharge summary detailing a patient's visit to the hospital.

Aspect is also crucial for lexical choice and tense selection in machine translation [Moens and Steedman, 1988, Klavans and Chodorow, 1992, Klavans, 1994, Dorr, 1992]. For example, aspectual classification is necessary for translation between certain pairs of languages because some languages have explicit *perfective* markers and others do not. The *perfective* marker is used in many languages, such as Bulgarian and Russian, to indicate completedness. Therefore, a system translating from a language without explicit perfective markers such as English to one with explicit perfective markers must first detect the aspectual category of an input phrase in order to determine the form of the output [Stys, 1991, Dorr, 1992].

In order to meet the needs of natural language applications, several researchers have developed models that incorporate aspectual class to assess temporal relationships and interpret temporal modifiers [Hwang and Schubert, 1991, Schubert and Hwang, 1990, Dorr, 1992, Passonneau, 1988, Moens and Steedman, 1988, Hitzeman et al., 1994]. However, these models rely on the ability to first automatically detect the fundamental aspectual category of a clause, which is a function of the main verb and other complement modifiers.

Further details regarding the need for aspectual classification by natural language interpretation, generation and summarization applications, as well as tasks in information retrieval, machine translation and specialized lexical databases are given in Chapter 2.

1.2 Why Linguistic Indicators?

Aspectual classification presents a challenging problem. The number of ways in which clausal constituents interactively influence aspectual class is unknown. However, syntax alone is not sufficient, and the lexical head of multiple constituents (e.g., the verb phrase and the direct object) can be influential factors [Pustejovsky, 1991]. In addition to specialized lexical knowledge describing open class constituents of the clause, aspectual classification requires a method to detect the interactions between these constituents which produce aspectual class.

Many clauses can be categorized based on their main verb only, as illustrated in the following chapters. Therefore, aspectually categorizing verbs is the first step in this direction. However, the primary category of a verb is often domain-dependent, so it is necessary to perform a specialized analysis for each domain.

Statistical approaches to language developed in recent years use lexical cooccurrence to automatically group or rank open class words according to semantic distinctions in general [Hatzivassiloglou and McKeown, 1993a, Lafferty and Mercer, forthcoming, Pereira et al., 1993a, Lankhorst and Moddemeijer, 1992, Schűtze, 1992], and aspectual class in particular [Klavans and Chodorow, 1992]. There are three advantages to such automation:

- The cost of *manually* generating a linguistic model for a particular classification problem, which is often prohibitive, is avoided.

- Biases introduced by a human lexicographer may be suboptimal.
- Automated approaches are extensible to multiple natural language classification problems, across multiple domains and multiple languages.

Naturally occurring text contains vast amounts of linguistically encoded information pertaining to aspectual classification. In particular, there is an array of *aspectual markers* with associated aspectual constraints. For example, only clauses that describe an **event** can appear with the *progressive* marker, e.g., *"I was eating breakfast."* Since each aspectual marker constrains the category of clauses it appears with, the occurrences and distributions of these markers may reflect the categories of the verbs and phrases they appear with.

However, the best way to make use of these markers is not obvious. In general, 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. Furthermore, aspectual transformations, described in the following chapter, can allow the clause to appear with a marker with apparently conflicting constraints on aspectual class. These factors makes it difficult for a system to detect the fundamental aspectual class of a clause based soly on the presence or absence of aspectual markers.

Therefore, as with most statistical methods for natural language, the linguistic constraints associated with markers are best exploited by a system that measures cooccurrence frequencies. In particular, we measure the frequencies of aspectual markers across verbs, and also across *multiple* clausal constituents. This way, the aspectual *tendencies* of verbs and phrases are measured. These tendencies are likely to correlate with fundamental aspectual class. For example, a verb that appears more frequently in the progressive is more likely to describe an event. The cooccurrence frequency of a linguistic marker is a *linguistic indicator*.

The development of a method to measure and combine linguistic indicators achieves the general benefits of statistical corpus analysis: it can be applied over multiple domains and languages, and can be extended to multiple semantic distinctions. Furthermore, once a system has been trained to distinguish by indicator value, it can automatically classify all verbs appearing in a corpus, including "unseen" verbs that were not included in the supervised sample. These features of numerical indicators are illustrated with our results and analysis.

There are several advantages to classifying verbs with *linguistic* indicators in particular, which measure the frequencies of linguistic phenomena such as aspectual markers, rather than indicators that measure cooccurrence with open class lexical items such as the head of the direct object. These advantages are described and illustrated in the chapters of this thesis.

1.3 Summary of Contributions

We have developed a system for aspectual classification using linguistic indicators. The values of indicators are computed automatically across corpora of text. A total of fourteen indicators were evaluated over unrestricted sets of verbs from two corpora of text. Furthermore, machine learning techniques were employed to combine multiple indicators in order to improve classification performance and to gain linguistic insights about the indicators and their interactions.

We conclude that linguistic indicators successfully exploit linguistic insights to provide a much-needed method for aspectual classification. Future work will extend this approach to other semantic distinctions in natural language.

This section summarizes the main contributions of this work:

- Fourteen linguistic indicators are evaluated for aspectual classification.
  - The analysis reveals a predictive value for several indicators that have not previously been conjectured to correlate with aspect in the linguistics literature.
  - When used in combination, indicators improve classification performance as compared to uninformed baseline approaches.
  - In this way, the first lower bounds of performance for the automatic classification of clauses according to two aspectual distinctions have been established.
  - Since these indicators are evaluated over two aspectual distinctions, it is possible to assess the statistical differences and similarities between the two.
- Multiple linguistic indicators are combined with machine learning.
  - Three machine learning techniques are compared for this task: decision tree induction, a genetic algorithm, and log-linear regression.
  - Each of these learning methods demonstrate an increase in classification performance. Their relative performances vary depending on the classification task, and the evaluatory measure of performance.
  - The models automatically derived by the learning methods are manually examined to reveal several linguistic insights regarding linguistic indicators and their interactions.
- Linguistic indicators are measured over combinations of clausal constituents such as the main verb and the direct object. This approach is shown to alleviate difficulties with verb sparsity and ambiguity.
- Semantically-motivated word groupings are shown to improve aspectual classification in several ways:
  - Classifying instances of the ambiguous verb *have* by way of the Wordnet semantic category of the direct object.
  - Statistically established groupings of verbs are shown to correlate with the aspectual category of verbs.
  - Indicators are measured over Wordnet categories of the direct object in order to alleviate verb ambiguity and direct object sparsity.
  - Manually derived groups of temporal nouns and adverbs help define and measure certain linguistic indicators.

  - High-level semantic categories of aspectually ambiguous verbs are manually derived from linguistic principles. These categories, which are postulated to include all statively ambiguous verbs, can help with aspectual disambiguation.

- The ways in which aspectual classification must be incorporated into several natural language applications are discussed in detail.

1.4 Thesis Overview

The organization of this thesis is simple. The problem of aspectual classification is introduced in Chapter 2. Aspectual markers and their constraints are detailed. The identification of these constraints serves three purposes:

- They validate the hierarchy of aspectual distinctions.

- They demonstrate how natural language applications can make use of aspectual classification.

- They provide a method of measuring how people tend to use verbs, i.e., *linguistic indicators.*

Our approach is described Chapter 3. The use of *linguistic indicators* is further motivated, and the method by which they are automatically measured is described. Furthermore, this chapter describes the variety of measures we employ to evaluate the automatic classification of clauses according to aspect. These measures reflect the varying demands among natural language applications.

The results of our evaluations of linguistic indicators are given in Chapters 4 and 5. Chapter 4 evaluates the indicators for distinguishing states from events, and Chapter 5 for distinguishing culminated from non-culminated events. Both of these chapters evaluate the indicators individually, and then in combination. Various evaluatory measures and methods are employed.

The application of automatic machine learning methods to coordinate and combine multiple linguistic indicators is described in Chapter 6. Four learning paradigms are described, and the resulting classification performance of each is evaluated. Furthermore, the models automatically derived by the learning methods are manually examined to reveal several linguistic insights regarding linguistic indicators and their interactions.

Finally, Chapter 7 concludes with a detailed summary of contributions, as well as future research directions.

Chapter 2

Aspect in Natural Language

The ability to distinguish **stative** clauses, e.g., *"She resembles her mother,"* from **event** clauses, e.g., *"She ran down the street,"* is a fundamental component of natural language understanding. These two high-level categories correspond to primitive distinctions in many domains, including, for example, the distinctions between *diagnoses* and *procedures* in the medical domain, and between *analyses* and *activity* in the financial domain.

Stativity is the first of three fundamental temporal distinctions that compose the *aspectual class* of a clause. The aspectual class of a clause reflects the underlying temporal structure of the event(s) described by the clause. For example, the aspectual class of *"The boy held the tiller"* is *process*, an event with no explicit temporal boundaries, while *"You will find us there"* is a *culminated process*, an event with a determined point of completion.

Aspectual classification is a necessary prerequisite for interpreting certain adverbial adjuncts, as well as identifying temporal constraints between sentences in a discourse [Moens and Steedman, 1988, Dorr, 1992, Klavans, 1994]. It is also crucial for lexical choice and tense selection in machine translation [Moens and Steedman, 1988, Klavans and Chodorow, 1992, Klavans, 1994, Dorr, 1992].

The number of ways in which clausal constituents interactively influence aspectual class is unknown. However, syntax alone is not sufficient, and the lexical head of multiple constituents (e.g., the verb phrase and the direct object) can be influential factors. Given the difficulty in identifying the precise factors determining the aspectual class of a clause, some researchers have used empirical analysis of corpora to develop probabilistic weights that aid in aspectual classification [Klavans and Chodorow, 1992].

This chapter has four sections. The first section introduces aspectual classification, detailing temporal constraints and entailments related to aspect. Then, we illustrate the ways in which clausal constituents can contribute to aspectual class, and describe a hierarchy of aspectually ambiguous verbs. Finally, I describe forms of aspectual coercion in order to pinpoint the particular level of classification approached with this thesis.

In the second section of this chapter, I provide the historical and philosophical framework behind aspectual classification, enumerating empirical support to validate the aspectual distinctions. The third section describes several natural language applications that rely

Table 2.1: Aspectual classes. This table comes from Moens and Steedman [1988].

	EVENTS		STATES
	punctual	extended	
Culminated	CULMINATION recognize	CULMINATED PROCESS build a house	understand
Non-Culm	POINT hiccup	PROCESS run, swim, walk	

on aspectual classification. The last section concludes with a summary, emphasizing the importance and problems of aspectual classification.

2.1 Aspect: Introduction and Related Work

Table 2.1 summarizes the three aspectual distinctions, which compose five aspectual categories. The first distinction, *stativity*, determines whether a clause is a *state* (e.g., *"Mark is happy"*) or an *event* (e.g., *"Renée ran down the street"*). Intuitively, a state is something that is true or false about the world, and an event is something that can *happen.*

Events are further distinguished along two other dimensions. First, *completedness*, determines whether an event reaches a completion point 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**. Second, *atomicity*, distinguishes *punctual* events from *extended* events. *Punctual* events, such as, *"She noticed the picture on the wall,"* are at the most specific level of temporal granularity, i.e. instantaneous, whereas, *"She ran to the store"* is *extended.*

By dividing events along the second and third dimensions we derive four classes of events, as shown in Table 2.1, *culmination*, *culminated process*, *process*, and *point.* We adopt the terminology set forth by Moens and Steedman [1988] to name the five aspectual categories.

These aspectual distinctions are motivated by a series of syntactic and entailment *constraints* which are described in the first three subsections below. Next, we look at how the constituents of a clause interact to affect the aspectual category of the clause. Then we describe semantic groups of verbs that are aspectually ambiguous. Finally, we introduce the concept of a clause's *fundamental* aspectual category.

2.1.1 Aspectual Markers and Constraints

Certain features of a clause, such as adjuncts and tense, are constrained by and contribute to the aspectual class of the clause [Vendler, 1967, Dowty, 1979, Pustejovsky, 1991, Passonneau, 1988, Klavans, 1994]. Tables 2.2 illustrates an array of linguistic constraints, as summarized by Klavans [1994]. Each entry in this table describes an aspectual *marker* and

Table 2.2: Aspectual markers and associated constraints on aspectual class, primarily from the summary of Klavans [1994].

If a clause can occur:	then it must be:
with a temporal adverb (e.g., *then*)	**Event**
in *progressive*	**Extended Event**
in *imperative*	**Event**
with *adverbs of intention*	**Event**
as a complement of *force/persuade*	**Event**
in a *pseudo-cleft* (e.g., *What she did was think.*)	**Event** (*agentive* only)
in a passivized phrase	**Event**
after *"What happened was..."*	**Event**
as a complement of *aspectual auxiliaries* (e.g., *stop*)	**Event**
redundantly, e.g. *"trying and trying"*	**Extended Event**
with a duration *in*-PP (e.g., *in an hour*)	**Culminated Event**
in the *perfect* tense	**Culminated Event** or **State**
as a complement of *finish*	**Culminated Process**
as a complement of a *durative adverbial* (e.g., *all day*)	**Extended Event** or **State**
with an indefinite implicit object (e.g., *We just ate* ε.)	**Event**
with *"...and still at it"*	**Event**

the constraints on the aspectual category of any clause that appears with that marker. For example, only *events* can occur in a pseudo-cleft, e.g.,

"What its parents did was run off" (**event**),

which contrasts with,

*"*What we did was know what is on the other side"* (**state**).

However, the pseudo-cleft tests for agentivity or purposefulness, so some **events** clauses are not permitted as a pseudo-cleft, e.g.,

*"*What she did was notice the picture on the wall"* (**event**).

As another example, a clause must be *extended* to appear in the progressive tense, e.g.,

"He was prospering in India" (**extended event**),

which contrasts with,

*"*You were noticing that I can hardly be blamed..."* (**atomic event**).

Since the perfect tense requires that the clause it occurs in must entail a consequent state, an event must be *culminated* to appear in the perfect tense, for example,

"Thrasymachus had made an attempt to get the argument into his own hands" (**culminated event**),

which contrasts with,

*"*He has cowered down in a paralysis of fear"* (**non-culminated event**).

Table 2.3: Aspectual entailments, as more comprehensively summarized by Klavans [1994].

If a clause occurring:	necessarily entails:	then it must be:
in past tense with *for-adverbial* temporal adjunct	past tense reading	**State** or **non-Culminated Event**
in past progressive tense	past tense reading	**non-Culminated Event**
in present (progressive) tense	present perfect reading	**Process** or **State**
as an argument of *stopped*	past tense reading	**non-Culminated Event** or **State**
in past tense with *in-adverbial* temporal adjunct	past progressive tense reading with *during-adverbial* adjunct	**Culminated Process**
as a complement of *start/finish*	the *habitual* reading	**Culmination**
in simple present tense	the *habitual* reading	**Event**

Also, a culminated event can be modified by a duration *in*-PP, as in,

"You found us there **in ten minutes***"* (**culminated event**),

but a non-culminated event cannot, e.g.,

*"*The boy held the bottle* **in ten minutes***"* (**non-culminated event**),

nor can a state, e.g.,

*"*You loved him* **in ten minutes***"* (**state**).

As a last example, *"I* **purposely** *confessed that I had doubts"*, where *purposely* is an adverb of intention, is intelligible. Therefore, the clause describes an event.

In general, the presence of these linguistic markers in a particular clause indicates a constraint on the aspectual class of the clause, but the absence thereof does not place any constraint. This makes it difficult for a system to aspectually classify a clause based on the presence or absence of a marker. Therefore, these linguistic constraints are best exploited by a system that measures their frequencies. This is the approach detailed in Chapter 3.

2.1.2 Aspectual Entailments

Table 2.3 lists a set of aspectual entailments. Each entry in this table describes a linguistic phenomenon, a resulting entailment, and the constraints on aspectual class that apply if the resulting entailment holds. For example, if a clause that appears as an argument of *stop*, e.g., *"He stopped prospering in India"*, entails the past tense reading, e.g., *"He prospered in India"*, then the clause describes a non-culminated event. As another example, the simple present reading of an event, e.g., *"He jogs,"* denotes the *habitual* reading, i.e., *"every day,"* whereas the simple present reading of a state, e.g., *"He appears healthy,"* implies, *"at the moment."*

2.1.3 Temporal Constraints and Connectives

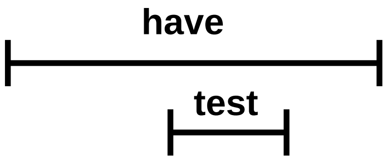
Q: *Why didn't the chicken cross the road?*

A: *It tried.*

Several researchers have developed models that incorporate aspectual class to assess temporal constraints and interpret temporal adverbials [Hwang and Schubert, 1991, Schubert and Hwang, 1990, Dorr, 1992, Passonneau, 1988, Moens and Steedman, 1988, Hitzeman et al., 1994, Parsons, 1990, Tenny, 1992]. For example, stativity must be identified to detect temporal constraints between clauses connected with *when*. For example, in interpreting,

She **had** *good strength when objectively* **tested**.[1]

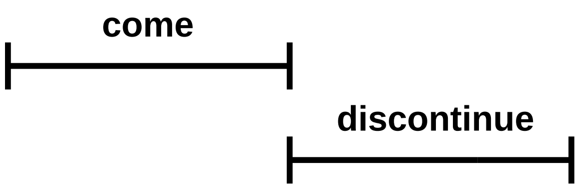
the *have*-**state** began before or at the beginning of the *test*-**event**, and ended after or at the end of the *test*-**event**:



However, in interpreting,

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

the *discontinue*-**event** began at the end of the *come*-**event**:



A similar phenomenon exists between sequential sentences [Hitzeman et al., 1994]. Consider:

*“John entered the room (***event***). Mary stood up (***event***).”*

*“John entered the room (***event***). Mary was seated behind the desk (***state***).”*

Temporal models also predict constraints on temporal adjuncts. For example, the following two sentences each have a durative temporal modifier, and a temporal at-location:

“At 3:00, I hiked the mountain in 2 hours.”

“*At 3:00, I reached the mountain top in 2 hours.”

The second sentence violates aspectual constraints since the at-location refers to the culmination point in time, and therefore constrains the time reference of the other modifier.

Such models also predict temporal relations between clauses combined with other connectives such as *before*, *after*, and *until*.

Models that incorporate aspect rely on the ability to automatically find the aspectual category of a clause. For example Passonneau describes an algorithm which depends on what is called *lexical aspect* [Passonneau, 1988], the aspectual information stored in the

[1]These examples of *when* come from the corpus of medical discharge summaries described in Chapter 4.

Table 2.4: Example clausal features and how they can influence aspectual class.

Feature:	Example:	Class:	Contrasting Example:	Class:
Verb	*John* **pet** *the lamb.*	**proc**	*John* **noticed** *the lamb.*	**culm proc**
Adverb	*Kathy hiccuped.*	**point**	*Kathy hiccuped* **continuously**.	**proc**
Predicate adj	*John drove the car.*	**proc**	*John drove the car* **ragged**.	**culm proc**
Particle	*John drove the car.*	**proc**	*John drove the car* **up**.	**culm proc**
Dir obj cat	*John saw* **Sue**.	**proc**	*John saw* **that Sue was happy**.	**culm**
Dir obj head	*Judith played the* **piano**.	**proc**	*Judith played the* **sonata**.	**culm proc**
Dir obj det	*John ate fries.*	**proc**	*John ate* **the** *fries.*	**culm proc**
Ind obj det	*Kathy showed people her car.*	**proc**	*Kathy showed* **the** *people her car.*	**culm proc**
Ind obj head	*Kathy showed* **people** *her car.*	**proc**	*Kathy showed* **Sal** *her car.*	**culm proc**
Prep obj head	*Judith looked around the* **store**.	**proc**	*Judith looked around the* **corner**.	**culm**
Prep obj det	*Kathy shot at deer.*	**proc**	*Kathy shot at* **the** *deer.*	**culm proc**
Tense	*Sal* **said** *that it helps.*	**culm**	*Sal* **says** *that it helps.*	**state**

lexicon, and Dorr augments Jackendoff's lexical entries with aspectual information [Dorr, 1992].

2.1.4 How Clausal Constituents Contribute to Aspectual Class

Although some verbs appear to connote only one aspectual class regardless of context, e.g., *stare* (**process**), many verbs are aspectually ambiguous. For example, *show* denotes a state in *"His lumbar puncture showed evidence of white cells,"* but denotes an event in *"He showed me the photographs."*

This ambiguity presents a difficulty for automatically classifying a verb because the aspectual class of a clause is a function of several clausal constituents in addition to the main verb [Passonneau, 1988, Moens and Steedman, 1988, Pustejovsky, 1991, Resnik, 1996, Olsen and Resnik, 1997]. Table 2.4 shows a set of *clausal features* that influence aspectual class. The importance of each feature is illustrated by showing two example sentences with distinct aspectual classifications. As this shows, different verbs are influenced by their clausal context in different ways.

The number of ways in which clausal constituents interactively influence aspectual class is unknown. However, syntax alone is not sufficient, and the lexical head of multiple constituents (e.g. the verb phrase and the direct object) can be influential factors. Specifically, the presence, syntactic category, lexical heads, and even plurality of certain arguments can influence the clause's aspectual class. For example, *"Sue played the piano"* signifies a process, while *"Sue played the sonata"* signifies a culminated process (this example comes from Moens and Steedman [1988]). As another example, *"I think that Mars is big"* describes a state, while *"I think about Mars"* describes a process.

Table 2.5: Groups of verbs that are statively ambiguous.

group:	example verbs	**event** sentence	**state** sentence
`communication`	*admit, allude, allude, appear, call, complain, confirm, convey, deny, describe, display, document, exhibit, externalize, indicate, note, present, recommend, report, reveal, say, show, specify, state, suggest*	*I* **said**, *"Hello."*	*I* **say** *it is correct.*
`cognition`	*associate, assume, attribute, base, care, consider, enjoy, include, judge, realize, recall, recognize, regard, remember, suspect, think, tolerate, understand, wish*	*I* **thought** *about them.*	*I* **think** *they are nice.*
`perception`	*feel, hear, look, perceive, see, smell, sound, taste, weigh*	*I* **felt** *the tablecloth.*	*I* **felt** *terrible.*
`location`	*accompany, allow, appear, block, clog, control, enter, exclude, extend, flow, follow, hold, include, lie, live (in NYC), obstruct, occlude, remain, rest, run, sit, stand, support,*	*I* **lay** *on the bed.*	*The book* **lies** *on the bed.*
`psych-movement`	*astonish, dismay, please, surprise, worry*	*You* **surprised** *me.*	*That* **suprises** *me.*
metaphorical	*work, run*	*I* **worked** *hard.*	*The machine* **works**.
carrier	*continue, remain*	*I* **continued** *to talk about it.*	*I* **continued** *to feel good.*

2.1.5 Types of Ambiguous Verbs

Some verbs are similar to one another in the way they combine with their clausal context to determine aspectual class. For example, *look* and *weigh* can each appear as events, e.g.,

I looked at the baby. (**event**)

I weighed the baby. (**event**)

and can also appear as states, as in,

The baby looked heavy. (**state**)

The baby weighed a lot. (**state**)

These examples suggest that the two verbs have similar sub-categorization frames which determine their aspectual category. There is also a relationship between their meanings, since they each describe a type of `perception` or `measurement`.

Because there is a correlation between subcategorization frame and semantic class [Levin, 1993], establishing a semantic hierarchy of aspectually ambiguous verbs will help determine the subcategorization frames of the verbs that fit into the hierarchy. Table 2.5 shows the top level of such a hierarchy of statively ambiguous verbs. Six groups are shown, each with a set of example verbs, and two contrasting sentences illustrating contrasting uses of an example verb from that group.

The first four groups shown contain semantically related verbs. Each verb in the first group, `communication`, can appear as either an event or state. Intuitively, this is because each verb can convey a communicative *act*, e.g.,

She showed me the photos. (**event**)

or, alternatively, a stable, unchanging situation, e.g.,

The xrays show no sign of growth. (**state**)

Verbs in the second group in Table 2.5, `cognitive`, can convey a mental *act*, e.g.,

When he mentioned bananas, she remembered Edward. (**event**)

or, alternatively, a mental state, e.g.,

I'll always remember Disney World. (**state**)

The third group, `perception`, is a subgroup of `cognition`. Similar groups of `perception` and `communication` verbs have previously been identified with respect to aspect in particular [Vendler, 1967, Dowty, 1979], and for general purposes beyond aspectual ambiguity [Levin, 1993]. The fourth group, `locative`, has previously been identified as "lay-verbs." [Dowty, 1979, Klavans, 1994]

The group *metaphorical* in Table 2.5 contains **event** verbs with idiomatic uses that are stative. These idiomatic uses correspond to a metaphorical interpretation of the **event** reading. For example,

I ran down the street. (**event**)

It runs in the family. (**state**)

Finally, *carrier* verbs simply reflect the aspectual class of their argument, as shown in Table 2.5.

2.1.6 Aspectual Transformations and Coercion

Several aspectual markers such as those shown in Table 2.2 transform the aspectual class of the clause they modify. For example, a duration *for*-PP, e.g. *"for 10 minutes,"* denotes the time duration of a (non-culminated) **process**, resulting in a **culminated process**, e.g.,

I stared at it. (**process**)

I stared at it for 10 minutes. (**culminated process**)

Some *aspectual auxiliaries* also perform an aspectual transformation of the clause they modify, e.g.,

I finished staring at it. (**culminated process**)

Aspectual coercion, a second type of aspectual transformation, can take place when a clause is modified by an aspectual marker that violates an aspectual constraint [Moens and Steedman, 1988, Pustejovsky, 1991]. In this case, an alternative interpretation of the clause is inferred which satisfies the aspectual constraint. For example, the *progressive* marker is constrained to appear with an **extended event**. Therefore, if it appears with an **atomic**

event, e.g.,

He hiccupped, (**point**)

the event is *transformed* to an **extended event**, e.g.,

He was hiccupping. (**process**)

in this case with the *iterated* reading of the clause. As a second example, the *perfect* marker must appear with a **culminated event** or a **state**. Therefore, if it appears with a **non-culminated event** such as,

John ran, (**process**)

the event must be interpreted as a **culmination** such as,

John has run, (**culminated process**)

For example, if John jogs every day for an hour, John's running has an implicit completion point. Moens and Steedman describe an algorithm which predicts possible coercion amongst aspectual classes, and begins to provide constraints for the circumstances under which such coercions take place [Moens and Steedman, 1988].

2.1.7 The First Problem: Fundamental Aspect

We define *fundamental* aspectual class as being the aspectual class of a clause *before* any aspectual transformations or coercions. That is, the fundamental aspectual category is the category the clause would have if it were stripped of any and all aspectual markers that induce an aspectual transformation, as well as all components of the clause's pragmatic context that induce a transformation. Fundamental aspectual class is therefore a function of the main verb and a select group of complements, as illustrated in the previous two subsections. It is the task of detecting fundamental aspect that I address in this thesis. Previous work in linguistics handles adjuncts separately from other clausal constituents in assessing aspectual class [Pustejovsky, 1995].

An understanding system can recognize the aspectual transformations that have affected a clause only after establishing the clause's fundamental aspectual category. Linguistic models motivate the division between a module which first detects fundamental aspect and a second which detects aspectual transformations [Hwang and Schubert, 1991, Schubert and Hwang, 1990, Dorr, 1992, Passonneau, 1988, Moens and Steedman, 1988, Hitzeman et al., 1994]. In principle, it is possible for this second module to detect aspectual transformations that apply to any input clause, despite the way in which the core constituents interact to produce its fundamental aspectual class.

2.2 Linguistic Reality and Cognitive Rationale

Semantic distinctions pose a great challenge for the field of language understanding. Refining a semantic hierarchy of language is an integral step in the advancement of natural language processing. However, intuitively-established semantic distinctions often risk being ill-defined and subjective. Therefore, the challenge is to empirically validate such formalisms.

This section surveys the history, rationale, and sources of empirical evidence for aspectual distinctions. I first provide a brief history of the origins of aspect, exploring the philosophical and linguistic insights that first motivated its investigation. Next, I describe the formal definitions set forth by Dowty, illustrating them with examples and motivating them with intuition. This type of intuitive and phenomenological investigation is then further refined with an overview of philosophical issues concerning aspect, and its relationship with cognitive linguistics. Finally, I describe the various sources of empirical evidence, both linguistic and statistical, that serve to validate aspectual distinctions.

2.2.1 The History of Aspect

Aspectual distinctions were first recognized in early philosophy literature. For example, in *Metaphysics* (1048b), Aristotle noted that some verbs entail an endpoint, and that this endpoint is a sort of conclusion that results in a new circumstance. This was the first examination of completedness. More recent philosophical publications have further refined these insights. Ryle [1949] further refined our understanding of completedness, distinguishing between *activities* (non-culminated processes) and *achievements* (culminations). However, the investigation of aspectual distinctions is limited in the philosophical literature, since philosophers investigate language primarily as a means towards understanding human cognition, knowledge and belief systems. Furthermore, philosophical insights into language are often based entirely on subjective intuition.

Linguistics, however, studies language as a field unto itself, and applies more rigorous methodology towards this investigation. In particular, linguistic investigations of aspect have taken form in the following paradigms:

- Discovering and testing phenomenological constraints and entailments that correspond to aspectual class.
- Developing a generative semantic model of sentence composition that integrates surface form, aspect, and deep semantics.
- Establishing temporal models that show the influence of aspect on temporal constraints between clauses and sentences.
- Establishing formal definitions of aspectual classes in terms of temporal interval semantics.
- Discovering linguistic correlations between verb subcategorization, verb aspect, and resolving verb ambiguity.
- Statistical analysis to evaluate hypothesized aspectual distinctions, constraints and entailments.

Each of these paradigms have served both to further refine our understanding of the aspect, and to further validate the insights of early philosophy. I provide further details on the first three of these paradigms in the rest of this subsection, and detail the rest in the following few subsections.

The primary linguistic insight that bridged the gap between philosophical intuition and formal definitions was the discovery of phenomenological entailments of aspectual class. These entailments, summarized above in Section 2.1.2, provide criteria by which to manually categorize individual clauses and sentences, provide entailments that can be implemented by language understanding systems (cf. Section 2.3 on applications of aspect), and further validate our set of aspectual distinctions. Kenny [1963] took the insights of philosophers to a more formal level with linguistic observations of entailment. For example, he noted that the progressive reading of a sentence necessarily entails the perfect reading for activities, but not for achievements. For example, *I am starring at him* implies *I have starred at him*, but, *I am building a house* does not necessarily imply *I have built a house.* Kenny also first recognized the concept of states, as linguistically distinct from activities. Further linguistic entailments, as listed in Table 2.3, were then primarily fleshed out by Vendler [1967] and Dowty [1979].

Another important investigation of aspectual semantics took place in developing the *generative* lexicon [Pustejovsky, 1991]. This is a semantic model that describes the mechanics by which lexical semantics interact in a sentence to produce a semantic interpretation of the sentence. For example, *bake-bread* and *bake-potato* are distinct in that, for the former, something new is being created, but for the latter, the object is pre-formed. Informally, whether the item is natural or manufactured is reflected in semantic entries for the direct object, and the entry for *bake* is designed to accept arguments of each type. This reflects a deeper interpretation model underlying the aspectual class they share (culminated). These concepts are further formalized for the generative model [Pustejovsky, 1991]. This type of generative model applies to deriving aspectual class in several ways. For example, *I drove the car* describes a non-culminated event, but, *I drove the car ragged* entails a *transition* (i.e., culminated event), since the lexical semantics of *ragged* denote the consequent state that results from the event's culmination. As another example, *to the store*, has a similar effect when modifying *I ran.* The significance of this work to the history of aspect is twofold. First, it introduces a lexically-based approach to resolving aspectual ambiguity. Although this still leaves open the lexical-bottleneck issue, it opens a door for statistical approaches by further refining the precise type of lexical knowledge required for aspectual classification. Second, it describes a linguistically plausible semantic framework that incorporates the aspectual hierarchy with other linguistically-motivated semantic hierarchies. The underlying semantic mechanics reflect our intuitive understanding of aspectual distinctions.

As shown above in Section 2.1.3, aspectual class is predictive of certain temporal constraints, e.g., between clauses connected with *when.* This type of observation was instrumental in refining and understanding aspectual distinctions. For example, a punctual event such as, *I noticed the picture on the wall*, might be followed by the question, *At what time?* However, an extended event such as, *I walked the dog* might be followed with *For how long?* [Moens and Steedman, 1988].

Table 2.6: Equivalency of various terminology used to describe aspectual classes and distinctions.

[Vendler, 1967, Dowty, 1979]	[Moens and Steedman, 1988]	[Pustejovsky, 1991]	[Siegel, 1999]
state	state	state	state
	event	event	event
activity	process	non-telic	process
accomplishment	culminated process	transition	culminated process
achievement	culmination	transition	culmination
	point		point
		stativity	stativity
	±conseq	telicity	completedness
	atomicity	durative	atomicity

The term *aspect* was first used by Greč (1827), when investigating the Russian language, to describe semantic distinctions other than simple tense distinctions. In general, *aspect* refers to multiple viewpoints or perspectives on the same situation [Binnick, 1991]. This is reflected by the discussions in the following few subsections. The philosophical analysis of aspect, below, argues that by providing multiple ways to view the same situation, aspect provides a fundamental linguistic mechanism to represent and enable comprehension of the relationship between entities and the events that involve them. Furthermore, the linguistic discussions below define aspectual classes in terms of viewpoints, e.g., temporal granularity distinguishes between punctual and extended perspectives on a non-instantaneous event.

The terminology used to describe aspectual distinctions and classes has varied widely during the development of this field, and a standard vocabulary has not yet been established. Table 2.6 lists the various terms that have been used to describe each concept. There is often confusion between the terms *achievements* and *accomplishment*. Moens and Steedman [1988] derived a new set of terminology that is more explicit in meaning in order to resolve this confusion. I have replaced *telicity* with *completedness* to continue in this effort. Note that Pustejovsky [1991] uses *transition* to refer to *culminated events* since this reflects the generative semantics of that aspectual category.

2.2.2 Formal Definitions for Aspectual Distinctions

Formal definitions of aspectual distinctions were set forth by Dowty [1979]. These definitions are in terms of Taylor's temporal interval semantics [Taylor, 1977], and are therefore described in terms of temporal intervals, as opposed to instants:

- **State:** If a state holds over a certain time interval, it is true for all subintervals within that time interval.
- **Process:** If a process holds over a certain time interval, it is true for all sub-intervals within that time interval, but only down to a certain threshold in the size of the sub-interval.

- **Culminated event:** If a culminated event holds over a certain time interval, it is false for all proper sub-intervals within that time interval.

Such formalization serves two purposes. First, since these definitions provide necessary and sufficient conditions for aspectual classes in terms of temporal semantics, they describe a deeper level of semantic understanding that can be automatically inferred after aspectual classification. This can be used, for example, to enhance the semantic model of a lexicon. Moreover, temporal interval semantics can be used to describe the effects of temporal modifiers. For example, *She ran the mile* **in five minutes** implies that both *She ran the mile* **in 3 minutes** and *She ran the mile* **in 10 minutes** are false.

Second, these formalizations enhance our intuitive understanding of aspectual distinctions. This helps validate the aspectual distinctions by eliminating a certain degree of subjectivity. Furthermore, it can serve to guide further linguistic investigation. For example, aspectual entailments, such as those shown in Table 2.3, reflect this underlying temporal model.

In this section, I illustrate each of these definitions with examples, providing a detailed, intuitive analysis. Primarily, this is a summary of Dowty's work, but some inferences were made in order to provide a more explicit interpretation of how the formal definitions are applied.

Stativity

If *I dislike cheese* (state) holds true from time $t1$ to time $t2$, it holds true for every smaller span of time between $t1$ and $t2$. In particular, there is no span of time so small that it does not hold true. That is, even as the length of the interval approaches zero, i.e., even if "time stops", the situation, *I dislike cheese*, still holds true.

In contrast to this, events require dynamicism – change through time. Therefore, if a small enough interval in time is examined, the assertion will no longer hold true. For example, it is easy to see that *I ran a country* (event) does not necessarily hold true for all subintervals. This is because there can be subintervals during which no events transpire that are related to running a country. Another example is (9), which may not appear as clear at the first glance.

(9) *I ran for exercise* (event)

This is because, at all intermediate intervals, (10) holds true:

(10) *I was (in the process of) running.*

However, since (10) is in the progressive, it takes a stative view (a type of aspectual transformation is at work here).

But when a small enough interval is considered in isolation, it does not qualify as *running*, so (9) cannot be considered a state. In particular, if *I ran* holds true from time $t1$ to time $t2$, it holds true for many intervals within $t1$ and $t2$. However, for a small enough time span, it does not hold true. For example, when considering in isolation the events that transpired during an individual nanosecond of the *run event*, one would hardly consider it to be *running.* Likewise, if you are asleep, you can't *think about mice* (event), but you can *think that mice are small* (state).

This brings up another complication. Consider, *I had mixed feelings,* and, *It is a dynamically-colored pattern*, both of which describe states. Since the states they describe necessarily imply dynamicism, isn't there a small enough subinterval for which, when considered in isolation, the stative sentence does not hold true? What differentiates these stative sentences from an event such as (9) is that they denote non-dynamic *qualities* that provide an explanation for the dynamic events described. Since these qualities hold true for arbitrarily small intervals of time, it is possible to say that a person *has mixed feelings* (in general), even over a time span during which the person's feelings do not change; during such a time span it is true that they had the potential to change. In contrast, *Her opinion oscillated* describes an event. This exemplifies the fact that aspectual class determines which of several alternative *viewpoints* of a situation are being expressed.

Completedness

Consider the culminated event, *I climbed the mountain.* If the duration of this event spans from time $t1$ to time $t2$, then the statement is not true for any proper sub-interval between $t1$ and $t2$. In general, since culminated events entail a completion point in time, they cannot hold true for a sub-interval in time that does not include this culmination.

One test for non-culminated events, shown in Table 2.3, is that the past progressive reading necessarily entails the past tense reading. For example, *I was jogging* implies, *I jogged.* This is not true for non-culminated events; *I was climbing the mountain* does not necessarily imply that *I climbed the mountain.*

This rule is reflected by the temporal interval semantics of completedness. Since the progressive tense entails, *is in the process of...*, the culmination is not necessarily entailed. Therefore, the entire temporal interval may not be completed, in which case the statement's fundamental aspectual class (e.g., *I climbed the mountain*, with no aspectual transformation), is not necessarily entailed.

Applying this semantic rule requires careful attention to detail, since some examples may initially appear misleading. For example, *I swam laps* is non-culminated by virtue of its indefinite, plural direct obect [Vendler, 1967]. However, one might argue that *I was swimming laps (until interrupted)* does not necessarily entail *I swam laps*, since the swimming may have terminated before even one lap had been swum. However, *I was swimming laps* can only denote a true state if multiple laps were being swum.

Atomicity

Dowty's definition of atomicity is less formal: "uninterrupted." Specifically, it stipulates that although a punctual event, e.g., *I died*, may actually transpire over a non-zero-length temporal interval, there are no other narrative events, implicit or explicit, that transpire within that interval. In particular, a culminated process denotes that there is an interruptible event leading to the culmination, but a (punctual) culmination does not enable interruption.

Aspectual class allows an event to be viewed as atomic, despite its actual temporal duration. For example, *The battalian crossed the border (in 20 minutes)* denotes a culmination, despite the knowledge that it requires a non-zero interval of time [Vendler, 1967]. Temporal granularity and operators that modify it are further discussed by Moens and Steedman [1988].

2.2.3 Linguistic Philosophy, Cognition, and Aspect

A philosophical investigation of the purpose and significance of linguistically-established semantic distinctions is necessary for advancing natural language understanding. As mentioned above, philosophical literature is where aspectual distinctions first appeared. Natural language and linguistics reflect our understanding of the world and our understanding of our process of understanding. Philosophical investigations serve to expand our intuitive understanding of the formal definitions of aspect so that, instead of appearing prescribed, their operation in language is evident. This in turn illustrates the deeper significance of aspectual distinctions, and can motivate additional linguistic and statistical methodology for further investigating aspect. It helps refine the integration of linguistics and statistics in language understanding.

In this section, I further explore the philosophical issues underlying linguistic work in order to refine the relevance of aspect classification to natural language understanding, and to reveal its potential when applied statistically. First, I explore the significance of aspect for human cognition from an intuitive standpoint. Then I discuss how linguistic evidence suggests that clauses are constrained to exactly one aspectual class per clause. I explain how this insight can be exploited with statistics to bootstrap deeper natural language understanding. Finally, I illustrate a challenge for linguistics by enumerating the obstacles towards defining and justifying many semantic distinctions in natural language understanding.

Stativity as a Cognitive Tool

Stupid is as stupid does. -Forest Gump

As rational agents, humans must find some way to organize their understanding of events in the world around them. The first step in this direction is to represent the relationships between entities, both abstract and concrete, and the events that involve them.

Stativity is the cognitive mechanism to generically represent this relationship. In particular, by making a stative interpretation of a situation, the speaker is inferring a characteristic that describes an entity. This characteristic is intended to predict potential and actual events than can or have involved the entity. For example, emotional states of agents, e.g., *I am angry*, imply potential actions, and physical states of objects, e.g., *It is hot*, imply ways in which the object interacts with its environment. Even locations, e.g., *The book is on the table*, imply events, specifically, interactions of molecules at the location the book meets the tables. Our world consists of events; stativity as a concept is a cognitive tool for interpreting these events.

Stativity as an abstraction frames inferred characteristics of entities. It can represent general causative relationships between enties and events, applied in either direction. For example, we infer from *She walked* that *She was able to walk.* In this case, a state is being inferred from an event. Being able to walk is a quality that we endow entities that do in fact walk. The inference can go the opposite direction, deriving an event from a state. For example, *I can see Harry* implies, *I see Harry.*

In general, any event can be viewed with a stative interpretation, e.g., *It was in a state of oscillation,* and, *She* is able to *ride a bike.* Likewise, certain aspectual coercions are facilitated by generating an event interpretation of a state, e.g., *Be brave* (state → event). In this way, stativity is a way of indicating different viewpoints, or aspects, of a situation.

Strongly-Typed Aspectual Classification

The strong hypothesis regarding aspect is that each clause is an instance of exactly one aspectual class. This is the underlying claim of most work in linguistic that examines semantic distinctions such as aspect, e.g., the definitions of aspectual constraints. It is supported by the presence of aspectual constraints and entailments, which are empirical indicators of a clause's class. Furthermore, this claim is supported philosophically: if stativity is a mechanism to enable a non-dynamic interpretation of dynamic situations (i.e., events), a speaker must make an explicit choice between these two viewpoints when generating a clause. Adopting this hypothesis allows for the most parsimonious linguistic model.

This claim does not imply that each clause denotes an actualized event or state. Rather, the aspectual class of a clause indicates the concept described by the clause, which itself may or may not be instantiated. For example, *I did not go home* is not instantiated by virtue of being negated. As another example, *I can jump over the fence* is not necessarily instantiated.

An alternative hypothesis regarding aspect is that aspectual distinctions are on a continuum. For example, Klavans and Chodorow place verbs on a scale called "degrees of stativity" according to the frequency with which each verb occurs with certain aspectual markers such as the progressive [Klavans and Chodorow, 1992]. This way, multiple numerical factors can be integrated to derive aspectual class, or, alternative, "aspectual ranking."

Alternatively, if the strong claim is adopted, statistically-measured aspectual rankings may be indicative of the ease with which aspectual coercion takes place. This way, rankings are used together with sentential and pragmatic context to predict coercions, providing a way to compute the one resulting aspectual class of a clause.

Linguistically-Based Statistics for Cognitive Science

Linguistics as a field has provided enormous insights that describe how the thoughts behind language are reflected by the structure of sentences. For example, certain surface forms such as aspectual markers are constrained by the semantic content of a sentence. These linguistic constraints represent a system that forces the speaker to select from one of a limited number of semantic categories. Therefore, the distributions of aspectual markers across corpora reflect the semantic categories of verbs and phrases. Such insights provide an unprecedented opportunity for cognitive science.

In particular, statistical analyses of the frequencies of aspectual markers may provide an initial, bootstrapping step towards automatically building large-scale, semantic knowledge basis that are cognitively plausible. Such knowledge bases could extend beyond the small number of high-level aspectual semantic distinctions.

Although aspectual constraints are not absolute, they are informative when examined on a statistical basis. In particular, the *fundamental* aspectual class of a clause is elusive due to the presence of aspectual transformations. However, since aspectual markers are syntactic manifestations, they are measurable. Despite the sporadic presence of aspectual transformations, these measurements are likely to correlate with semantic categories since transformations are exceptions, not the rule.

It may be possible to use the results of such statistical analyses in a reverse-engineering fashion to build up pragmatic, world knowledge about certain domains. For example, if *have-fever* appears in the progressive only infrequently, this indicates that it is stative. It may then be possible to automatically deduce that *fever* describes a quality.

Semantic Distinctions: Subjective or Objective

A challenge for language interpretation in general is establishing a well-founded hierarchy of semantic distinctions, such as aspect. Since such distinctions are founded on intuition, work in semantic language classification requires more extensive justification beyond evaluating classification performance than many other applications in automated classification. For example, visual face recognition is generally thought to be a viable problem, since the objects being classified, people, are agreed to be distinct entities in the real world. However, linguistic distinctions are not as clearly agreed apon. On the other hand, perhaps distinctions between people are not as objective as often assumed. For example, false identities could result in pictures of the same person with inconsistent labels in supervised learning data.

There are numerous factors contributing to why semantic distinctions in language are not as widely accepted as viable classification problems in comparison to image classification, including:

- Image classification distinguishes between physical entities, while semantic distinctions are between abstract entities.
- The distinction between physical objects are conscious, while semantic distinctions in language are not necessarily conscious.
- Physical objects are named with well-founded terminology in natural language, but there is not yet a consensus on names for many semantic linguistic classes.

2.2.4 Empirical Support for Aspectual Distinctions

Despite its importance, providing evidence to validate semantic distinctions in language provides a major challenge for natural language processing as a field. Work in linguistics are often prone to theoretical arguments of an informal nature that can lead to specious distinctions. However, since many semantic distinctions are not conscious, it is necessary to use a well-formed regiment of linguistic and statistical methods to elicit and evaluate these distinctions. This subsection describes these methods as applied to aspect.

Linguistic Evidence

Linguistic insights that involve aspect both illustrate and support the aspectual semantic distinctions. This includes aspectual constraints (Section 2.1.1), aspectual entailments (Section 2.1.2), and temporal models that incorporate aspect, e.g., for *when* (Section 2.1.3) [Dowty, 1986].

Moreover, many natural languages have explicit *perfective* markers that indicate completedness, e.g., Bulgarian, Russian, Check, Polish, Finish and the past tense of French. Perfective markers are illustrated below in Section 2.3.1. Thus, although non-Slavic languages such as German and English do not have explicit morpho-syntactic markers such as the perfective to indicate aspectual class [Grimm, 1883], the presence of aspectual distinctions is validated by virtue of explicit syntactic manifestations in other languages.

There is a correlation between verb subcategorization, meaning and aspect which further supports the aspectual hierarchy. In particular, many semantically-related groups of verbs that share subcategorization behavior denote identical aspectual relationships [Dowty, 1979]. For example, the following verbs, which take animate subjects, denote states: *love, hate, dislike, know, have.* Likewise, the following intransitive verbs denote culminated processes, in particular, changes of physical state: *darken, warm, cool, sink, improve.* Dowty describes a large number of such verb groups [Dowty, 1979]. Furthermore, as shown above in Section 2.1.5, many statively ambiguous verbs operate similarly in each aspectual class. This provides evidence that the underlying semantic model incorporates aspectual distinctions.

Statistical Evidence

Statistical investigations can validate aspectual distinctions in several ways. For example, the frequency with which aspectual markers appear with a verb is linguistically hypothesized to correlate with aspectual class. This is because the presence of aspectual markers places constraints on aspectual class. For example, verbs that appear in the *progressive* frequently are likely to denote events. In this thesis, the aspectual class of verbs and phrases is determined by the frequency with which verbs and verb-complement combinations appear with aspectual markers, as described in Chapter 3. Note, however, that this correlation is not predicted to be 100% accurate. This is due to several factors:

- As discussed in Section 2.1.1, while the *progressive* indicator is linguistically linked to atomicity, it is only indirectly linked to completedness. Its usefulness for predicting whether a verb is **culminated** or **non-culminated** may be due to the fact that non-extended (i.e., **atomic**) verbs are more likely to be **culminated** than **extended**.

- Many ambiguous `location` verbs can appear with the progressive, even in their stative sense, e.g. *"The book was lying on the shelf."*
- 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****."*
- Aspectually ambiguous verbs conflate multiple aspectual classes.
- Infrequent verbs are measured sparsely.
- Noise to due parsing errors.

Despite these caveats, a positive correlation between marker frequency and verb class provides empirical support for aspectual distinctions. Measuring such a correlation is one of the primary investigations of this thesis; results are given in Chapters 4 and 5.

Another method to validate aspectual distinctions is to measure agreement between two linguists on the aspectual class of clauses. A high agreement would indicate that aspectual distinctions have been defined in formal enough terms as to eliminate subjectiveness. Moreover, if an aspectual constraint is used as a linguistic test to accrue each linguist's marking, this method has the additional advantage of, effectively, evaluating the aspectual constraint. For example, *what happened was...*, tests for stativity. The results of this test show 91% agreement. (cf. Section 5.1.1).

Additionally, showing highly accurate automatic prediction of aspectual class, which is covered in Chapters 4, 5 and 6, indicates that manually supervised aspectual distinctions are reflected in objective reality. If a computer automatically predicts the classes assigned by a human, this shows that the human is doing something systematic. In particular, since automatic classification is based on objective measurements (marker frequencies), and it is highly correlated with manual marking, the manual markings must reflect an external reality. That is to say, intuitively, if an automated process comes up with roughly the same set of labels as a human, the human was following a non-random system which is objectively validated. The objective data can by itself establish that there is a separation of words according to something closely related to what we call "aspect", and thus aspectual distinctions are not specious.

To look at this more formally, automatic labeling is based solely on empirical measurements that are collected in an unsupervised fashion, and is thus objective. Manual labeling follows a linguistic model which is meant to explain the objective labeling, but is carried out in a subjective manner. The existence of a correlation between the two indicates that the subjective, manual process is in fact reflected in objective reality. Moreover, this correlation supports the validity of the linguistic model as an explanation for the empirical data.

Finally, Klavans and Passonneau have conducted experiments that evaluate aspectual constraints and entailments statistically by repeated manual evaluation over small corpora

[Passonneau and Klavans, 1994]. This type of of work could be expanded. For example, many human subjects could answer questions such as the following, and compare responses (which is outside the scope of this thesis): If it is true that "X was building a house," is it necessarily true that "X built a house"? (No.) If it is true that "X was playing the piano," is it necessarily true that "X played the piano"? (Yes.)

2.3 Applications of Aspectual Classification

Aspectual classification is a necessary component for systems that perform certain natural language interpretation, natural language generation, summarization, information retrieval and machine translation tasks. For example, once a language interpretation system has identified the aspectual categories of input clauses, aspectual entailments such as those listed in Table 2.3, and temporal constraints between connected clauses, such as those described in Section 2.1.3, can be inferred. These temporal entailments and constraints are a prerequisite for inferences pertaining to many tasks involving temporal relations, e.g., detecting upward or downward trends in the stock market domain, or inferring sequences of medical procedures in medical domains (see below). On the discourse level, changes in aspect may inform the task of recognizing subdialogs or other discourse structures [Wiebe et al., 1997] (cf., e.g., Grosz and Sidner [1986] and Nakhimovsky [1988]), which in turn helps anaphora resolution.

A natural language generation system must explicitly determine the aspectual class of each clause it generates in order to adhere to aspectual constraints and to generate appropriate aspectual entailments. For example, if a system constructs a clause with an aspectual marker that has an associated aspectual constraint which the clause violates, the clause will either be unintelligible, or will undergo aspectual coercion. Aspectual coercion is desirable for certain sentences, but must be planned according to certain linguistic constraints [Moens and Steedman, 1988]. Therefore, the fundamental aspectual class of each constructed clause must be known by the generation system in order to avoid unintelligible clauses or unwanted aspectual coercion, or to select desirable aspectual coercions. Likewise, a generation system must be aspect-aware to take advantage of or to avoid unwanted results of aspectual entailments such as those shown in Table 2.3. Finally, these same conditions apply to selecting connectives such as *when* and *while*, since these place temporal constraints on the clauses they connect, depending on aspectual class.

Aspectual classification is necessary for certain text summarization tasks that involve temporal relationships. For example, consider a system that summarizes time durations. A prepositional phrase denoting a duration with *for*, e.g., *"for a minute,"* describes the duration of a state, e.g., *"She felt sick for two weeks,"* or the duration of the state that results from a culminated event, e.g., *"She left the room for a minute."* That is, correctly identifying the use of *for* depends on identifying the stativity of the clause it modifies. A language understanding system that incorrectly classifies *"She felt sick for two weeks"* as a non-culminated event will not detect that *"for two weeks"* describes the duration of the *feel*-state.

Aspectual classification is also necessary for certain tasks in information retrieval. For example, aspect can inform genre classification, since different genres have different frequencies of certain aspectual categories (Jussi Karlgren, personal communication). In particular, initial investigations show a higher proportion of **non-culminated events** in narratives than in medical reports or the Wall Street Journal. Genre identification is an integral part of certain text retrieval tasks, and could help increase parsing, POS-tagging, and word sense disambiguation accuracies [Kessler et al., 1997].

Recent work has incorporated aspect to develop natural language interfaces for temporal databases [Androutsopoulos et al., 1998]. For example, a database of airflight information describes the locations and actions of aircraft across temporal intervals. In such a domain, *cooperative responses* [Kaplan, 1982] are enabled by aspectual classification. For example, *Did BA737 taxi to gate 2?* could be responded to by, *BA737 was taxiing to gate 2, but never reached it.* In particular, the system must recognize that the query at hand denotes a culminated event.

Finally, as described below, aspect can help augment an on-line lexical database which can then be used to automatically expand queries with related concepts and terminology.

In the remainder of this section we further detail three applications that require aspectual classification: machine translation, processing medical reports, and augmenting the Medical Entities Dictionary.

2.3.1 Machine Translation

Aspectual classification is necessary for translation between certain pairs of languages because some languages have explicit *perfective* markers and others do not. A *perfective* marker is used in many languages, such as Bulgarian, Russian, Check, Polish, Finish and the past tense of French, to indicate completedness. For example, in Finish, the direct object changes to indicate completedness, so *"read book"* is translated as either *"Luin kirjaa"* (incomplete) or *"Luin Kirjan"* (complete). Perfectiveness is closely linked to *completedness*, so a system translating from a language without explicit perfective markers such as English to one with explicit perfective markers must first detect the aspectual category of an input phrase in order to determine the form of the output [Stys, 1991, Dorr, 1992].

Additionally, aspect must be considered when selecting a preposition for machine translation [Moens and Steedman, 1988]. For example,

"John arrived late at work **for** *several years."*[2]

describes a **non-culminated event**, since John's repeated arrivals do not have one culminating point in time. As a result, its translation to French is,

"**Pendant** *des années Jean est arrivé en retard au travail."*

However,

"John left the room **for** *a few minutes."*

describes a **culminated event**. Here, the preposition *for* describes the duration of the

[2]This translation example is from Moens and Steedman [1988].

state that results from the event. Therefore, the translation is,

"Jean a quitté la chambre **pour** *quelques minutes."*

Here, a difference in completedness accounts for the different preposition choice in French. German also requires such a distinction between prepositions.

Detecting *stativity* is also necessary for machine translation. For example, when translating from English to Spanish,

"She is resembling her mother more and more each day."[3] (**event**)

shows the result of coercing a **state** to an **event** by way of the *progressive* marker. The Spanish translation, however, does not need to appear in the *progressive* because *parecerse* describes an **event** [Klavans and Chodorow, 1992], as in the translation:

"Cada dia ella se parece más a su madre."

2.3.2 Processing Medical Reports

Many applications that process medical reports require aspectual classification because a patient's medical progress and history are established as a series of states and events that are temporally related. One task is to automatically complete a database entry for a patient by processing a medical report [Friedman et al., 1994, 1995], detailing, for example, a visit to the hospital. Medical discharge summaries consist of unstructured paragraphs of text that typically address the following information regarding a patient's visit to the hospital:

- history of present illness
- medical summary and symptoms
- diagnosis
- hospital course and procedure
- symptomatic and diagnostic outcome

From this information, the task is to complete the entries of a template record that formalizes information describing the visit. In particular, the record would itemize symptoms treatments, and their temporal relationships. Therefore, temporal constraints must be assessed. Moreover, each clause must be categorized as to whether it describes a symptom or treatment.

In this subsection, I illustrate how aspect could be used to distinguish, and infer temporal relationships between symptoms and treatments in medical reports in order to complete a temporal template. Then I demonstrate several specific applications of such a template.

[3] This translation example is from Klavans and Chodorow [1992].

Temporal Relationships in the Medical Domain

Distinguishing between **states** and **events** is necessary to determine temporal relationships between procedures and symptoms in the medical domain. For example, consider the temporal relationships between clauses connected with *until*, e.g.,

(14) *The right utero-ovarian ligament* **was** *coagulated* <u>*until*</u> **desiccated** *with bipolar forceps.*[4]

In this case, the *be*-**state** ends at the same time the *desiccate*-**event** ends. In the contrasting example,

(15) *Irrigation was* **performed** <u>*until*</u> *the flaunt* **was** *clear.*

the *perform*-**event** ends at the same time the *be*-**event** begins. Medical examples illustrating similar phenomena for the connective *when* were shown in Section 2.1.3.

Distinguishing between **states** and **events** is also necessary to interpret durative time modifiers. For example, in (16),

(16) *The patient had low blood pressure* **for 15 minutes**.

the modifier *"for 15 minutes"* denotes the duration of the *have*-**state**. As described in Section 2.3.1 (pertaining to machine translation), classification according to stativity is necessary to interpret durative *for*-PPs.

Distinguishing between **culminated** and **non-culminated events** is also necessary to determine such temporal relationships. For example, consider the temporal relationship between the clauses connected with *when* in,

The small bowel **became** *completely free* <u>*when*</u> *dissection was* **continued**.[5]

In this case, the *become*-**culmination** takes place at the onset of the *continue*-**process**. However, in

The small bowel became completely free <u>*when*</u> *dissection was* **performed**.

the *become*-**culmination** takes place at the completion of the *perform*-**culminated process**.

Aspect for Distinguishing between Symptoms and Treatments

Most applications that make use of a temporal template describing a hospital visit will need to distinguish procedures and treatments performed by medical providers from the symptoms and indications displayed by patients. As it turns out, this very distinction corresponds to that between events and states, respectively. In particular, action taken by a medical provider, e.g., *prescribe medication* or *perform surgery*, is necessarily dynamic. In contrast, a patient's health is generally viewed in non-dynamic terms, e.g., *have fever* or *be dead*.

The use of aspectual classification for this type of semantic distinction is elaborated

[4] These examples of *until* are from the corpus of medical discharge summaries described in Chapter 4.

[5] These examples of *when* were inspired by samples from medical discharge summaries.

below in Section 2.3.3 on the Medical Entities Dictionary.

Applying the Temporal Database

The applications for a system that automatically fills out a temporal template describing a hospital visit are manifold. For one, the temporal relationships between symptoms and treatments allow causal relationships to be inferred. For example, in (14) above, the termination of the *dessicate*-procedure corresponds with the termination of the *be-coagulated*-indication. However, in (15) above, the termination of the *perform-irrigation*-procedure corresponds with the onset of the *be-clear*-indication. Aspectual classification is an integral part of assessing these temporal relationships.

These temporal relationships can then be used to infer causal relationships between a provider's treatment, and a patient's indication. In particular, these two contrasting examples describe an indication that is apparently *terminated* by a certain medical procedure, and another that is *created* by a medical procedure. Determining this type of causal relationship is crucial for determining the success of procedures that have been applied to a patient. This can, in turn, guide further choices of treatment of that patient.

Additionally, interpreting durative time modifiers and temporal constraints are necessary for automatic patient diagnosis. In particular, such information can be used by an expert system to infer probable diagnoses of a patient [DuMouchel et al., 1995, Hripcsak et al., 1995]. For example, from (16) above, a diagnosis system could derive that the patient had been hypertensive, since this is an appropriate diagnosis for a patient that has low blood pressure for an extended period of time. Furthermore, temporal relationships between drug prescriptions and dosage frequencies would make it possible to automatically query for patients on certain drugs, e.g., when a certain product or dosage is recalled. Moreover, potentially fatal interactions between temporally overlapping drug prescriptions could be automatically detected.

By accruing a database of hospital visit over multiple patients, preliminary medical research could be automated. For example, the success rates of medical treatments for certain symptoms or diagnoses could be examined statistically over a large number of medical reports in order to provide clues for more in-depth medical studies. This could significantly expand medical research, since currently such data is only available when a process for data collection is planned and implemented for a particular procedure.

Finally, since patient indications are distinguished from provider treatments, a query system to retrieve only indications or treatments in particular can be implemented.

2.3.3 Medical Entities Dictionary

Distinguishing clauses according to aspect could help automatically augment the Medical Entities Dictionary (MED). MED, developed at the Columbia Presbyterian Medical Center, is a terminological semantic network that categorizes medical terminology and concepts [Cimino et al., 1994]. A database such as MED has many applications in the medical domain, such as informing medical query systems (e.g., query expansion with related terms), and interpreting medical reports.

One developmental effort for MED is its automatic expansion [Cimino and Clayton, 1994]. The high-level of MED's semantic hierarchy includes the distinction between procedure and diagnosis. This distinction maps to the division between **events** and **states**. Therefore, aspectual categorization of verbs could inform the automatic expansion of MED. Furthermore, this procedure could be specialized for additional distinctions supported by MED. For example, since aspectual categorization further distinguishes **events** according to completedness, aspectual classification may also help automatically distinguish between medical procedures that have a specified completion, such as surgical procedures, and other treatments that are open-ended, such as drug or radioactive therapies. Additionally, corpus-based techniques specialized for distinctions in medical terminology could be developed, based on our approach to aspectual classification. For example, lexical co-occurrence with certain medical terms such as *perform* and *relapse* may help distinguish medical terminology according to the hierarchy implemented by MED.

2.4 Conclusions

The aspectual class of a clause reflects the underlying temporal structure of the event(s) described by the clause. Aspectual classification is a necessary prerequisite for interpreting certain adverbial adjuncts, as well as identifying temporal constraints between sentences in a discourse [Moens and Steedman, 1988, Dorr, 1992, Klavans, 1994]. It is also crucial for lexical choice and tense selection in machine translation [Moens and Steedman, 1988, Klavans and Chodorow, 1992, Klavans, 1994, Dorr, 1992]. Aspectual classification is a necessary component for applications that perform certain natural language interpretation, natural language generation, summarization, information retrieval and machine translation tasks. In particular, we have shown the relevance of aspectual classification to several medical applications.

Certain features of a clause known as *aspectual markers* are constrained by and contribute to the aspectual class of the clause [Vendler, 1967, Dowty, 1979, Pustejovsky, 1991, Passonneau, 1988, Klavans, 1994]. In addition, many temporal entailments result from the aspectual category of a clause. These constraints and entailments linguistically motivate the aspectual hierarchy of five categories. Furthermore, they are directly applicable as constraints in language understanding and generation systems.

Models that incorporate aspect rely on the ability to automatically find the aspectual category of a clause. Several researchers have developed models that incorporate aspectual class to assess temporal constraints and interpret temporal adverbials [Hwang and Schubert, 1991, Schubert and Hwang, 1990, Dorr, 1992, Passonneau, 1988, Moens and Steedman, 1988, Hitzeman et al., 1994]. For example, several aspectual markers have been noted to transform the aspectual class of the clause they modify. *Aspectual coercion* can take place when a clause is modified by an aspectual marker that violates an aspectual constraint [Moens and Steedman, 1988, Pustejovsky, 1991]. In this case, an alternative interpretation of the clause is inferred which satisfies the aspectual constraint.

An understanding system can recognize the aspectual transformations that have affected a clause only after establishing the clause's *fundamental* aspectual category, which is a function of the main verb and several other constituents, as illustrated in the previous two sections. Linguistic models motivate the division between a module which first detects fundamental aspect and a second which detects aspectual transformations [Hwang and Schubert, 1991, Schubert and Hwang, 1990, Dorr, 1992, Passonneau, 1988, Moens and Steedman, 1988, Hitzeman et al., 1994].

The number of ways in which clausal constituents interactively influence aspectual class is unknown. However, syntax alone is not sufficient, and the lexical head of multiple constituents (e.g., the verb phrase and the direct object) can be influential factors.

The presence of a linguistic marker in a particular clause indicates a constraint on the aspectual class of the clause, but the absence thereof does not place any constraint. This makes it difficult for a system to aspectually classify a clause based on the presence or absence of a marker. Therefore, these linguistic constraints are best exploited by a system that measures their frequencies. This is the approach detailed in the following chapter.

Six semantic categories of statively ambiguous verbs were derived: `communication`, `cognition`, `perception`, `location`, *metaphorical* and *carrier* verbs. Establishing a semantic hierarchy of aspectually ambiguous verbs will help predict a verb's subcategorization frame because of the correlation between subcategorization frame and semantic class.

Chapter 3

Linguistic Indicators for Aspectual Classification

Two things are required to automatically find the fundamental aspectual class of a clause:

- Specialized lexical knowledge describing open class constituents of the clause such as the main verb and the direct object
- A method to detect the interactions between these and other constituents

Many clauses can be categorized based on their main verb alone, as illustrated in the following 2 chapters. Therefore, aspectually categorizing verbs is a first step towards classifying clauses. Statistical approaches to language developed in recent years use lexical cooccurrence to group or rank words according to semantic distinctions in general [Hatzivassiloglou and McKeown, 1993a, Lafferty and Mercer, forthcoming, Pereira et al., 1993a, Lankhorst and Moddemeijer, 1992, Schűtze, 1992], and aspectual class in particular [Klavans and Chodorow, 1992]. This way, since the primary category of a verb is often domain-dependent, a specialized analysis can be performed for each domain.

Naturally occurring text contains vast amounts of implicit information pertaining to aspectual classification. Since each aspectual marker described in the previous chapter constrains the category of clauses it appears with, the occurrences and distributions of these markers may reflect the categories of the verbs and phrases they appear with.

However, the best way to make use of these markers is not obvious. In general, while the presence of a linguistic marker in a particular clause indicates a constraint on the aspectual class of the clause, the absence thereof does not place any constraint. Furthermore, aspectual transformations from a clause's fundamental category can allow the clause to occur with a constrained marker that appears to conflict with the fundamental aspectual class of the clause. These factors makes it difficult for a system to detect the fundamental aspectual class of a clause based soly on the presence or absence of aspectual markers.

Therefore, as with most statistical methods for natural language, the linguistic constraints associated with markers are best exploited by a system that measures cooccurrence frequencies. In particular, we measure the frequencies of aspectual markers across verbs, or across *multiple* clausal constituents. This way, the aspectual *tendencies* of verbs and phrases are measured. These tendencies are likely to correlate with fundamental aspectual class. For example, since we know that clauses modified by a duration *in*-PP are constrained to signify a **culminated process**, e.g., *"I ate my lunch* **in ten minutes**,*"* verbs that occur more frequently with this type of modifier are more likely to denote **culminated processes**.

The development of a method to measure and combine the frequencies of markers achieves the general benefits of statistical corpus analysis: it can be applied over multiple domains and languages, and can be extended to multiple semantic distinctions. Furthermore, once a system has been trained to distinguish by indicator value, it can automatically classify all verbs appearing in a corpus, including "unseen" verbs that were not included in the supervised sample. Also, the tradeoff between recall and precision can be selected for particular applications by selecting an appropriate classification threshold. These benefits of numerical indicators are illustrated with our results and analysis.

We will call the cooccurrence frequency of a linguistic marker a *linguistic indicator*. There are several advantages to classifying verbs with *linguistic* indicators in particular, which measure the frequencies of linguistic phenomena such as aspectual markers, rather than indicators that measure cooccurrence with open class lexical items such as the head of the direct object:

- The difficulty with sparsity for measuring the cooccurrence frequency between two open class lexical items (e.g., most verb-object combinations are rare) is alleviated, since many markers such as the *perfect* are prevalent across all verbs.
- The linguistic constraints associated with aspectual markers are exploited without the a-priori need to detect aspectual coercion. This solves a "bootstrapping" problem, since you need the fundamental aspectual category of a clause to detect any coercion affected by an aspectual marker, as discussed in Section 2.1.7.
- Linguistic markers are related to particular semantic distinctions such as aspectual categorization, whereas the semantic distinctions reflected by lexical cooccurrence are less specific, and may correspond to the least common denomenator over a broad set of semantic distinctions.
- Linguistic indicators constrained with respect to one particular aspectual distinction may also correlate with other aspectual distinctions, since aspectual distinctions are likely to be mutually correlated. For example, since **points** (e.g., *hiccup*) are rare, non-extended (i.e., **atomic**) **events** are likely to be **culminations**. Therefore, an indicator that discriminates **events** according to atomicity is likely to also discriminate between **culminated** and **non-culminated events**.
- Results with linguistic indicators may be less domain-dependent, since the linguistic constraints associated with markers relate to generally-applicable semantic distinctions such as aspectual classification. On the other hand, the relationship between

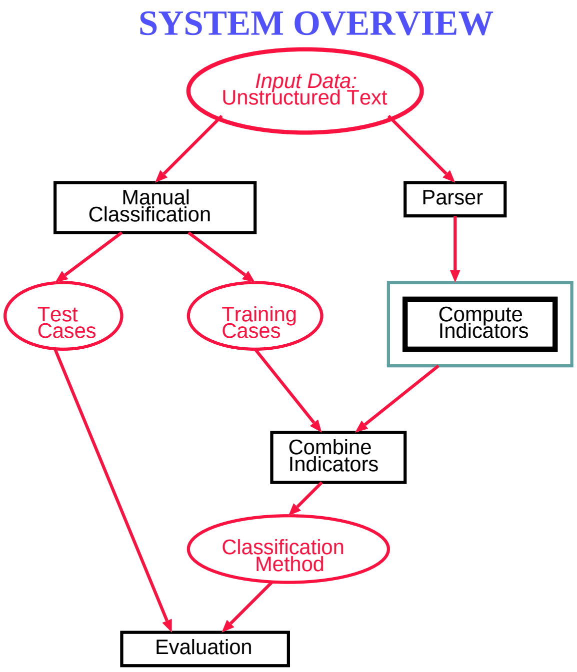


Figure 3.1: System overview for aspectual classification.

aspectual class and the frequencies of open class lexical items is likely to vary between domains.

- This framework is expandable:
  - Additional indicators can be introduced by measuring the frequencies of additional aspectual markers.
  - Indicators can also be measured over clausal constituents other than the main verb to alleviate verb ambiguity and sparsity, as described below.
  - Alternate machine learning methods may further improve the process of combining multiple indicators, as described in the "Future Work" portion of this thesis (Section 7.6).

This chapter describes a set of 14 linguistic indicators and their use for detecting aspectual class. The outline of this chapter roughly corresponds with the rightmost path of the system overview shown in Figure 3.1. First we describe our use of the English Slot Grammar to syntactically parse a corpus. Then we describe the set of linguistic indicators measured and evaluated for aspectual classification. Next we describe learning methods tested to combine multiple indicators for an improvement in classification performance; these methods are described in further detail in Chapter 6, "Machine Learning for Combining Linguistic Indicators." Then we describe methods to integrate clausal constituents other than the main verb into our system. Next, we describe various measures we use to

```
Number of parses = 2.
0.25 : phrase([1,6|8], s(return,1), verb(fin([X1|sg],past,X2)),
[slot(subj(n),ob,subj(n):5,t),slot(obj(n),op,X3,X4),
 slot(comp(p(to)),op,comp(p(to)):8,t)], ,
mods(
   subj(n) : phrase([1,5|5], s(count,3), noun(cn,[nom|sg],nwh),
   [slot(nobj(n),op,X5,X6)],     ,
   mods(
      ndet : phrase([1,1|1], s(his,1), det(sg,poss), , , mods([],[]))
      .
      nadj : phrase([2,2|2], s(white,1), adj(X7,nwh), , , mods([],[]))
      .
      nnoun : phrase([3,3|3], s(blood,1), noun(cn,[X8|sg],nwh),
      [slot(nobj(n),op,X9,X10)], , mods([],[]))
      .
      nnoun : phrase([4,4|4], s(cell,1), noun(cn,[X11|sg],nwh), , , mods([],[]))
      .
   ,)).
,
   comp(p(to)) : phrase([7,7|8], s(to,1), prep(to,X12,8),
   [slot(objprep(sg), ,
   mods(,
      objprep(sg) : phrase([8,8|8], s(normal,2), noun(cn,[acc|sg],X12),
      [slot(nobj(p(for | to)),op,X13,X14)], ,
      mods([],[]))
      .))
   .))))
```

Figure 3.2: ESG parse of, *"His white blood cell count returned to normal."*, shown in Prolog format.

evaluate classification performance. This system overview diagram is revisited in further detail in Chapter 4.

In the final two sections of this chapter, we discuss previous work in statistical language classification and summarize our approach.

3.1 Grammatical Parsing

The automatic identification of individual constituents within a clause is necessary to detect the presence of aspectual markers, to identify the main verb of each clause, and to identify other constituents that will be used to classify a clause such as the direct object. The English Slot Grammar (ESG) [McCord, 1990] has previously been used on corpora to accumulate aspectual data [Klavans and Chodorow, 1992]. ESG is particularly attractive for this task since its output describes a clause's deep roles, detecting, for example, the deep subject and object of a passivized phrase.

For each sentence ESG produces a parse structure. For example, the parse of *"His white blood cell count returned to normal."* is shown in Figure 3.2. This parse is hierarchical; each clause in the sentence is headed by a `phrase()` with a `verb()`, as seen at the top of the example parse shown. For the purposes of evaluating our approach we then produce, for each clause, a flat structure of values corresponding to a list of 25 features, shown in Table 3.1. For example, the values for *"His white blood cell count returned to normal"* are:

Verb this clause is a complement of
Main Verb
Tense
Perfect (*yes* or *no*)
Progressive (*yes* or *no*)
Subject head
Subject Determiner
Subject Number
Object
Object Determiner
Object Number
Indirect object head
Indirect object Determiner
Indirect object Number
Adverb
Predicate Adjective
Particle
Complement Preposition
Complement Prepositional phrase Object head
Complement Prepositional phrase Object Determiner
Complement Prepositional Object Number
Adjunct Preposition
Adjunct Prepositional phrase Object head
Adjunct Prepositional phrase Object Determiner
Adjunct Prepositional Object Number

Table 3.1: Constituents extracted for each clause from ESG's parses.

Main Verb:	*return*
Tense:	*past*
Perfect:	*no*
Progressive:	*no*
Subject head:	*count*
Subject Determiner:	*possessive*
Subject Number:	*singular*
Complement Preposition:	*to*
Complement Prepositional phrase Object head:	*normal*
Complement Prepositional Object Number:	*singular*

All other values are empty for this sentence.

Since many sentences contain multiple clauses, our procedure recursively scans ESG's parse tree to produce this flat structure for each phrase headed by a main verb; intermediate locations in the structure corresponding to support verbs such as *will, can* and *must*, are bypassed. However, *have*, when used as a support verb, is recorded in order to mark subordinate phrases as being in the perfect tense.

As an example, the sentence,

"At this time she is in the hospital awaiting an AV graft to be placed by the vascular surgery team."

contains three clauses:

Clause 1:	
Main Verb:	*place*
Tense:	*infinite*
Perfect:	*no*
Progressive:	*no*
Subject head:	*team*
Subject Determiner:	*definite*
Subject Number:	*singular*

Clause 2:	
Main Verb:	*await*
Tense:	*past/present participle*
Perfect:	*no*
Progressive:	*no*
Object:	*graft*
Object Determiner:	*indefinite*
Object Number:	*singular*

Clause 3:	
Main Verb:	*be*
Tense:	*present*
Perfect:	*no*
Progressive:	*no*
Subject head:	*she*
Subject Number:	*singular*
Predicate Adjective:	*hospital*
Adjunct Preposition:	*at*
Adjunct Prepositional phrase Object head:	*time*
Adjunct Prepositional phrase Object Determiner:	*definite*
Adjunct Prepositional Object Number:	*singular*

In order to keep the size of this flat structure constant, only the first of a series of conjoined constituents (other than clauses) is encoded. For example, the predicate adjective of *"She is smart, tall and fastidious"* would be *smart*.

3.2 Linguistic Indicators

Our goal is to exploit the aspectual constraints associated with aspectual markers such as those described in the previous chapter by counting their frequencies in a corpus. For example, it is likely that event verbs will occur more frequently in the progressive than state verbs, since the progressive is constrained to occur with event verbs. Therefore, the frequency with which a verb occurs in the progressive *indicates* whether it is an event or stative verb.

However, aspectual transformations and other linguistic anomalies introduce noise in this scheme. Since a clause's aspectual class can be coerced from its fundamental class, the constraints implied by the existence of an aspectual marker do not necessarily hold for the clause's fundamental aspectual class. For example, although a durative *in*-PP is constrained to modify **culminated events**, the event it modifies can be coerced from a *state*, as in,

When she started exercising, she felt better **within ten minutes**.
(**state → culminated event**)

Additional examples illustrating transformations for the *progressive* and for certain adverbs are given below.

Furthermore, certain linguistic combinations allow these constraints to be overridden. For example, although the perfect tense is generally constrained to appear with **culminated events** or **states**, any relative clause can appear in the perfect, as in,

(This is the place where) he has thought about it. (**process**)

These problems can be alleviated by measuring the *tendencies* and *correlations* between occurrences of these markers and the aspectual category of the verbs they occur with. That is, although there are exceptions to the constraints, we hypothesize that marker frequency will correlate with the aspectual distinction that they constrain.

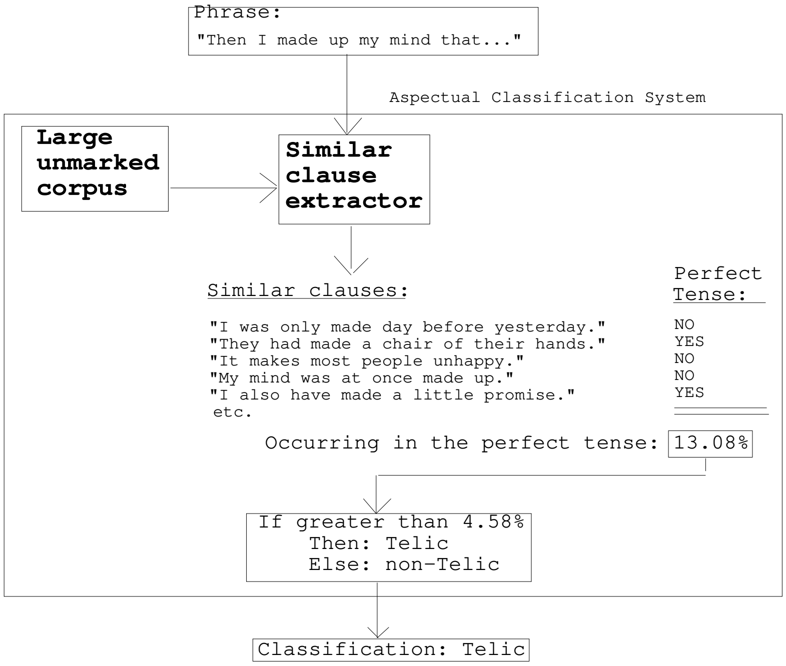


Figure 3.3: Aspectually classifying a clause with an individual linguistic indicator. The input clause is classified according to the frequency of an aspectual marker (the *perfect* tense) among clauses with the same verb.

3.2.1 Classification with Individual Indicators

An individual indicator can be used to classify verbs by simply establishing a threshold; if a verb's indicator value is below the threshold, it is assigned one class, otherwise it is assigned the alternative class. Figure 3.3 illustrates the approach. The clause being input to the system, *"Then I made up my mind that...,"* is shown at the top. This is fed directly to a module which extracts a set of *similar* clauses from an unmarked corpus. In this example, the similar clauses all have the same verb as the input clause, *make*. The frequency with which these clauses occur with the *aspectual marker* (the perfect tense) comes to 13.08%. This is then compared to the a threshold, 4.58%, and the input clause is classified as **Telic** (i.e., **culminated**).

Section 3.3 describes how all 14 indicators can be combined to improve classification performance.

3.2.2 Linguistically Motivated Indicators

The first column of Table 3.2 lists the 14 linguistic indicators evaluated to classify verbs. Each indicator has a unique value for each verb. For example, the first indicator, `frequency`, is simply the 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

Linguistic 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 to the hospital.*
past/pres participle	*...blood pressure going up.*
duration *in*-PP	*She built it* **in an hour**.
perfect	*Thay* **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 3.2: Fourteen linguistic indicators evaluated for aspectual classification.

the linguistic marker indicated. 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., passivized phrases such as, *"She was admitted to the hospital"*).

Eight of these indicators measure linguistic phenomena with corresponding linguistic constraints, i.e., aspectual markers: *perfect*, *progressive*, duration *in*-PP, no subject, and the four adverb groups. Details regarding these eight indicators are described below, and their corresponding constraints are illustrated. Other aspectual markers listed in the previous chapter present difficulties for implementation at this stage. For example, clauses with an indefinite implicit object, e.g. *"We just ate ε,"* are constrained to be **events** [Olsen and Resnik, 1997, Resnik, 1996]. However, such a clause can only be detected with the help of a subcategorization lexicon. Such lexicons exist (e.g., COMLEX, and ESG's lexicon), but require another stage of implementation nonetheless. Additionally, *force* and *persuade* constrain their complement to be an **event**. However, these words appeared very infrequently with clausal complements in our corpora.

Progressive Voice

The *progressive* marker is constrained to appear with **extended events**, e.g.,

She was counting sheep. (**process**)

but not,

She was recognizing her little lamb.* (culmination**)

This frequency of this marker was first measured by Klavans and Chodorow [1992] to place verbs on a scale called "degrees of stativity".

afternoon, attosecond, century, day, eve, evening, femtosecond, fortnight, generation, half-century, half-hour, hour, hr, instant, lifespan, lifetime, microsecond, millennium, millisecond, min, minute, moment, month, morn, morning, msec, nanosecond, night, period, picosecond, quarter-century, quarter-hour, s, season, sec, second, semester, shift, spell, term, trimester, week, weekend, year

Table 3.3: Nouns that describe durations used to detect durative *in*- and *for*-PPs.

The *progressive* marker is not entirely reliable as a predictor of aspectual class. For example, many ambiguous `location` verbs, described in Section 2.1.5, can appear with the progressive, even in their stative sense, e.g.,

The book is lying on the table. (**state**)

Also, aspectual coercion such as *iteration* can allow a **punctual event** to appear with the progressive, e.g.

She was sneezing for a week.
(**point** → **process** → **culminated process**)

Perfect Tense

The perfect tense is constrained to appear with **culminated events** or **states**, e.g.,

I have reached the top. (**culmination**)

but not,

I have stared at clouds.* (process**)

However, any clause in the progressive can appear in the perfect, even if it is **non-culminated**, e.g.,

"I have been running." (**process**)

Likewise, any clause with the main verb modified by *not* or *never* can appear in the perfect, e.g.,

"I have never run." (**process**)

Therefore, when distinguishing between **culminated** and **non-culminated events**, as in all of Chapter 5 and parts of Chapter 6, only occurrences of the perfect tense that appear in clauses that are not in the progressive tense and in which the main verb is not modified by *not* or *never* are counted. However, when distinguishing between **events** and **states**, all occurrences of the perfect are counted, since this linguistic constraint is not predicted to distinguish according to stativity in particular.

Duration *in*-Prepositional Phrase

Clauses can only be modified by a duration *in*-PP if they describe a **culminated event**, e.g.,

The computer figured it out **in a millisecond**. (**culminated process**)

but not,

The computer hung* **in a millisecond. (**process**)

To detect these durative modifiers, Wordnet was used in a partially-automatic form to derive the list of nouns shown in Table 3.3. These nouns come primarily from the set of nouns that Wordnet lists as hyponyms of the concepts "time period" and "time unit."

An *in*-PP with a definite object may connote a time location instead of a time duration, e.g.,

He had a fever **in the afternoon**. (**state**)

Therefore, we do not count PP's with an object that has a definite determiner.

Similar prepositional phrases with *within* or *over* can also connote a duration, e.g.,

Technology improved tenfold **over a year**. (**culminated process**)

Therefore, we include *within* and *over* as well as *in*.

Finally, since ESG's parse differentiates between complement and adjunct prepositional phrases, we look for durative *in*-PP's in either slot.

No Deep Subject

Both passivized phrases and imperatives are constrained to appear with **events**, e.g.,

The building was destroyed. (**culminated process**)

and,

Eat that! (**culminated process**)

but not,

Her mother was resembled by her.* (state**)

nor,

Resemble your mother!* (state**)

To estimate the occurrences of passivized or imperative clauses, we measure the frequency of clauses with no deep subject.

Adverb Groups

We have identified four semantic groups of adverbs with corresponding aspectual constraints on clauses they modify. These groups are shown in Table 3.4. These groups were identified

`temporal`: (**culminated events**)
again, already, always, annually, at, at, daily, finally, first, frequently, hebdomadally, hourly, immediately, just, monthly, never, oft, oft-times, often, oftenly, oftentimes, once, repeatedly, suddenly, then, thrice, twice, weekly, yearly
`continuous`: (**processes**)
assiduously, boundlessly, ceaselessly, constantly, continually, continuously, endlessly, endlessly, eternally, everlastingly, evermore, forever, immeasurably, incessantly, indefinitely, infinitely, interminably, perennially, perpetually, perseveringly, persistently, relentlessly, unceasingly, unendingly, unrelentingly
`manner`: (**extended events**)
accurately, attentively, cautiously, conscientiously, curiously, delicately, diligently, extensively, fastidiously, furiously, impatiently, intently, mindfully, patiently, protectively, steadily, studiously, thoughtfully, vigilantly, watchfully
`evaluation`: (**culminated process**)
abominably, abysmally, alright, atrociously, awfully, badly, befittingly, better, dreadfully, fitly, fittingly, horribly, inescapably, inevitablely, nicely, painstakingly, poorly, precisely, rottenly, satisfactorily, scrupulously, suitablely, terribly, unavoidablely, well

Table 3.4: Four groups of adverbs, and their constraints on **events**.

by consulting Wordnet, a thesaurus, and examining frequent adverbs from our corpora.

Temporal adverbs. The first group, `temporal`, consists largely of *iteration* adverbs (i.e. cardinal, frequency or cyclic adverbs [Hwang and Schubert, 1991]) that transform a clause by creating a repeated or iterated reading, e.g.,

John arrived late **repeatedly**. (**culmination → process**)

Since an event must be completed before it is repeated, these adverbs are constrained to modify **culminated events**, e.g.,

John glanced at his watch **repeatedly**. (**culmination**)

but not,

John starred at the president* **repeatedly. (**process**)

However, coercion can take place before the iteration transformation, allowing an iteration adverb to modify a **non-culminated process**, e.g.,

John thought about doughnuts **oft-times**.
(**process → culminated process → process**)

The `temporal` group also includes other types of adverbs that constrain the clause to connote a completed event, e.g., *already*, *just* and *then*. As shown in the next chapter, *then* dominates this group since it is much more frequent that the other adverbs over our corpus.

Continuous adverbs. Adverbs in the `continuous` group are constrained to modify (non-culminated) **processes**, e.g.,

She starred at the clouds **incessantly**. (**process**)

but not,

She built the house* **incessantly. (**culminated process**)

However, coercion via iteration can allow, e.g.,

She sneezed **continuously**. (**point → process**)

as pointed out by Moens and Steedman [1988].

Manner adverbs. Adverbs in the `manner` group are constrained to modify **extended events**. These adverbs describe the quality with which something is *intentionally* done [Klavans and Chodorow, 1992, Klavans, 1994]. What's more, these qualities can only apply to a process with time duration, i.e. **extended**, e.g.,

She built the house **diligently**. (**culminated process**)

but not,

She reached the top* **diligently. (**culminated**)

nor,

She resembled her mother* **vigilantly. (**state**)

Evaluation adverbs. Adverbs in the `evaluation` group are constrained to modify **culminated processes**. These adverbs evaluate the quality with which something is completed.

Like `manner` adverbs, the quality can only apply to a process with a time duration. For example,

She performed **abominably** *on the test.* (**culminated process**)

but not,

She reached the top* **horribly. (**culmination**)

nor,

She appears fatigued* **horribly. (**state**)

3.3 Combining Indicators with Learning

Given a verb and its 14 indicator values, our goal is to use all 14 values in combination to classify the verb as a state or an event. Once a function for combining indicator values has been established, previously unobserved verbs can be automatically classified according to their indicator values. This section briefly describes three machine learning methods employed to this end. Further details are provided in Chapter 6, "Machine Learning for Combining Linguistic Indicators."

Log-linear regression. As suggested by Klavans and Chodorow [1992], a weighted sum of multiple indicators that results in one "overall" indicator may provide an increase in classification performance. This method embodies the intuition that each indicator correlates with the probability that a verb describes an event or state, but that each indicator has its own unique scale, and so must be weighted accordingly. One method to determine these weights is log-linear regression [Santner and Duffy, 1989].

Genetic programming. An alternative to avoid the limitations of a linear combination is to generate a non-linear function tree that combines multiple indicators. A popular method for generating such function trees is a genetic algorithm [Holland, 1975, Goldberg, 1989]. The use of genetic algorithms to generate function trees [Cramer, 1985, Koza, 1992] is frequently called genetic programming. A threshold must be selected for both linear and function tree combinations of indicators. This way, overall outputs can be discriminated such that classification performance is maximized.

Decision trees. Another method capable of modeling non-linear relationships between indicators is a decision tree. Each internal node of a decision tree is a choice point, dividing an individual indicator into ranges of possible values and each leaf node is labeled with a classification (state or event). The most popular method of decision tree induction, employed here, is recursive partitioning [Quinlan, 1986, Breiman et al., 1984], which expands the tree from top to bottom.

3.4 Incorporating Multiple Clausal Constituents

There are two potential advantages to considering multiple constituents of a clause when performing aspectual classification. First, because of verb ambiguity, considering multiple constituents when classifying a clause may improve classification accuracy. And second, sparsity problems arise when a verb is very infrequent. For such a verb, it is difficult to accurately measure the frequency of linguistic markers. Therefore, it may improve accuracy to categorize a clause with a rare verb by measuring indicators over other constituents that correlate with aspectual class.

Wordnet can help model the aspectual composition of a clause. The ways in which clausal constituents combine to produce fundamental aspectual class are difficult to ascertain, and probably rely on a complex array of lexical features of open class words such as the main verb and direct object [Pustejovsky, 1991, Resnik, 1996]. One way to model this is to find semantic categories of constituents such as the direct object, and categorize ambiguous verbs based on this category. We have applied this method to distinguish instances of *have* based on a semantic category assigned to the direct object by the Wordnet database. This approach is evaluated in the next chapter in Section 4.3. Wordnet is described in the first subsection below.

As a contrasting approach, measuring linguistic indicators over word combinations such as verb-object has the advantage that the interactions between lexical items are measured statistically in lieu of a model that predicts these interactions. This approach is described in the second subsection below.

3.4.1 Incorporating Wordnet Classes

As suggested by Rebecca Passonneau (personal communication), Wordnet, an on-line semantic database of words, can be used to semantically classify clausal constituents such as the direct object to aid in aspectual classification. We have evaluated methods to incorporate the Wordnet category of the direct object to improve the performance of automatic aspectual classification.

Wordnet groups nouns into 25 high-level categories at the top of its semantic hierarchy (these categories are listed in Table 4.8 in the next chapter) [Miller et al., 1993]. Many nouns have multiple entries corresponding to multiple senses. For the purposes of this evaluation, we take the first category listed and discard the rest. Pronouns such as *him* and *it* are assigned their own category, `pronoun`.

3.4.2 Measuring Indicators over Multiple Constituents

Measuring indicators over combinations of constituents is important for highly ambiguous verbs. Indicators measured over an ambiguous verb will simply find the average frequency of aspectual markers over instances of the verb with both aspectual categories. In this case, the indicator value will not discriminate well. However, when incorporating additional constituents, the indicator will be measured over a subset of instances that bear a similarity to the input clause. This way, it is more likely that the indicator will be measured over instances with the same aspectual category as the input clause.

The approach is to aspectually classify clauses by the frequency of aspectual *markers* among *similar* clauses. *Clausal similarity* is a heuristic approach to grouping together clauses with a common fundamental aspectual class. The basic assumption is that clauses with certain features in common will tend to have the same fundamental aspectual class. We have previously evaluated this approach to distinguish between **culminated** and **non-culminated events** [Siegel, 1997]. The extraction of similar sentences from a corpus is also used in example-based machine translation [Nagao, 1984]; the translation of a sentence is guided by the translation of a similar sentence in a bilingual corpus. Additionally, the benefits of clustering data into equivalence classes is further discussed by Ueberla [Ueberla, 1994].

The Wordnet category of an input clause can alleviate sparsity difficulties for combinations of infrequent words. For example, extracting verb-object combinations from a corpus can lead to the difficulty than many combinations will not occur frequently enough to accurately measure linguistic indicators. In this case it may be advantageous to incorporate wordnet categories by measuring indicators over combinations of verbs and the Wordnet category of the direct object.

3.5 Evaluation

This thesis evaluates the ability to detect the fundamental aspectual class of clauses using linguistically-based numerical indicators. To evaluate such indicators over an unrestricted set of verbs in a corpus, clauses for evaluations are selected uniformly across the corpus. Over these clauses, we first test for positive correlations between individual indicators and aspectual class with t-tests. Then, indicators are evaluated for classification performance, both individually and in combination.

As described below in the first subsection, evaluation is performed over sets of unique verbs, and over sets of verb instances (i.e., clauses). The second and third subsections below describe two measures other than accuracy for evaluating classification performance beyond accuracy – F-measure and recall tradeoff.

3.5.1 Verb Sets and Instances

Classification performance in this thesis is primarily evaluated over verb instances, i.e. clauses in which the verb appears as the main verb. Because of this, the same verb may appear multiple times in the training and testing sets. This measure is beneficial in several ways:

- Measured classification performance reflects the true distribution of the verbs – some are more frequent than others.
- Ambiguous verbs may appear with multiple aspectual categories, reflecting the true distribution of the data.
- Clausal constituents other than the verb can be incorporated to help resolve ambiguity and alleviate verb sparsity.

However, we also evaluate our approach over sets of unique verbs. Training and testing across verb sets complements our evaluation of the linguistic indicators in several ways:

- Fewer supervised examples must be generated.
- Non-intersecting training and test verb sets are possible.
- Both training and testing may be more domain-independent, since verb frequencies vary between domains.
- Resulting lexicon can be evaluated over word sets.

3.5.2 F-Measure

Classification performance can be designed and evaluated according to a variety of objectives. Classification is most frequently optimized for overall classification accuracy, which is an important criterion for many applications. However, some applications weigh certain classes more heavily in which case other optimization criteria are more appropriate [Brodley, 1995, Cardie and Howe, 1997]. One example is the F-measure, computing a tradeoff between the recall and precision of a designated class [Van Rijsbergen, 1979]. In this work, we weigh recall and precision equally, in which case F-measure is simply defined as,

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

that is, recall times precision, divided by the average of recall and precision.

3.5.3 Favorable Recall Tradeoff

A skewed distribution presents a difficulty for evaluating classification performance. When evaluating over a two class classification problem in which one class dominates, there is a simple accuracy baseline method accomplished by simply categorizing all instances as the dominant class. This baseline will be greater than 50%, and therefore leaves a smaller window for potential improvement in accuracy. In this case, discriminators for this classification problem may not receive the credit they deserve when evaluated simply for classification accuracy.

An alternative method which alleviates this problem is to measure what tradeoff is achieved between the recalls of the dominant and non-dominant (i.e., majority and minority) categories, with no loss in accuracy [Cardie and Howe, 1997]. Such a *favorable recall tradeoff* is non-trivial if a non-zero recall for the non-dominant class is achieved with no loss in accuracy. This tradeoff is not possible for an uninformed approach such as that achieving the accuracy baseline. This is because an uninformed approach that classifies any non-dominant instances correctly can do so only by misclassifying a proportional amount of dominant instances, thus decreasing accuracy.

Both an improvement in F-measure and a favorable recall tradeoff present advantages for applications that weigh the identification of non-dominant (i.e., minority) instances, e.g., **non-culminated** clauses, more heavily. For example, a prepositional phrase denoting a duration with *for*, e.g., *"for a minute,"* describes the duration of a non-culminated event, e.g., *"She ran for a minute,"* or the duration of the state that results from a culminated event, e.g., *"She left the room for a minute."* That is, correctly identifying the use of *for* depends on identifying the aspectual class of the clause it modifies. A language understanding system that incorrectly classifies *"She ran for a minute"* as **culminated** will not detect that *"for a minute"* describes the duration of the *run* event. If this system, for example, summarizes the duration of events, it is particularly important to correctly classify non-culminated events. In this case, our approach is advantageous. As another example, it is advantageous for a medical system that identifies patient diagnoses to identify stative clauses, since there is a correspondence between **states** and medical diagnoses.

3.6 Related Work in Statistical Disambiguation

The aspectual classification of a clause has thus far been primarily approached from a knowledge-based perspective. For example, Pustejovsky's *generative lexicon* describes semantic interactions between clausal constituents which effect aspectual class [Pustejovsky, 1991]. Additionally, Resnik demonstrates the influence of implicit direct objects [Olsen and Resnik, 1997].

The application of automatic corpus-based techniques to aspectual classification is in its infancy. Klavans and Chodorow [1992] pioneered the application of statistical corpus analysis to aspectual classification by placing verbs on a scale according to the frequency with which they occur with certain aspectual markers from Table 2.2. This way, verbs are automatically ranked according to their "degree of stativity."

Passonneau [1988] and Passonneau and Klavans [1994] empirically measure the accuracy of various aspectual linguistic constraints which can help classify those clauses that appear with an aspectual marker. This way, the predictive accuracy of each aspectual marker is measured. This is an important first step towards making use of aspectual markers, either to directly classify clauses that contain them, or to use them statistically, measuring their frequencies, i.e., as linguistic indicators.

Various statistical and machine learning approaches have been applied over manually marked corpora for disambiguation problems. For example, decision tree induction (i.e. *recursive partitioning*) has been applied to word sense disambiguation [Black, 1988]. Decision tree induction has also been applied to other disambiguation problems, including determiner prediction [Knight, forthcoming], coordination parsing [Resnik, 1993], syntactic parsing [Magerman, 1993], and disambiguating clue phrases [Siegel, 1994, Siegel and McKeown, 1994]. Further related work applying machine learning to language problems is described in Section 6.1.

There is syntactic and semantic data implicit in the frequency with which certain words occur proximately with one another in large corpora (lexical co-occurrence statistics) which can be applied to automatic disambiguation [Smadja and McKeown, 1991]. For example, the lexical co-occurrence data describing the words near an ambiguous word can be used together for word sense disambiguation [Schűtze, 1992] [Leacock et al., 1993]. Further,

training set:				**which verbs:**	**aspectual distinction:**	**results section:**
instances	**constituents:**					
	verb only			all but *be/have*	stativity	4.2
				all **events**	completedness	5.5.1
	multiple	**knowledge source:**				
		linguistic indicators	**open-class constituents:**			
			lexical item (direct obj)	infrequent	completedness	5.6.1
				all **events**	completedness	5.6.1
			Wordnet cat (direct obj)	all **events**	completedness	5.6.2
		Wordnet cat (direct obj)		*"have"*	stativity	4.3
verbs	**supervised learning:**					
	yes	**test set:**				
		non-overlapping set		all **events**	completedness	5.5.2
		instances		all **events**	completedness	5.5.2
		non-overlapping instances		reduced set	completedness	5.5.2
	no			frequent	stativity	6.3.1

Table 3.5: Measuring linguistic indicators and evaluating aspectual classification methods; the experiments reported in this thesis.

co-occurrence data can be used to automatically cluster (i.e. classify) lexical items [Hatzivassiloglou and McKeown, 1993a, Lafferty and Mercer, forthcoming, Pereira et al., 1993a, Lankhorst and Moddemeijer, 1992], and the classes of words in the vicinity of an ambiguity can be accessed by a decision tree to improve disambiguation, e.g., for syntactic parsing [Magerman, 1993].

3.7 Summary

Table 3.5 breaks down our experiments by the *methods* used to measure and evaluate linguistic indicators. This serves as a helpful summary and index, since the three "results" chapters of this thesis, Chapters 4, 5 and 6, divide our work primarily by *aspectual distinction*, and secondarily by learning method. Therefore, this breakdown spans multiple chapters and serves as an index to list which aspectual distinction each method was applied to, and in which section the results are reported.

The first column of Table 3.5 distinguishes between the two types of training sets described in the previous sections: a set of verb instances (i.e., clauses), and a set of (unique) verbs. When training over instances, the second column distinguishes between classifying according to indicators measured over the main verb only, versus incorporating multiple constituents of a clause to classify it. The central results of this thesis are measured over the first of these two options, "verb only." As shown in the far right columns, these results are measured for both stativity and completedness, over unrestricted sets of verbs.

Variations on this basic approach of verb-only indicators augment our evaluation of linguistic indicators and of other methods of aspectual classification. Continuing with a depth-first traversal of Table 3.5's hierarchy, methods that incorporate multiple clausal constituents differ according to how they use the constituents. For the first method, classification is performed according to indicators measured over a set of constituents. This method, detailed in Chapter 5, is further distinguished according to whether indicators are measured over the head of the direct object directly, or measured over the Wordnet semantic category of this noun.

The second method of incorporating multiple constituents also incorporates the Wordnet category of the direct object, in this case to distinguish *have*-clauses according to stativity. However, in this case, instead of measuring linguistic indicators, the Wordnet category is used directly for classification.

Two additional variations are applied when training over a set of unique verbs. Note that there is no way to incorporate multiple constituents in this case, since instead of classifying individual clauses, we are classifying, or *grouping*, individual verbs. The second column of Table 3.5 distinguishes this work according to whether the grouping method was supervised, that is, whether it made use of training data that had been manually marked. As shown, the results of supervised learning are evaluated over three types of test sets that are further detailed in Chapter 5. The unsupervised learning method clusters verbs according to meaning, but is not designed to group words according to any particular semantic distinction. However, the resulting groups are evaluated according to their ability to predict stativity.

Chapter 4

States vs. Events in Natural Language

The ability to distinguish **state** clauses, e.g., *"She resembles her mother,"* from **event** clauses, e.g., *"She ran down the street,"* is a fundamental element of natural language understanding. These two high-level categories correspond to fundamental distinctions in many domains, including, for example, the distinctions between *diagnoses* and *procedure* in the medical domain, and between *analyses* and *activity* in the financial domain. Furthermore, this distinction is crucial for detecting temporal relations in discourse, and for interpreting temporal adverbials, as detailed in Chapter 2. However, the performance of automatically classifying clauses according to stativity has never been evaluated over naturally-occurring text corpora.

In this chapter I evaluate the application of linguistic indicators to classifying verbs according to stativity. Clauses are classified according to indicator values computed over their main verb only, since verb ambiguity is rare in our domain (indicators measured over mutliple constituents are evaluated in Chapter 5). Fourteen linguistic indicators are evaluated over clauses selected uniformly from a text corpus. In this way, we are measuring classification performance over an unrestricted set of verbs. The choice and computation of such indicators is detailed in the previous chapter. We show that the values of at least seven indicators correlate significantly with verb class. Furthermore, these indicators improve classification performance when combined with machine learning methods.

In Section 4.1, I first describe the data over which linguistic indicators are evaluated for distinguishing **state** and **event** verbs, namely, a corpus of medical discharge summaries. An initial analysis of this corpus reveals that most clauses can be classified by their verb alone, with the exception of the verb *have*. Therefore, aspectual classification is reduced to the problem of classifying verbs. However, the remaining challenge is to automatically categorize this open-class group of words according to their behavior in particular domains. Then, we derive upper and lower bounds in classification accuracy.

Next, in Section 4.2, the linguistic indicators are evaluated. For evaluation, clauses (*verb instances*) are classified by the indicator values of the main verb only; although additional clausal constituents such as temporal adverbials are used to *measure* indicator values, only the main verb is considered for *classifying* each input clause. First, the correlation and classification accuracy of each individual indicator is measured. Then, multiple indicators are combined with machine learning methods to increase classification performance. The application of learning methods to this and other natural language problems is further detailed in Chapter 6.

Finally, in Section 4.3, I address the problem of automatically classifying clauses with *have* as the main verb, i.e., *have*-clauses. These clauses are different than most in that their class depends heavily on not only the main verb, but also on the direct object. This section demonstrates that the high-level semantic category of the direct object derived from Wordnet, an on-line lexical database, increases classification performance.

4.1 Corpus: Medical Discharge Summaries

Our experiments are performed across a corpus of 3,224 medical discharge summaries, with a total of 1,159,891 words. As discussed in Section 2.3, a medical discharge summary describes the symptoms, history, diagnosis, treatment and outcome of a patient's visit to the hospital. Each summary consists of unstructured text, divided into sections with titles such as: "History of Present Illness," "Hospital Course," "Brief History," "Medical Summary," "Procedure," and "Discharged."

The text under these titles was extracted and parsed with the English Slot Grammar, 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. Although partial parses are available for sentences that produced errors, these parses were discarded for this study.

Other sections in the summaries, such as "Physical Examination," were ignored since they report the numerical results of certain medical tests, interspersed with incomplete sentences such as, "Chem 7 within normal limits, hemolyzed potassium resulting in a level of 5.3."

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, composing 23.9% of the corpus, always denote a **state**. Therefore, we can focus our efforts on the remaining clauses. Clauses with *have* as their main verb, composing 8.0% of the corpus, describe both **events** and **states**, and are addressed separately in Section 4.3 at the end of this chapter.

In this section, I describe a supervised set of clauses classified by hand and detail the procedure used for this manual marking. These examples are used for training and evaluating our system. Next, to measure verbal ambiguity, I examine which verbs occur in both **state** and **event** classes among these marked clauses. Finally, I detail the upper and lower bounds in accuracy for classifying this set of clauses, which are a factor of the data's distribution and an intrinsic limitation to our approach. These bounds can be used to evaluate results, e.g., by provide a baseline for comparison.

Verb	Occurrences as **States**	Occurrences as **Events**
see	1	21
give	1	19
reveal	1	9
tolerate	2	8
say	1	4
state	1	2
supplement	1	1
describe	1	1
lie	1	1
remain	9	1
show	28	1

Table 4.1: Ambiguous verbs among the 739 training examples.

4.1.1 Manual Marking for Supervised Data

As a basis for evaluating our approach, 2,069 clauses from the parsed corpus were manually marked according to their stativity. In contrast to the entire parsed corpus (97,973 clauses), each clause in this set of supervised data must be judged by a linguist. This subset was selected unformly from clauses in the corpus that had main verbs other than *be*, which always denotes a **state** (see Section 2.1.5).

As a linguistic test for marking, each clause was tested for readability with *"What happened was..."*, an aspectual marker introduced in Chapter 2.[1] In order to find the *fundamental* aspectual class of each clause, modifiers that result in aspectual transformations, such as durative *for*-PPs, and in exceptions, such as *not*, were ignored. A comparison between human markers for this test was performed over a different corpus, and is reported in the next chapter, Section 5.1.1.

Of these, 385 clauses were rejected because of parsing problems (verb or direct object incorrectly identified). 206 clauses had *have* as the main verb, and are dealt with separately in Section 4.3. This left 1,478 clauses, which were divided equally into training and testing sets of 739 clauses each.

4.1.2 Ambiguous Verbs

Most clauses can be classified by their main verb only. Although verbs can appear as either **states** or **events**, as detailed in Section 2.1.5, verb ambiguity in the medical discharge summaries is unusual for verbs other than *be* and *have*; most these words are highly dominated by one sense. Of the 739 clauses included in the training set, 235 verbs occurred. As shown in Table 4.1, only 11 of these verbs were observed as both **states** and **events**. Most of these were highly dominated by one sense. For example, *show* appears primarily as a **state**.

[1]This was suggested by Judith Klavans (personal communication).

	n	States	Events	Baseline Acc.
be	23,409	100.0%	0.0%	100.0%
have	7,882	69.9%	30.1%	69.9%
other verbs	66,682	16.2%	83.8%	83.8%

Table 4.2: Breakdown of verb occurrences.

As shown in Table 4.1, only five verbs - *say*, *state*, *supplement*, *describe*, and *lie*, were not dominated by one class over 80% of the time. Further, each of these was observed less than 6 times a piece, which makes the estimation of sense dominance inaccurate.

This observation illustrates the importance of corpus-based techniques for lexical knowledge acquisition, since our corpus analysis has revealed idiosyncratic uses of verbs in the medical domain; other domains may include non-dominated verbs in addition to *have*. Additionally, 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 Upper and Lower Bounds in Accuracy

As shown in Table 4.2, 83.8% of clauses with main verbs other than *be* and *have* are **events**. Therefore, simply classifying every verb as an **event** achieves an accuracy of 83.8% over the 739 test cases. However, this approach classifies all **state** clauses incorrectly, achieving a *stative recall* of 0.0%. This method serves as a baseline for comparison since we are attempting to improve over an *uninformed* approach.[2]

One limitation to our approach places an upper bound on classification accuracy. Our approach examines only the main verb, since linguistic indicators are computed for verbs only, as described below in Section 4.2. For example, a verb occurring 3 times as an event and 2 times as a state will be misclassified at least 2 times. This limits classification accuracy to a maximum of 97.4% over the test cases due to the presence of verbs with multiple classes, similar to those listed in Table 4.1 for the training cases.

4.2 Indicators for Stativity

Figure 4.1 shows the system overview, as shown in Chapter 3, with the addition of details regarding the medical discharge summary corpus used in this study. As this shows, the values for linguistic indicators are computed across the entire parsed corpus. This is a fully automatic process. Then, the 739 training examples are used to derive a mechanism for combining multiple indicators. The combination of indicators achieves an increase in classification performance. This increase in performance is then validated over the 739 unseen test cases.

[2]Similar baselines for comparison have been used for many classification problems [Duda and Hart, 1973], e.g., part-of-speech tagging [Church, 1988, Allen, 1995].

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 participle	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 4.3: Indicators discriminate between two classes.

Verb	C	`freq`	*"not"* or *"never"*	*temporal* adverb	no subj
show	S	2,131	1.55%	0.52%	18.07%
admit	E	1,895	0.05%	1.11%	91.13%
discharge	E	1,608	0.50%	1.87%	96.64%
feel	S	1,177	4.61%	1.20%	52.52%

Table 4.4: Example verbs and their indicator values.

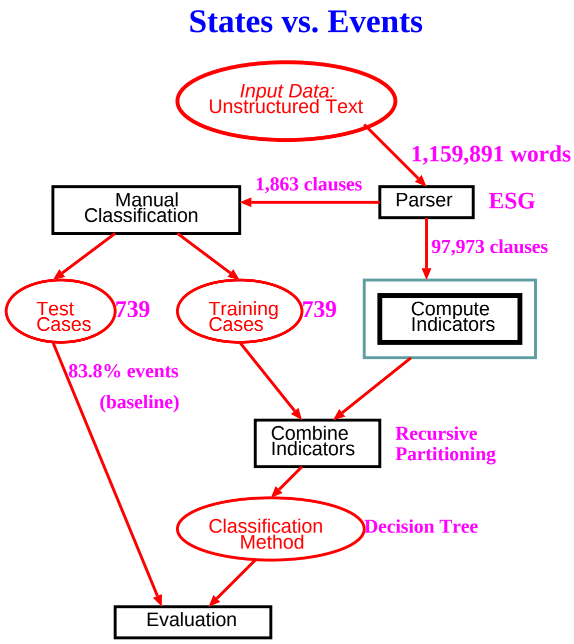


Figure 4.1: System overview with details of medical discharge summary data.

The values of the 14 indicators listed in Table 4.3 were computed, for each verb, across the 97,973 parsed clauses from our corpus of medical discharge summaries. As described in Section 3.2, each indicator has a unique value for each verb which corresponds to the frequency of the aspectual marker with the verb (except verb `frequency`, which is an absolute measure over the corpus). Table 4.4 illustrates some additional detail by listing indicator values for four example verbs.

This section evaluates the correlation and classification performance of these indicators. Evaluations are performed by classifying clauses (*verb instances*) by the indicator values of the main verb only; although additional clausal constituents are used to *measure* indicator values, only the main verb is considered to *classify* each input clause. I first examine the correlation between the values of each individual indicator, and verb class (**state** or **event**). Second, I investigate which individual indicators can be used to increase classification accuracy over the baseline of 83.8%. Last, I briefly evaluate the ability for learning methods to combine multiple indicators to improve classification performance. The discussion and experimentation of machine learning methods are further detailed in Chapter 6.

4.2.1 Predictive Correlations of Individual Indicators

The second and third columns of Table 4.3 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 modified by these adverbs. These values are computed over the 739 training cases in order

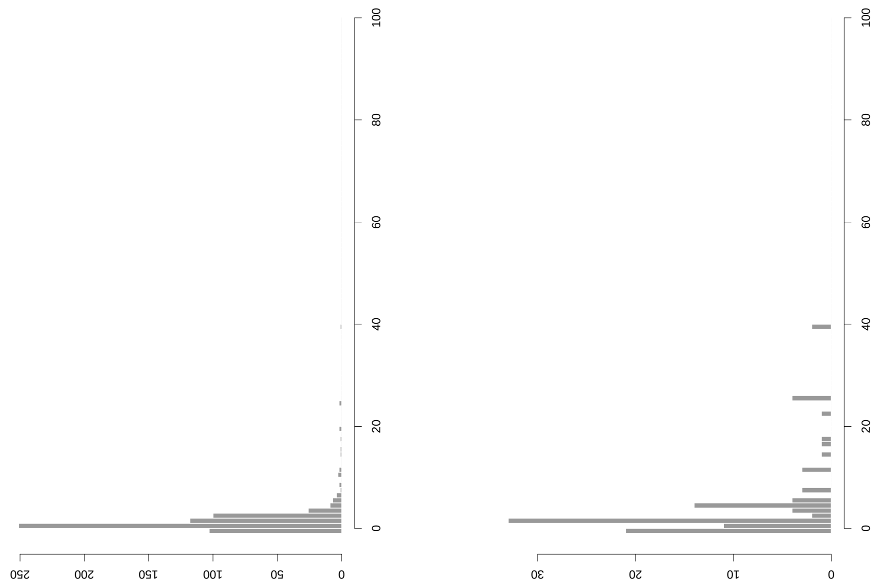


Figure 4.2: Histograms of *not/never* indicator for **events** (left) and **states** (right).

to avoid biasing the classification experiments in the sections below, which are evaluated over the unseen test cases. As described above, the training set (and test set) excludes clauses with *be* or *have* as the main verb.

The differences in stative and event means are statistically significant ($p < .01$) for the first seven of the 14 indicators listed in Table 4.3. The fourth column shows the results of t-tests that compare indicator values over stative verbs to those over event verbs for each indicator. For example, there is less than a 0.05% chance that the differences between stative and event means for the first seven indicators listed is due to chance. The differences in average value for the bottom seven indicators were not confirmed to be significant with this small sample size (739 training examples).

This analysis has revealed correlations between verb class and five indicators that are not traditionally linked to stativity on a linguistic basis. Of the top seven indicators shown to have positive correlations with stativity, three have been linguistically motivated, as shown in Table 2.2. The other four were not previously hypothesized to predict aspectual class: (1) Verb `frequency`, (2) occurrences modified by *"not"* or *"never"*, (3) occurrences with no true subject, and (4) occurrences in the past or present participle. Furthermore, the last of these seven, occurrences in the perfect tense, was not previously hypothesized to predict stativity in particular.

Several indicators only help distinguish a subset of verbs, and are therefore incomplete when used by themselves. In particular, the correlations between these indicators and verb class are caused by a small number of outliers. This is illustrated by examining histograms of indicator values, broken down by verb class, as shown in Figures 4.2 through 4.6. The values

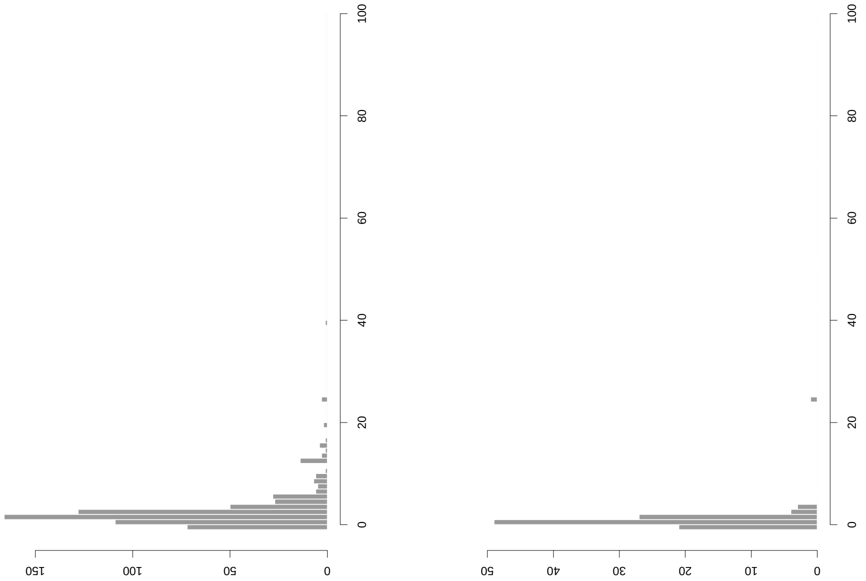


Figure 4.3: Histograms of temporal adverb indicator for **events** (left) and **states** (right).

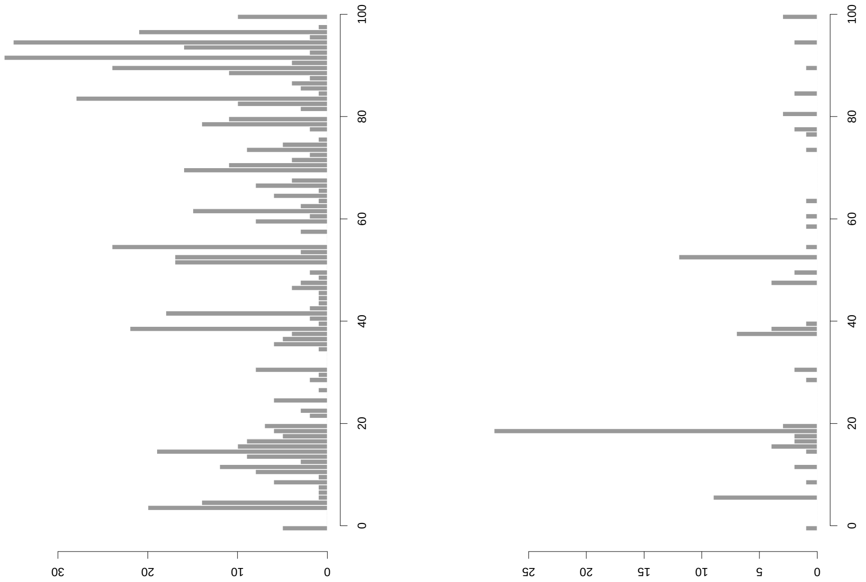


Figure 4.4: Histograms of “no subject” indicator for **events** (left) and **states** (right).

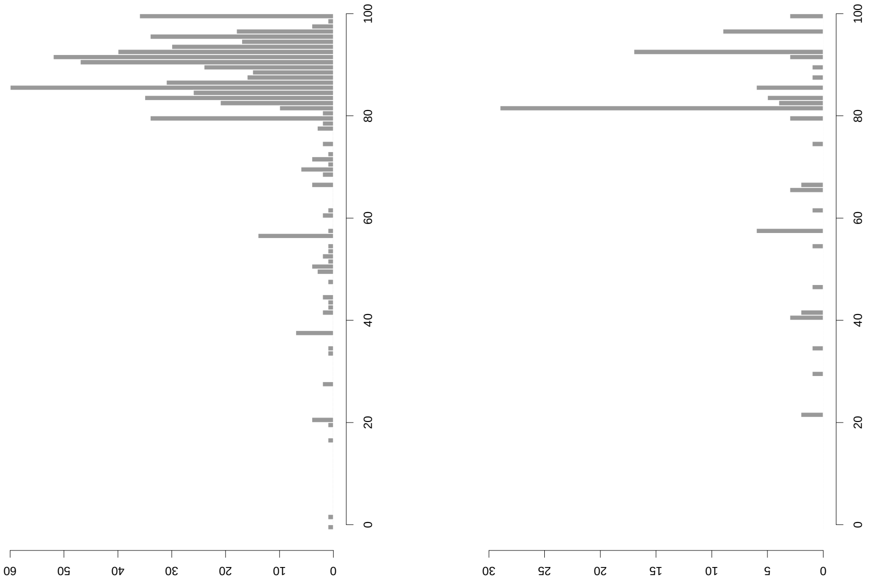


Figure 4.5: Histograms of past/pres participle indicator for **events** (left) and **states** (right).

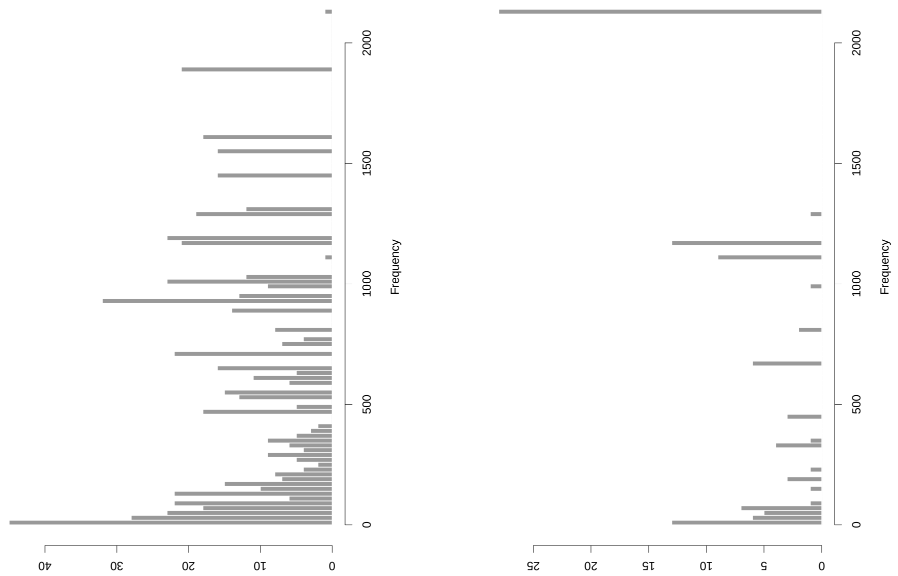


Figure 4.6: Histograms of `frequency` indicator for **events** (left) and **states** (right).

adverb	n	adverb	n
then	991	weekly	2
never	148	hourly	1
just	91	yearly	1
finally	82	again and again	0
first	67	oftenly	0
immediately	61	oftentimes	0
daily	44	annually	0
frequently	41	thrice	0
always	30	at last	0
often	27	hebdomadally	0
twice	28	monthly	0
suddenly	23	at once	0
already	22	oft	0
once	17	oft-times	0
repeatedly	14		

Table 4.5: Frequencies of temporal adverbs over entire corpus.

in these histograms are computed over the 739 training cases only. For example, Figure 4.2 shows two histograms, depicting the distribution of values for the *not/never* indicator over the training cases. As listed in Table 4.3, the average value over **events**, 1.56%, is lower than the average value over **states**, 4.44%. An examination of the corresponding histograms reveals that this difference in average value is due primarily to the several outlying data points shown in the **stative** histogram, since most values lie together towards the left end in both histograms. A similar phenomenon is observed for the *temporal* adverb indicator, as shown in Figure 4.3 (average values: 2.70% over **events**, 1.00% over **states**). Additionally, this interpretation may also be applicable, though not as clearly, to the `frequency` and past/present participle indicators, as shown in Figures 4.6 and 4.5, respectively.

The two indicators that involve adverbs and are significantly correlated are each dominated by individual adverbs, when measured across this corpus. Table 4.5 shows the 29 adverbs in the temporal adverb class, as motivated in Section 3.2.2, and the frequency with which each adverb modified the main verb among the 97,973 parsed clauses. The adverb *then* occurs much more frequently than any other, accounting for 58.6% of this indicator's value. Similarly, *not* accounts for 91.6% of the occurrences of the *not/never* indicator, since *not* occurs 1,619 times as a modifier of the main verb, whereas *never* occurs only 148 times. Therefore, these two indicators primarily measure the frequency of the adverbs *then* and *not*, respectively.

4.2.2 Classification Accuracy with Individual Indicators

A positive correlation between indicator value and verb class does not necessarily mean an indicator can be used to increase classification accuracy over the baseline of 83.8%.

	overall accuracy	States recall	States precision	Events recall	Events precision
decision tree	93.9%	74.2%	86.4%	97.7%	95.1%
genetic programming	91.2%	47.4%	97.3%	99.7%	90.7%
log-linear	86.7%	34.2%	68.3%	96.9%	88.4%
baseline	83.8%	0.0%	100.0%	100.0%	83.8%

Table 4.6: Comparison of three learning methods and a performance baseline.

This is because of the dominance of **events** among the testing examples; a threshold to distinguish verbs that correctly classifies more than half of each class will have an accuracy lower than the baseline if the number of states correctly classified is less than the number of events misclassified. To examine this, each indicator was tested individually for its ability to improve classification accuracy over the baseline by establishing the best classification threshold over the training data. The performance of each indicator was validated over the testing data using the same threshold.

Only the `frequency` indicator succeeded in significantly improving classification accuracy. Both `frequency` and occurrences with *not* and *never* improved classification accuracy on the training data over the baseline obtained by classifying all clauses as events. To validate this improved accuracy, the thresholds established over the training set were used over the test set, with resulting accuracies of 88.0% and 84.0%, respectively. Binomial tests show the first of these to be a significant improvement over the baseline of 83.8%, but not the second.

This improvement in accuracy was achieved simply by discriminating the popular verb *show* as a **state**, but classifying all other verbs as **events**. This is illustrated by examining the histograms in Figure 4.6. The spike at the far right of the **state** histogram represents occurrences of the verb *show*. The threshold selected is right at this spike, so *show* is correctly classified, in addition to all **event** verbs, which is an improvement over the baseline, which misclassifies all occurrences of *show*.

4.2.3 Combining Indicators with Machine Learning

Three machine learning methods were employed to combine the fourteen indicators: decision trees, genetic programming and log-linear regression. These methods, as well as experimental results, are further detailed in Chapter 6. In this section, I preview the overall results, illustrating that machine learning allows us to take full advantage of our wide range of indicators by coordinating and combining them.

All three machine learning methods successfully combined indicator values, improving classification accuracy over the baseline measure. As shown in Figure 4.7, and detailed in Table 4.6, the decision tree's accuracy was 93.9%, genetic programming's function trees had an average accuracy of 91.2% over seven runs, and log-linear regression achieved an 86.7% accuracy. Binomial tests showed that both the decision tree and genetic programming

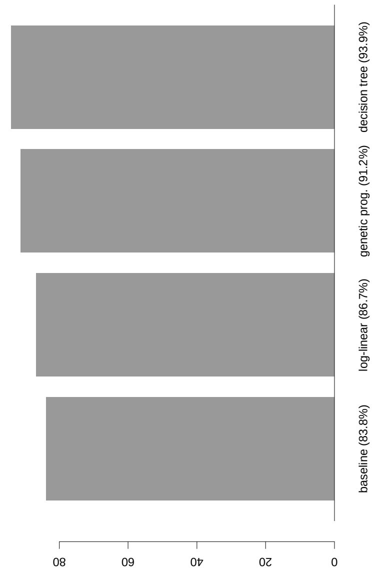


Figure 4.7: Accuracy of three learning paradigms for stativity.

achieved a significant improvement over the 88.0% accuracy achieved by the `frequency` indicator alone. Therefore, we have shown that machine learning methods can successfully combine multiple numerical indicators to improve the accuracy by which verbs are classified.

The differences in accuracy between the three methods are each significant ($p < .01$). Therefore, these results highlight the importance of how linear and non-linear interactions between numerical linguistic indicators are modeled. These differing models are contrasted in detail in Chapter 6.

Classification performance is equally competitive without the `frequency` indicator. Since `frequency` is the only indicator that can increase accuracy by itself, and since it is the first discriminator of the decision tree, it may appear that `frequency` highly dominates the set of indicators. This could be problematic, since the relationship between verb `frequency` and verb category may be particularly domain dependent, in which case `frequency` could be less informative when applied to other domains. However, 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. Therefore, our results are not entirely dependent on the `frequency` indicator.

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 Figure 4.8, and detailed in Table 4.6, stative recalls of 74.2%, 47.4% and 34.2% were achieved by the three learning methods, 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. This difference in recall is more dramatic than the accuracy improvement because of the dominance of event clauses

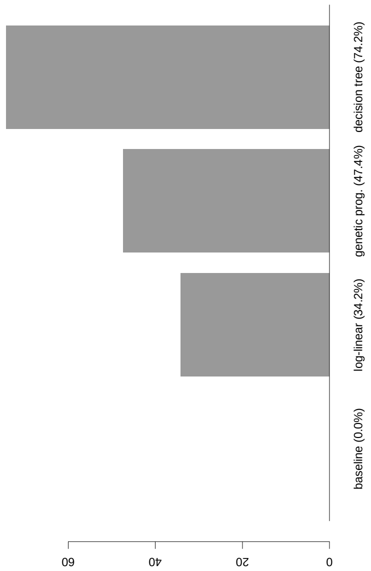


Figure 4.8: Stative recall of three learning paradigms for stativity.

	overall accuracy	States recall	States precision	Events recall	Events precision
train-test	93.9%	74.2%	86.4%	97.7%	95.1%
baseline	83.8%	0.0%	100.0%	100.0%	83.8%
test-train	93.2%	83.8%	72.7%	94.8%	97.2%
test-train baseline	85.8%	0.0%	100.0%	100.0%	85.8%

Table 4.7: Train-test sensitivity analysis for decision tree learning, with updated baseline.

in the test set.

This favorable tradeoff between recall values presents an advantage for applications that weigh the identification of stative clauses more heavily than that of event clauses. For example, a prepositional phrase denoting a duration with *for*, e.g., *"for a minute,"* describes the duration of a state, e.g., *"She felt sick for two weeks,"* or the duration of the state that results from a culminated event, e.g., *"She left the hospital for a week."* That is, correctly identifying the use of *for* depends on identifying the stativity of the clause it modifies. A language understanding system that incorrectly classifies *"She felt sick for two weeks"* as a non-culminated event will not detect that *"for two weeks"* describes the duration of the *feel*-state. If this system, for example, summarizes durations, it is important to correctly identify states. In this case, our approach is advantageous.

Sensitivity analysis further validates the ten-point increase in accuracy achieved over the baseline. Although a binomial test shows that 93.9% accuracy is a significant increase over the baseline of 83.8%, this tells little about how sensitive this high measurement is on the particular test set used for evaluation. To test this, the training and test sets were swapped; a decision tree was induced over the test set, and evaluated over the training set. As shown in Table 4.7, the resulting accuracy (over the "training" set), 93.2%, was comparable to that in the original experiment. For this evaluation, there is a slightly higher accuracy baseline, since 85.8% of the training cases are events. Nonetheless, accuracy still marks a significant improvement over this baseline accord to a binomial test ($p < .001$). Note that improved results may be expected with this reversal of training and test sets, since a portion of the stativity training data was used to develop the set of linguistic indicators. Indeed, there is an even wider tradeoff in recall, with 83.8% stative recall achieved, as compared to the original stative recall of 74.2%. These results suggest that performance is not sensitive to the particular testing set over which the linguistic indicators were evaluated.

4.3 Classifying the Verb *Have*

The verb *have* occurs as the main verb of clauses frequently and is aspectually ambiguous in medical discharge summaries, occurring 69.9% of the time as a **state**, and 30.1% of the time as an **event**. Therefore, it is important to examine the context in which *have* is used to determine its meaning. In particular, the direct object of *have* is the major determiner for the stativity of a *have*-clause, as illustrated in Section 2.1.5.

This section evaluates the classification performance of a linguistically motivated rule that considers the semantic category of the direct object of *have*. The category of nouns is established by utilizing *Wordnet* [Miller et al., 1993]. I show that the high-level category of direct objects of *have* correspond greatly with the stativity class of clauses. Furthermore, these categories improve the automatic classification of *have*-clauses.

In this section, I first describe the use of Wordnet to derive object categories. Next I show a simple classification rule based on Wordnet classes, which is manually established with linguistic principles governing aspect and is supported across the training data. Finally, as results, I detail the correspondence of Wordnet classes with clausal stativity, and evaluate the classification performance of the classification rule.

4.3.1 Wordnet

To derive semantic categories for the direct object of each clause, I employ the on-line lexicon *Wordnet* [Miller et al., 1993]. Wordnet groups nouns into 25 high-level categories at the top of its semantic hierarchy. These categories are listed in Table 4.8. Our use of Wordnet to classify nouns is detailed in Section 3.4.1.

Wordnet is able to handle the technical domain of medical discharge summaries. This is largely because the nouns that appear frequently as direct objects of *have* are widely-known or non-technical terms. This is illustrated in Table 4.9, which shows the most popular objects of *have*. Acting as the direct object, these 13 nouns account for 25.7% of the 7,882 occurrences of *have* in the discharge summaries. Table 4.9 also shows the frequency of each

Wordnet class	Occurrences as **States**	Occurrences as **Events**
location	0	1
event	2	5
act	6	6
artifact	5	3
phenomenon	2	1
entity	2	1
attribute	3	1
measure	3	1
N/A	5	1
cognition	11	1
state	19	0
time	9	0
substance	5	0
relation	3	0
person	2	0
communication	1	0
causalagent	1	0
possession	1	0
group	1	0
food	1	0
shape	0	0
natural object	0	0
feeling	0	0
animal	0	0
plant	0	0
motivation	0	0

Table 4.8: Wordnet classes of direct objects of *have* in the training data.

direct object	n	Wordnet class	classification of clause
history	624	`time`	**state**
episode	280	`event`	**event**
pain	192	`cognition`	**state**
fever	123	`cognition`	**state**
temperature	113	`attribute`	***state**
allergy	109	`state`	**state**
movement	106	`act`	***event**
course	96	`act`	***event**
<none>	91	`<none>`	***state**
symptom	81	`cognition`	***state**
complaint	73	`state`	***state**
seizure	72	`event`	**event**
nausea	67	`cognition`	***state**

Table 4.9: Frequent objects of *have*, their Wordnet category, and the aspectual class of *have*-clauses with the object. Asterisks (*) denote classifications that were intuitively derived, since these examples did not occur in the training cases.

noun as an object of *have* across the entire corpus of discharge summaries, and the high level semantic category assigned by Wordnet. The last column shows the classification of *have*-clauses with each noun as a direct object. This classification is the most frequent across the training data, except for those marked with an asterisk, which did not appear in the training data. The class listed for these cases is based on informal intuition.

However, there are some problematic cases that are not easily handled with Wordnet. The direct object of 10.9% of the *have*-clauses are unknown to Wordnet ("N/A"). This includes medical terminology, e.g. *anticonvulsants* and *vitrectomy*, as well as complex expressions parsed by the English Slot Grammar, such as *"between 32 and 29"*. Furthermore, Wordnet categories and lexical ambiguity are occasionally misleading for the task of classifying *have*-clauses. For example, *"The patient had Medicaid"* is a **state**, but Wordnet classifies *Medicaid* as an `act` (which normally corresponds with **event** – see the next subsection below). This of course could be solved via sensitivity to capitalization. As a reverse examples, *"The patient had an enema"* is an **event**, but Wordnet lists *enema* as `artifact` before `act` (as described in Section 3.4.1, only the first Wordnet sense is used). As a final example, *"The patient had a urine culture"* is an **event**, but Wordnet's first sense of *culture* is `group`.

4.3.2 Classification Rule

A simple rule for classifying *have*-clauses, shown in Table 4.10, was derived manually. This rule classifies *have*-clauses based on the semantic category of their direct object. In particular, clauses with direct objects belonging to the category `event`, `act`, `phenomenon`,

If object is a(n):	**then** class is:	n	Frequent nouns
`act`	**event**	1,157	*movement* (106) *course* (96) *difficulty* (66) *scan* (61) *admission* (60)
`event`		655	*episode* (280) *seizure* (72) *pulse* (28) *recurrence* (25) *onset* (24)
`phenomenon`		242	*pressure* (52) *x-ray* (30) *flatus* (21) *response* (19) *intake* (15)
`communication`		194	*sign* (25) *resolution* (22) *effusion* (18) *section* (17) *electrocardiogram* (12)
`possession`		59	*loss* (27) *amount* (15) *residual* (5) *insurance* (4) *cut* (3)
`food`		17	*bun* (5) *coffee* (2) *vitamin* (1) *sugar* (1) *scotch* (1)
`cognition`	**state**	1,146	*pain* (192) *fever* (123) *symptom* (81) *nausea* (67) *test* (54)
`state`		875	*allergy* (109) *complaint* (73) *infection* (56) *disease* (56) *problem* (40)
N/A		860	*echocardiogram* (51) *hematocrit* (41) *ultrasound* (34) *stenosis* (29)
`time`		636	*history* (624) *rhythm* (8) *past* (3) *gestation* (1)
`artifact`		415	*catheter* (20) *stool* (19) *tube* (17) *output* (16) *PPD* (15)
`attribute`		349	*temperature* (113) *shortness* (46) *tenderness* (26) *level* (22) *sound* (16)
`entity`		209	*chest* (20) *head* (13) *abdomen* (13) *artery* (12) *shunt* (7)
`measure`		205	*count* (41) *increase* (18) *bout* (15) *fall* (12) *day* (9)
`substance`		182	*blood* (29) *thallium* (15) *sodium* (11) *urine* (10) *fluid* (9)
`relation`		116	*change* (40) *rate* (32) *function* (12) *aspirate* (5) *relationship* (3)
`person`		115	*child* (13) *aide* (13) *son* (8) *patient* (8) *temp* (6)
`group`		84	*culture* (41) *series* (7) *meeting* (6) *progression* (4) *panel* (4)
`location`		49	*area* (8) *post* (7) *left* (6) *state* (4) *lab* (4)
`feeling`		48	*appetite* (18) *relief* (7) *chill* (6) *preference* (3) *feeling* (3)
`pronoun`		39	*which* (18) *this* (8) *her* (4) *them* (3) *it* (3)
`animal`		12	*dog* (3) *paceer* (2) *pet* (1) *fetus* (1) *emu* (1)

Table 4.10: Classification rule for direct object of *have*-clauses.

	overall accuracy	States recall	States precision	Events recall	Events precision
using Wordnet	79.6%	84.7%	85.9%	67.7%	65.6%
baseline	69.9%	100.0%	69.9%	0.0%	100.0%

Table 4.11: Classification performance for *have*.

`communication`, `possession`, or `food` are classified as **events**, and all others are classified as **states**. This rule is evaluated below in the next subsection.

This classification rule was manually constructed from linguistic observations of the selectional restrictions on aspectual class (no automatic learning). As outlined in Section 2.1.5, the class of a *have*-clause is determined primarily by the category of the direct object. In particular, if the direct object is a nominalized **event**, the clause describes an **event**. For this reason, it is clear why the Wordnet categories `event`, `act`, `phenomenon` and `communication` have been selected. The category `possession` was selected since, as shown in Table 4.10, most occurrences of `possession` as a direct object of *have* are instances of *loss*, e.g. *"The patient had blood loss"* describes an **event**. The category `food` was selected to cover idioms such as *"The patient had lunch"* (**event**).

Furthermore, this classification rule is quantitatively supported over the training data. For each Wordnet class, Table 4.8 shows the distribution of **event** and **state** *have*-clauses with a direct object belonging to that class. As this table shows, each Wordnet category linked to **states** with our rule occurs at least as frequently in **stative** clauses as they do in **event** clauses within the training set, with the exception of `communication`, `possession` and `food`. However, these categories only occur one time each in the training data, which is far too infrequent to present a significant quantitative counter-argument to linguistic intuition. Therefore, the rule is maintained as-is for evaluation over the test cases.

4.3.3 Evaluation

There is a strong correlation between the Wordnet category of a direct object, and the stative class of the *have*-clauses that it appears in. When using the classification rule established in the previous subsection, the Wordnet categories that appear more than five times in the test data correctly predict the class of *have*-clauses with an average precision of 82.7%. In particular, `act` and `event` predict **event** *have*-clauses 85.7% and 66.7% correctly, respectively. Likewise, **states** are predicted with `artifact` (62.5% precision), `cognition` (88.2%), `state` (93.3%) and `time` (100.0%).

69.9% of the 103 *have*-clause test cases are **states**. Therefore, simply classifying each of these clauses as a **state** achieves an accuracy of 69.9%. However, this approach classifies all **event** clauses incorrectly, achieving an *event recall* of 0.0%. This method serves as a baseline for comparison since we are attempting to improve over an *uninformed* approach.[3]

[3] Similar baselines for comparison have been used for many classification problems [Duda and Hart, 1973], e.g., part-of-speech tagging [Church, 1988, Allen, 1995].

The classification rule using Wordnet classes significantly improves classification accuracy over the baseline measure. As shown in Table 4.11, an accuracy of 79.6% was achieved. A binomial test showed that this accuracy improvement over the baseline is significant ($p < .05$).

The increase in the number of **event** clauses correctly classified, i.e. event recall, illustrates an event greater improvement over the baseline. As shown in Table 4.11, an event recall of 67.7% was achieved by the classification rule, as compared to the 0.0% event recall achieved by the baseline, while no loss in overall accuracy was suffered. The baseline does not classify any **event** clauses correctly because it classifies all clauses as **states**. This difference in recall is more dramatic than the accuracy improvement because of the dominance of **state** clauses in the test set. This favorable tradeoff between recall values presents an advantage for applications that weigh the identification of stative clauses more heavily than that of event clauses.

This improvement in classification performance demonstrates the significance of the direct object in determining the stative class. Furthermore, we have shown that a high-level semantic categorization of the direct object can be used to predict the class of a clause. The categorization scheme implemented by Wordnet is informative enough to show a significant improvement in classification performance, despite difficulties presented by lexical ambiguity and the technical terminology of the medical domain.

4.4 Conclusions

When used in combination, linguistic indicators significantly improve the classification of an unrestricted set of verbs according to stativity. A total of fourteen indicators, each measured automatically across a corpus of text, were combined. Each of three machine learning techniques successfully combined the indicators to improve classification performance. The best of the three, decision tree induction, achieved a classification accuracy of 93.9%, as compared to the uninformed baseline's accuracy of 83.8%. Furthermore, genetic programming and log-linear regression also achieved improvements over the baseline.

This improvement in classification performance is more dramatically illustrated by the favorable tradeoff between stative and event recall achieved by all three learning methods, which is profitable for tasks that weigh the identification of states more heavily than events. Stative recalls of 74.2%, 47.4% and 34.2% were achieved by the three learning methods, as compared to the 0.0% stative recall achieved by the baseline.

This analysis has revealed correlations between verb class and five indicators that are not traditionally linked to stativity on a linguistic basis. Four were not previously hypothesized to predict aspectual class: (1) Verb `frequency`, (2) occurrences modified by *"not"* or *"never"*, (3) occurrences with no true subject, and (4) occurrences in the past or present participle. One of these four, verb `frequency`, individually increased classification accuracy from the baseline method to 88.0%. Furthermore, the *perfect* tense was not previously hypothesized to predict stativity in particular.

Several indicators only help distinguish a subset of verbs, and are therefore incomplete when used by themselves. In particular, the correlations between these indicators and verb class are caused by a small number of outliers. This was illustrated by examining histograms of indicator values. To alleviate this problem, machine learning is used to automatically combine multiple indicators, as detailed in Chapter 6.

A semantic category associated with the direct object of *have* can be used to predict the aspectual class of a *have*-clause. We demonstrated a correlation between the Wordnet category of a direct object, and the stative class of the *have*-clauses that it appears in. A simple rule for classifying *have*-clauses based on the Wordnet category of the direct object was manually derived using linguistic observations. This classification rule using Wordnet achieved an accuracy of 79.6% as compared to the baseline of 69.9%. Furthermore, an event recall of 67.7% was achieved by the classification rule, as compared to the 0.0% event recall achieved by the baseline. This improvement in classification performance demonstrates the significance of the direct object in determining the stative class. Furthermore, we have shown that a high-level semantic categorization of the direct object can be used to predict the class of a clause.

Chapter 5

Culminated vs. Non-Culminated Events in Natural Language

Completedness determines whether an event reaches a culminating point in time at which a new state is introduced. For example *"The boy held the tiller"* is **non-culminated**, an event with no explicit temporal boundaries, while *"You will find us there"* is **culminated**, an event with a determined point of completion.[1] This categorization is crucial to many natural language problems, including the determination of temporal constraints between events in a discourse [Moens and Steedman, 1988, Passonneau, 1988] as well as lexical choice and tense selection in machine translation [Moens and Steedman, 1988, Klavans and Chodorow, 1992, Klavans, 1994, Dorr, 1992]. For a complete discussion, see Chapter 2.

In this chapter I present the application and resulting classification performance of linguistic indicators automatically computed over a corpus of unstructured text. The choice and computation of such indicators was detailed in Chapter 3. We show that the values of at least six indicators correlate significantly with verb class. Furthermore, these indicators improve classification performance when combined with machine learning methods.

In Section 5.1, I first describe the data over which linguistic indicators are evaluated for distinguishing **culminated** and **non-culminated** clauses, namely, a corpus of novels. We derive upper and lower bounds in classification accuracy.

Next, in Sections 5.3 and 5.4, the correlations and classification accuracy of each individual indicator is measured. Then, multiple indicators are combined with machine learning methods to increase classification performance in Section 5.5. The application of learning methods to natural language problems is further detailed in Chapter 6.

Finally, in Section 5.6, I evaluate linguistic indicators measured over clausal constituents other than the main verb, providing evidence that this technique helps address the problems of sparsity and verb ambiguity.

[1]Most examples shown are derived from the corpora used in this study.

Title	Author	Translator
Plato's Republic	Plato	Benjamin Jowett
Terminal Compromise	Winn Schwartau	
The Nigger of the "Narcissus"	Joseph Conrad	
The Time Machine	H.G. Wells	
Typee	Herman Melville	
Utopia	Sir Thomas More	
The Sign of Four	Arthur Conan Doyle	
The Jungle Book	Rudyard Kipling	
The Wonderful Wizard of Oz	L. Frank Baum	
The Call of the Wild	Jack London	

Table 5.1: Novels provide a corpus for the study of aspectual classification. Available from Project Bartleby by telnet to `info.columbia.edu`.

5.1 Corpus: Novels in English

The corpus over which linguistic indicators were computed to distinguish **culminated** from **non-culminated** clauses consists of novels in the English language. Novels provide two advantages for this task. First, because of their narrative nature, they exhibit a relatively equal distribution of **culminated** and **non-culminated events**. This is in contrast with, for example, the Wall Street Journal, which has a very small proportion of **non-culminated events**. Second, novels are domain-independent, covering a wide range of topics. Therefore, corpus analysis performed over novels is not able to focus on the way words are used in any one particular specialty. In this way, our results are not specialized, and may extend to multiple domains.

Our experiments are performed across a corpus of ten novels, listed in Table 5.1, comprising 846,913 words. These novels were parsed with the English Slot Grammar, resulting in 75,289 clauses that were parsed fully, with no self-diagnostic errors. With the same settings used for parsing medical reports, ESG produced error messages on 22,505 of this corpus' 59,816 complex sentences. The values of the 14 indicators listed in Table 5.2 were computed, for each verb, across the parsed clauses. Note that in this Chapter, the *perfect* indicator differs in that we ignore occurrence of the perfect in clauses that are also in the progressive, as linguistically motivated in Section 2.1.1.

In this section, I describe the set of clauses classified by hand and detail the procedure used for this manual marking. Then I detail the upper and lower bounds in accuracy for classifying this set of clauses.

5.1.1 Manual Marking for Supervised Data

To evaluate the performance of our system, we manually marked 884 clauses from the parsed corpus according to their aspectual class. These 884 were selected evenly across the 10 novels from parsed clauses that do not have *be* as the main verb, since we are testing a distinction between **events**. Of these, 109 were rejected because of parsing problems (verb or direct object incorrectly identified), and 160 rejected because they described **states**. This left 615 **event** clauses over which to evaluate classification performance.

We used linguistic tests that were selected for this task by Passonneau [1988]. First, if a clause can be read in a *pseudo-cleft*, it is non-stative (i.e. an *event*), e.g., *"What its parents did was run off,"* versus *"*What we did was know what is on the other side."* As an additional check for stativity, we also tested the clause with *"What happened was..."* Second, 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-culminated**) entails that *"We talked just like men"*, but *"The woman was building a house"* (**culminated**) does not necessarily entail that *"The woman built a house."*

Cross-checking between linguists shows 91% agreement. In particular, in a pilot study manually annotating 89 clauses from the corpus of novels, two linguists agreed 81 times. Informal analysis suggests the remaining disagreement could be further divided in half by a few simple refinements of the annotation protocol. Furthermore, of 57 clauses agreed to be **events**, 46 were annotated in agreement with respect to completedness.

The verb *say*, which is a frequent **point**, i.e., **non-culminated** and **non-extended**, poses a challenge for manual marking. **Points** are misclassified by the test for completedness described above since they are **non-extended** and therefore cannot be placed in the progressive, e.g.:

*"*She was saying, 'I have a problem.' "*

However, *say* is usually **non-culminated**, e.g.:

*"*She has said, 'I have a problem.' "*

Therefore, *say*, which occurs nine times in the test set, was marked incorrectly as **culminated**. After some initial experimentation, we switched the class of each occurrence of *say* in our supervised data to **non-culminated**.

The division into training and test sets was derived such that the distribution of classes was equal between the two sets. This eliminated the possibility of deviant distributions between the training and test sets, which would otherwise be likely given the small size of our sample. This was important since preliminary experiments had indicated difficulty in demonstrating a significant increase in classification accuracy for completedness. However, the subsequent change to *say* made the distribution slightly uneven. This division also split the manually annotated data chronologically, so that the training and test cases contained equal numbers of examples that were manually marked at both the beginning and the end of the marking process. This has the advantage of equalizing any change in human bias that developed during our annotation process.

This process resulted in 307 training cases (196 **culminated**) and 308 test cases (195 **culminated**).

5.1.2 Lower Bound in Accuracy

Since 63.3% of test cases are **culminated events**, simply classifying every clause as **culminated** achieves an accuracy of 63.3% over the 308 test cases (Baseline A). However, this approach classifies all **non-culminated** clauses incorrectly, achieving a *non-culminated recall* of 0.0%. This method serves as a baseline for comparison since we are attempting to improve over an *uninformed* approach.[2]

5.1.3 Verb Ambiguity

The prevalence and class distributions of ambiguous verbs is difficult to estimate over this data because many verbs only appear once. For examle, 157 training cases have main verbs that appear only once in the training set. Ambiguity can only be measured over verbs occurring multiple times.

However, some ambiguous verbs do occur multiple times in the training and test sets. Because of this, for portions of this chapter that measure indicators over the main verb only and evaluate over verb instances, there is a maximum classification accuracy of 93.2% over the test cases due to the presence of ambiguous verbs in the test set.

5.2 Comparison: Completedness vs. Stativity

Several factors suggest that automatic classification according to completedness is a more challenging task than classification according to stativity:

- Comparisons between human markers show more disagreement when marking according to completedness.
- Verb sparsity makes verb ambiguity more difficult to measure and approach.
- Linguistic evidence suggests that multiple constituents besides the main verb and direct object play a larger role in determining completedness than stativity.
- Linguistic evidence suggests that a larger context than the clause itself plays a larger role in determining completedness.

5.3 Predictive Correlations of Linguistic Indicators

The second and third columns of Table 5.2 show the average value for each indicator over stative and event clauses, as measured over the training examples. For example, 7.87% of **culminated** clauses appear in the perfect tense, but only 2.88% of **non-culminated** clauses appear in the perfect. These values were computed strictly over the 307 training

[2]Similar baselines for comparison have been used for many classification problems [Duda and Hart, 1973], e.g., part-of-speech tagging [Church, 1988, Allen, 1995].

Linguistic Indicator	Culminated Mean	Non-Culminated 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 participle	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 5.2: Indicators discriminate between two classes.

cases in order to avoid biasing the classification experiments in the sections below, which are evaluated over the unseen cases.

The differences in **culminated** and **non-culminated** means are statistically significant ($p < .05$) for the first six of the 14 indicators listed in Table 5.2. The fourth column shows the results of t-tests that compare indicator values over **culminated** verbs to those over **non-culminated** verbs for each indicator. For example, there is less than a 0.05% chance that the differences between **culminated** and **non-culminated** means for the first six indicators listed is due to chance. The differences in average value for the bottom eight indicators were not confirmed to be significant with this small sample size (307 training examples).

This analysis has revealed correlations between verb class and four indicators that are not traditionally linked to completedness on a linguistic basis. Of the six indicators shown to have positive correlations with completedness, three have been linguistically motivated, as shown in Table 2.2. The other three were not previously hypothesized to predict aspectual class: (1) occurrences modified by a *manner* adverb, (2) occurrences in the past or present participle, and (3) occurrences with no deep subject. Furthermore, occurrences in the progressive voice were not previously hypothesized to predict completedness in particular.

5.4 Classification Performance with Individual Indicators

A positive correlation between indicator value and verb class does not necessarily mean an indicator can be used to increase classification accuracy over the baseline of 63.3%. This is because of the dominance of **culminated events** among the testing examples. To examine

indicator	overall accuracy	Culminated recall	Culminated precision	Non-culminated recall	Non-culminated precision
perfect	64.6%	71.8%	72.2%	52.2%	51.8%
manner adverb	67.2%	98.5%	66.2%	13.3%	83.3%
`frequency`	64.6%	95.9%	64.9%	10.6%	60.0%
past/pres participle	64.0%	95.4%	64.6%	9.7%	55.0%
past tense	64.6%	98.5%	64.4%	6.2%	70%
duration *for*-PP	65.3%	100.0%	64.6%	5.3%	100.0%
evaluation adverb	64.0%	99.0%	63.9%	3.5%	66.7%
baseline	63.3%	100.0%	63.3%	0.0%	100.0%

Table 5.3: Classification performance of individual indicators, optimized for accuracy.

this, each indicator was tested individually for its ability to improve classification accuracy over the baseline by establishing the best classification threshold over the training data. The performance of each indicator was validated over the testing data using the same threshold.

No indicator was shown to significantly improve classification accuracy over the baseline. As shown in Table 5.3, which shows the testing results of the indicators that improved accuracy during training, the highest accuracy is attained by the *manner* adverb indicator, which achieves 67.2% over the test cases. However, because of the small sample size of 308 test cases, an accuracy of at least 68.8% would be necessary to confirm a statistically significant improvement in classification accuracy over the baseline of 63.3%.

However, a favorable tradeoff in recall values was attained with no loss in accuracy, as shown in Table 5.3. For example, the perfect indicator achieved a non-culminated recall of 52.2%, in contrast to the 0.0% non-culminated recall achieved by the baseline.

As an additional evaluation of the classification performance of individual indicators, each indicator was evaluated as to its ability to improve non-culminated F-measure. Table 5.4 lists the four indicators that achieved some increase in non-culminated F-measure over the training cases. For each indicator, this table shows classification performance, using the threshold established during training, over the test cases.

Individual indicators achieved an increase in culminated recall, as well as overall accuracy, with no loss in non-culminated F-measure. For example, the perfect indicator achieved a non-culminated F-measure of 54.8, which is comparable to the baseline's F-measure of 53.7, while achieving an improved culminated recall of 61.5%, in contrast to the baseline's culminated recall of 0.0%.

5.5 Combining Indicators with Machine Learning

Three machine learning methods were employed to combine the fourteen indicators: decision trees, genetic programming and log-linear regression. These methods, as well as experimental results, are further detailed in Chapter 6, "Machine Learning for Combining Linguistic

indicator	overall accuracy	Culminated recall	Culminated precision	Non-culminated recall	Non-culminated precision	Non-culminated **F-measure**
perfect	62.0%	61.5%	74.1%	62.8%	48.6%	54.8
temporal adverb	54.5%	37.9%	79.6%	83.2%	43.7%	57.3
pres tense	51.9%	35.4%	75.8%	80.5%	41.9%	55.2
continuous adverb	39.6%	7.2%	73.7%	95.6%	37.4%	53.7
baseline	36.7%	0.0%	100.0%	100.0%	36.7%	53.7

Table 5.4: Classification performance of individual indicators, optimized for non-culminated F-measure.

training set:				**which verbs:**	**aspectual distinction:**	**results section:**
instances	**constituents:**					
	verb only			all **events**	completedness	5.5.1
	multiple	**knowledge source:**				
		linguistic indicators	**open-class constituents:**			
			lexical item (direct obj)	infrequent	completedness	5.6.1
				all **events**	completedness	5.6.1
			Wordnet cat (direct obj)	all **events**	completedness	5.6.2
verbs	**supervised learning:**					
	yes	**test set:**				
		non-overlapping set		all **events**	completedness	5.5.2
		instances		all **events**	completedness	5.5.2
		non-overlapping instances		reduced set	completedness	5.5.2

Table 5.5: Measuring linguistic indicators and evaluating methods to classify according to completedness; the experiments reported in this chapter.

Indicators." In this section, I preview the overall results, illustrating that machine learning allows us to take full advantage of the wide range of indicators by coordinating and combining them.

Classification performance can be designed and evaluated according to a variety of objectives. Classification is most frequently optimized for overall classification accuracy, which is an important criterion for many applications. However, some applications weigh certain classes more heavily in which case other optimization crteria are more appropriate. One example is the F-measure, computing a tradeoff between the recall and precision of a designated class. Alternatively, the tradeoff obtained between recall and precision can be examined in detail. These issues and their relevance to applications of aspectual classification were introduced in detail in Section 3.5.

Table 5.5 summarizes the various methods for evaluating combined lingusitic indicators employed in this Chapter. The first column of Table 5.5 distinguishes between two types of training sets: a set of verb instances (i.e., clauses), and a set of (unique) verbs. When training over instances, the second column distinguishes between classifying according to indicators measured over the main verb only, versus incorporating multiple constituents of a clause to classify it. The central results of this thesis are measured over the first of these two options, "verb only."

Variations on the basic approach of verb-only indicators augment our evaluation of linguistic indicators. Continuing with a depth-first traversal of Table 5.5's hierarchy, classification is performed according to indicators measured over a set of constituents. This method, detailed in the section following this one, is further distinguished according to whether indicators are measured over the head of the direct object directly, or measured over the Wordnet semantic category of this noun.

An additional variation is applied when training over a set of unique verbs. Note that there is no way to incorporate multiple constituents in this case, since instead of classifying individual clauses, we are classifying, or *grouping*, individual verbs. As shown, the results of supervised learning are evaluated over three types of test sets that are further detailed below.

In this Section, training is performed across verb instances, which is also the approach taken in the previous chapter, "States vs. Events in Natural Language," and taken thus far in this chapter. Then I contrast these results with those obtained by training over a verb set, i.e., no duplication.

5.5.1 Training over Verb Instances

In this section we evaluate the linguistic indicators when combined to optimize two distinct criteria. First, indicators are combined in order to increase accuracy. The results of this approach are examined in detail according to the tradeoff in recall values achieved. Then we evaluate the improvement of **non-culminated** F-measure, which combines the **non-culminated** recall and precision values.

	overall accuracy	Culminated recall	Culminated precision	Non-Culminated recall	Non-Culminated precision
log-linear	70.5%	83.1%	73.6%	48.7%	62.5%
GP (4 runs)	68.6%	77.3%	74.2%	**53.6%**	57.8%
decision tree	68.5%	86.2%	70.6%	38.1%	61.4%
Baseline A	63.3%	100.0%	63.3%	0.0%	100.0%
Baseline B	49.0%	46.4%	63.3%	**53.6%**	36.7%

Table 5.6: Comparison of three learning methods, optimized for accuracy, and two performance baselines.

Optimizing Accuracy

Log-linear regression successfully combined indicator values, improving classification accuracy over the baseline measure. As shown in Table 5.6, the log-linear's accuracy was 70.5%. A binomial test shows this is a significant improvement over the 63.3% accuracy achieved by Baseline A. The accuracies attained by genetic programming and decision tree induction, 68.6% and 68.5% respectively, are also higher than that of Baseline A. However, a binomial test shows that these values are not significantly higher than the baseline. This may be due to our small test sample size. Therefore, a larger set of manually annotated data would be required to verify the significance of the increases in accuracy.

The increase in the number of **non-culminated** clauses correctly classified, i.e. non-culminated recall, illustrates an even greater improvement over the baseline. As detailed in Table 5.6, non-culminated recalls of 48.7%, 53.6% and 38.1% were achieved by the three learning methods, as compared to the 0.0% non-culminated recall achieved by Baseline A. Baseline A does not classify any non-culminated clauses correctly because it classifies all clauses as **culminated**. This difference in recall is more dramatic than the accuracy improvement because of the dominance of culminated clauses in the test set. Note that it is possible for an *uninformed* approach to achieve the same non-culminated recall as genetic programming, 53.6%, by arbitrarily classifying 53.6% of all clauses as **non-culminated**, and the rest as **culminated**. However, as shown in Table 5.6, the average performance of such a method (*Baseline B*) loses in comparison to genetic programming, e.g., in overall accuracy (49.0%) and non-culminated precision (36.7%).

Optimizing Non-culminated F-measure

Genetic programming achieved an improvement in non-culminated F-measure over baseline performance. The detailed average performance over five runs of the genetic algorithm optimizing for the F-measure is shown in Table 5.7. Our system achieved an F-measure of 61.9, in contrast to the best possible F-measure by an uninformed approach (Baseline C), 53.7, obtained by classifying every instance as **non-culminated**. In addition, our system's accuracy of 69.3% surpasses the baseline's accuracy of 36.7%.

This batch of genetic algorithm runs achieved an even further tradeoff between culmi-

	overall	Culminated		Non-culminated		
	accuracy	recall	precision	recall	precision	**F-measure**
GP (5 runs)	69.3%	69.9%	79.1%	68.1%	56.8%	61.9
Baseline C	36.7%	0.0%	100.0%	100.0%	36.7%	53.7

Table 5.7: Comparison of genetic programming and a performance baseline, optimized for non-culminated F-measure.

nated and non-culminated recall beyond that achieved by the runs optimizing for accuracy. As shown in Table 5.7, the average non-culminated recall was 68.1%.

These results present an advantage for applications that weigh the identification of non-culminated clauses more heavily than that of culminated clauses. For example, a prepositional phrase denoting a duration with *for*, e.g., *"for a minute,"* describes the duration of a non-culminated event, e.g., *"She ran for a minute,"* or the duration of the state that results from a culminated event, e.g., *"She left the room for a minute."* That is, correctly identifying the use of *for* depends on identifying the aspectual class of the clause it modifies. A language understanding system that incorrectly classifies *"She ran for a minute"* as **culminated** will not detect that *"for a minute"* describes the duration of the *run* event. If this system, for example, summarizes the duration of events, it is particularly important to correctly classify non-culminated events. In this case, our approach is advantageous.

5.5.2 Training over a Verb Set

The previous sections, as well as the entirety of Chapter 4, "States vs. Events in Natural Language," have evaluated classification performance of linguistic indicators across clauses, i.e., verb instances. Likewise, the training examples used to combine indicators with machine learning were sets of clauses in which an individual verb may occur multiple times as the main verb.

A contrasting approach is to train and test across a unique set of verbs, in which each verb only occurs once. In this case, we are simply grouping lexical items into their aspectual categories. Training and testing across verb sets complements our evaluation of the linguistic indicators in several ways:

- Fewer supervised examples must be generated.
- Non-intersecting training and test verb sets are possible. This way, the system to learns how to combine indicators over one set of verbs, and then is evaluated over a completely different set of verbs.
- Both training and testing may be more domain-independent, since verb frequencies vary between domains.
- The resulting lexicon can be evaluated over word sets.

	overall accuracy	Culminated recall	Culminated precision	Non-Culminated recall	Non-Culminated precision
log-linear	68.3%	78.7%	74.6%	48.2%	54.0%
decision tree	64.0%	78.7%	70.2%	35.7%	46.5%
baseline	65.9%	100.0%	65.9%	0.0%	100.0%

Table 5.8: Comparison of two learning methods and a performance baseline, evaluated over a set of 164 verbs.

In this section, training over a verb set is evaluated in three ways. First, evaluation is done over a non-intersecting test set of verbs. Next, evaluation is over a test set of verb instances, with some overlap. Last, evaluation is over a non-intersecting test set of verb instances (no overlap). Each of these sets are detailed below.

Evaluation over a Verb Set

The sets of verb instances discussed thus far were processed in order to create a training set of verbs, and a non-intersecting test set of verbs. This way, learning is applied over a group of verbs, and evaluation is performed over a completely different group verbs. First, multiple occurrences of each verb were collapsed into one training or test case, using the most frequent classification (**culminated** or **non-culminated**) as the classification of that verb. Verbs with an equal number of instances of each class, i.e., fully ambiguous, were eliminated. Half the verbs occurring in both the training and test sets, were eliminated from each set, ensuring that there was no overlap. This process resulted in a training set of 164 verbs (96 **culminated**), and a non-intersecting test set also of 164 verbs (108 **culminated**).

A favorable tradeoff in recall values was attained, with comparable accuracy. As shown in Table 5.8, log-linear regression correctly classified 48.2% of the non-culminated clauses on average, and a decision tree classified 35.7% correctly, both improving over the baseline's non-culminated recall of 0.0%.

The accuracy attained by log-linear regression was 68.3%, as shown in Table 5.8. Although this accuracy is higher than that of the baseline, based on a binomial test this is not significant, which may be due to our small sample size.

Evaluation over Verb Instances

The complete test set of 308 verb instances was also used to evaluate training over a verb set. This way, classification performance is measured such that the true frequency distribution of verbs is accounted for, while retaining the above-mentioned advantages of manually marking a verb set. In this case a new training set of verbs was created from the original training set of verb instances. To do this, multiple occurrences of an individual verb were collapsed into one, with the class (culminated or non-culminated) that dominated *over the training*

	overall accuracy	Culminated recall	Culminated precision	Non-Culminated recall	Non-Culminated precision
log-linear	68.2%	83.6%	71.2%	41.6%	59.5%
decision tree	71.4%	83.1%	74.7%	51.3%	63.7%
Baseline A	63.3%	100.0%	63.3%	0.0%	100.0%

Table 5.9: Comparison of two learning methods and a performance baseline, evaluated over verb instances. Compare to Table 5.6

	overall accuracy	Culminated recall	Culminated precision	Non-Culminated recall	Non-Culminated precision
log-linear	67.2%	81.5%	71.0%	42.5%	57.1%
decision tree	59.1%	69.2%	67.2%	41.6%	43.9%
Baseline A	63.3%	100.0%	63.3%	0.0%	100.0%

Table 5.10: Comparison of two learning methods and a performance baseline, evaluated over verb instances with unseen verbs. Compare to Table 5.10

cases only intact (verbs with equal occurrences of each class were eliminated). This ensures that the test cases are not used to determine the class of verbs in the training cases. This process resulted in a training set of 199 verbs.

In this experiment, some verbs occurring in the test set also appear in the training set, though not necessarily with the same class. However, the test set remains an unbiased distribution of examples from the corpus that is therefore representative of verb ambiguity, and that did not influence training.

In this experiment, decision tree induction attained the highest accuracy reported in this chapter over the full set of 308 test cases, 71.4%, which a binomial test shows to be significantly greater than that of Baseline A. These results are shown in Table 5.9. The average accuracy attained by log-linear regression was 68.2%. Although this accuracy is higher than that of the baseline, based on a binomial test this is not significant, which may be due to our small sample size.

A favorable tradeoff in recall values was attained, with comparable accuracy. As shown in Table 5.8, log-linear regression correctly classified 41.6% of the non-culminated clauses on average, and a decision tree classified 51.3% correctly, both improving over the baseline's non-culminated recall of 0.0%.

Non-overlapping Training Set

To evaluate the linguistic indicators and their combination with machine learning over unseen verbs, but taking into account the frequencies of individual verbs, the training set of 199 verbs in the previous section was further reduced by eliminating verbs that occur in the test set. This resulted in a training set of 131 verbs of which 76 were **culminated**.

A favorable tradeoff in recall values was attained, with comparable accuracy. As shown in Table 5.10, log-linear regression correctly classified 42.5% of the non-culminated clauses on average. improving over the baseline's non-culminated recall of 0.0%. The accuracy attained by log-linear regression was 67.2%, which is comparable to that obtained when training over the entire set of 307 verb instances (see Table 5.6). Although this accuracy is higher than that of the baseline, based on a binomial test this is not significant, which may be due to our small sample size.

Decision tree induction overfitted on this small set of 131 training cases. This is further discussed in Section 6.3.3.

5.6 Indicator Values over other Clausal Constituents

Incorporating other clausal constituents in addition to the main verb contributes to aspectual classification in two respects. First, clauses with rare or infrequent verbs pose the difficulty that there may be too few occurrences over which to measure linguistic indicators. In this case, the indicators measured over other clausal features may be more reliable, since several features besides the main verb correlate with aspectual class. Second, since many verbs are aspectually ambiguous, it is often necessary to account for constituents other than the main verb. One approach to this end is to measure indicator values across multiple constituents, as described in Section 3.4.

This section provides experimental results supporting these hypotheses. First, indicator values measured over the direct object are shown to correlate with aspectual class. Further, it is shown that classification performance achieved using these values is superior to that of the *verb*-indicators over clauses with infrequent verbs. Second, preliminary results show advantages to measuring indicators over multiple constituents in order to resolve verb ambiguity.

Training and validation in this section is performed over clauses, i.e., instances of verbs. This way, results reflect the distributions of verbs over real data. Further, this is necessary for additional clausal constituents to be considered in order to classify clauses with ambiguous or infrequent verbs.

5.6.1 Direct Object

The direct object of a clause influences its aspectual class [Pustejovsky, 1991, Resnik, 1996, Olsen and Resnik, 1997]. This has been supported linguistically with examples (see Section 2.1.4), and empirically, by using the Wordnet category of direct objects to classify *have*-clauses (Section 4.3).

The 14 indicators measured thus far over the main verb can alternatively be measured over other clausal constituents, e.g., the direct object. In particular, the indicator values corresponding to an input clause are computed by measuring the frequency with which the direct object occurs with each indicating phenomenon, e.g., the perfect tense, across the corpus of parsed clauses.

clausal constituent	overall accuracy	Culminated recall	Culminated precision	Non-culminated recall	Non-culminated precision	Non-culminated F-measure
main verb	52.9%	59.7%	60.0%	43.2%	42.9%	43.0
direct object	51.9%	32.5%	69.4%	**79.6%**	45.3%	57.7
Baseline E	44.8%	20.4%	59.3%	**79.6%**	41.2%	54.3
Baseline D	41.2%	0.0%	100.0%	100.0%	41.2%	58.4

Table 5.11: Comparison of indicators measured over main verb vs. over direct object, combined with genetic programming, evaluated over 114 instances of infrequent verbs, optimized for **non-culminated** F-measure.

In this case, the main verb is not used when computing the aspectual class of a clause. However, since one advantage of incorporating clausal features other than the main verb is to deal with infrequent verbs, this procedure could be applied in a system that classifies according to the main verb for frequent verbs, but uses the direct object to classify clauses with infrequent verbs. Additionally, the computation of indicators over multiple constituents is evaluated below in Section 5.6.2.

In this section I first show that indicators measured over the direct object outperform those measured over the main verb for clauses with infrequent verbs. Then I evaluate indicators measured over the direct object across clauses with main verbs of any and all frequencies, showing that these indicators outperform baseline measures.

Evaluation over Infrequent Verbs

To evaluate the linguistic indicators and their combination with machine learning over infrequent verbs, training and test sets were created from the original sets by eliminating each clause with a main verb that occurs more than 50 times as a main verb across the corpus of 75,289 parsed clauses. This way, all training and testing occurs on clauses with rare verbs. This resulted in a testing set of 114 clauses (67 **culminated**).

These results are from combining a smaller, preliminary subset of 7 indicators: progressive, perfect, present tense, past/present participle, duration *in*-PP, duration *for*-PP, and *temporal* adverb.

For clauses with rare verbs, indicators measured over the direct object outperformed baseline methods as well as indicators measured over the main verb. Table 5.11 shows the results of 2 batches of genetic programming runs, each optimizing for **non-culminated** F-measure. The first batch, combining indicator values measured over the main verb, achieved a **non-culminated** F-measure of 43.0. This was surpassed by the second batch, combining indicator values measured over the direct object, which achieved a score of 57.7. Although Baseline D, optimized for **non-culminated** F-Measure, meets this performance with a score of 58.4, a comparison to Baseline E, which matches direct object indicators in non-culminated recall (79.6%), does not do as well in all other performance measurements shown. Therefore, direct object indicators achieve a more favorable tradeoff in recall values

	overall	Culminated		Non-Culminated	
	accuracy	recall	precision	recall	precision
GP (7 runs)	62.7%	82.9%	66.5%	**27.8%**	48.6%
Baseline A	63.3%	100.0%	63.3%	0.0%	100.0%
Baseline F	55.9%	72.2%	63.3%	**27.8%**	36.7%

Table 5.12: Comparison of genetic programming and two performance baselines, optimized for accuracy, combining indicators computed over the direct object.

	overall	Culminated		Non-culminated		
	accuracy	recall	precision	recall	precision	**F-measure**
GP (6 runs)	55.1%	45.0%	73.8%	**72.4%**	43.3%	54.2
Baseline C	36.7%	0.0%	100.0%	100.0%	36.7%	53.7
Baseline G	44.0%	27.6%	63.3%	**72.4%**	36.7%	48.7

Table 5.13: Comparison of genetic programming and two performance baselines, optimized for non-culminated F-measure, combining indicators computed over the direct object.

than any uninformed baseline approach.

These results provide evidence that classifying clauses with infrequent verbs according to indicator values corresponding to the direct object improves over an uninformed approach, and is a more judicious approach than classifying caccording to indicator values computed over the main verb.

Evaluation over All Instances

For further evaluation, the classification performance of indicators measured over the direct object was evaluated over all instances by training and testing over the entire set of clauses. In this case, we are not restricting ourselves to clauses with infrequent verbs. Although the main verb is the primary factor in assessing aspectual class, showing that the direct object also informs aspectual classification provides evidence that it can be used for infrequent verbs and provides empirical data to complement our linguist intuition that the direct object influences the aspectual class of a clause.

In these experiments, values for infrequent objects, those occurring less than 10 times in the corpus, are computed over the entire corpus of parsed clauses.

When training and testing over all instances, indicator values measured over the direct object attained more favorable recall tradeoffs than baseline methods, both when optimizing for accuracy, and when optimizing for **non-culminated** F-measure (using genetic programming). As shown in Table 5.12, when optimizing for accuracy, the average non-culminated recall was 27.8%. Baseline F, which matches this non-culminated recall, does not perform as well in non-culminated precision nor in accuracy. Likewise, Table 5.13 shows that when optimizing for **non-culminated** F-measure, a further tradeoff in recall values is attained.

clausal features	runs	overall accuracy	Culminated recall	Culminated precision	Non-Culminated recall	Non-Culminated precision
verb	4	68.6%	77.3%	74.2%	**53.6%**	57.8%
verb, obj class	6	70.9%	80.4%	75.3%	54.4%	61.7%
verb, obj class , prep	4	71.2%	79.1%	76.3%	57.5%	61.5%
	Baseline A	63.3%	100.0%	63.3%	0.0%	100.0%
	Baseline B	49.0%	46.4%	63.3%	**53.6%**	36.7%

Table 5.14: Comparison of indicators computed over three clausal feature sets, combined with genetic programming, optimizing for accuracy.

5.6.2 Multiple Constituents

Since the aspectual class of a clause is determined by several clausal constituents in addition to the main verb, it is important that a classification system incorporate multiple constituents. This is especially true for clauses with a main verb that is aspectually ambiguous. In addition to the main verb and direct object, both examined thus far, we will also incorporate the adjunct preposition, which has been shown linguistically to influence aspectual class. These issues are detailed in Section 2.1.

In this section we evaluate indicators measured over multiple constituents, comparing the results to those of indicators measured over the main verb only. First, classification is optimized for accuracy, and then optimized for **non-culminated** F-measure. The methods by which indicators are computed over multiple clausal constituents is detailed in Section 3.4.

These results provide evidence that multiple constituents help resolve the difficulty of classifying clauses headed by aspectually ambiguous verbs. Although more sparsely measured, the performance achieved by indicators measured over multiple constituents is comparable to that of indicators measured over the verb only (for all three learning methods), with some evidence of improvement. Therefore, the noise introduced by more sparsely measuring indicators over multiple constituents is compensated for by the ability to resolve aspectually ambiguous verbs. Future work measuring performance over larger testing sets of clauses, and using larger parsed corpora to more accurately measure indicators over multiple constituents, may more clearly demonstrate improvements when measuring indicators over multiple constituents.

Optimizing Accuracy

When combined using genetic programming, indicators computed over the main verb, direct object class and adjunct preposition achieved the highest accuracy, when training over verb instances, reported in this chapter, 71.2%. Overall, the classification performance of indicators measured over multiple constituents is comparable to that of indicators measured over the main verb only, as shown in Table 5.14. Both accuracy and non-culminated recall improved over verb-only indicators when measuring indicators over the main verb and the direct object class, and further improve when also incorporating the adjunct preposition,

clausal features	overall accuracy	Culminated		Non-Culminated	
		recall	precision	recall	precision
verb	68.5%	86.2%	70.6%	38.1%	61.4%
verb, obj class	71.1%	81.5%	75.0%	53.1%	62.5%
verb, obj class, prep	66.6%	75.4%	72.8%	51.3%	54.7%
Baseline A	63.3%	100.0%	63.3%	0.0%	100.0%

Table 5.15: Comparison of indicators computed over three clausal feature sets, combined with decision tree induction, optimizing for accuracy.

clausal features		overall accuracy	Culminated		Non-culminated		
			recall	precision	recall	precision	**F-measure**
verb (min=1)	GP (7 runs)	69.0%	69.9%	78.7%	67.4%	56.5%	61.4
verb (min=10)	GP (12 runs)	69.3%	66.6%	81.5%	74.0%	56.2%	63.8
verb obj prep	GP (6 runs)	64.3%	54.4%	83.6%	**81.6%**	50.9%	62.7

Table 5.16: Comparison of two clausal feature sets, optimized for non-culminated F-measure. Here, genetic programming combines a reduced set of 7 linguistic indicators.

but these improvements can not be proven statistically significant due to the limited size of our testing set. Note that both the duration *in*-PP and duration *for*-PP indicators are measured over the verb only, for all three batches of GP runs to compensate for how infrequently they occur in the corpus, as compared to other indicators.

Indicators measured over the main verb and direct object class achieved a more favorable recall tradeoff than those measured over the verb only, with comparable performance in accuracy. For example, as shown in Table 5.15, indicators measured over the main verb and direct object class and combined with a decision tree achieved a non-culminated recall of 53.1%, as compared to 38.1% achieved by the verb-only indicators. This relative improvement also occurred when combining with log-linear regression.

Optimizing F-measure

The experiments evaluating indicators measured over multiple constituents, optimizing for **non-culminated** F-measure, differ in two ways. First, these are results from combining a smaller, preliminary subset of 7 indicators: progressive, perfect, present tense, past/present participle, duration *in*-PP, duration *for*-PP, and *temporal* adverb. Second, both the duration *in*-PP and duration *for*-PP indicators are measured over the verb only to compensate for how infrequently they occur in the corpus, as compared to other indicators.

The most dramatically favorable tradeoff in recall reported in this chapter was achieved by indicators computed over multiple constituents. These indicators were combined with genetic programming, optimizing for **non-culminated** F-measure. As shown in Table 5.16, indicators computed over the verb, object, and preposition achieved a non-culminated recall of 81.6%, which improves over that achieved by indicators computed over the verb only, 74.0%, with comparable performance in **non-culminated** F-measure (62.7).

Two comparison batches of GP runs are shown in Table 5.16, both measuring indicator values over the main verb only. The first computes indicator values in the usual way, while the second computes values for infrequent verbs (less than 10 occurrences) over the entire corpus, achieving a more favorable recall tradeoff.

5.7 Conclusions

When used in combination, linguistic indicators significantly improve the classification of an unrestricted set of verbs according to completedness. A total of fourteen indicators, each measured automatically across a corpus of text, were combined with machine learning. The best method, log-linear regression, successfully combined indicator values, achieving a classification accuracy of 70.5%, in comparison to 63.3% achieved by the baseline.

This improvement in classification performance is more dramatically illustrated by a favorable tradeoff between culminated and non-culminated recall, as well as an improvement in non-culminated F-measure, each achieved by the learning methods. These improvements are profitable for tasks that weigh the identification of non-culminated events more heavily than culminated events. For example, a non-culminated recall of 53.6% was achieved by genetic programming, as compared to the 0.0% non-culminated recall achieved by the baseline. Additionally, genetic programming achieved an improvement in non-culminated F-measure, achieving a score of 61.9, in contrast to 53.7 achieved by the best uninformed baseline.

Evaluating indicators individually, this analysis has revealed correlations between verb class and six indicators, four of which are not traditionally related to completedness on a linguistic basis. Three were not previously hypothesized to predict any aspectual distinction: (1) occurrences modified by a *manner* adverb, (2) occurrences in the past or present participle, and (3) occurrences with no deep subject. Furthermore, occurrences in the progressive voice was not previously hypothesized to predict completedness in particular.

Additional tests confirm that linguistic indicators are viable for a range of evaluatory criteria. When training over a set of verbs, evaluation was performed over the full test set of verb instances, as well as a test set of instances that excludes any verbs that occurred in the training set. Results were favorable, including, e.g., a 42.5% non-culminated recall over the latter test set, with no decrease in accuracy over the baseline.

Finally, further tests confirm the viability of measuring linguistic indicators over multiple clausal constituents. This is beneficial for resolving verb ambiguity, and for alleviating verb sparsity. For example, one experiment showed an increase in non-culminated recall to 81.6% (the highest of all tests in this chapter) by measuring linguistic indicators over multiple constituents. In this case, indicators were measured over the verb, direct object Wordnet category, and preposition, improving performance over the verb-only, which achieved 67.4% non-culminated recall.

Chapter 6

Machine Learning for Combining Linguistic Indicators

In the previous two chapters we explored the utility of individual indicators for aspectual classification. Despite a positive correlation between indicator values and aspectual class, the classification performance of individual indicators is limited. Therefore, in this chapter we will explore the ability of machine learning to automatically combine multiple indicators to improve aspectual classification.

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 was illustrated in Chapter 4, "States vs. Events in Natural Language," by showing that some indicators help for only a subset of verbs, a phenomenon visualized as *outliers* on histograms. 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:

- As discussed in Section 2.1.1, while the *progressive* indicator is linguistically linked to atomicity, it is only indirectly linked to completedness. Its usefulness for predicting whether a verb is **culminated** or **non-culminated** may be due to the fact that non-extended (i.e., **atomic**) verbs are more likely to be **culminated** than **extended**.
- As discussed is Section 3.2.2, many ambiguous `location` verbs can appear with the progressive, even in their stative sense, e.g. *"The book was lying on the shelf."*
- 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****."*
- Some indicators do not measure phenomena that is linguistically constrained by any aspectual category, such as *present tense*, durative *for*-PPs, `frequency` and *not/never*.

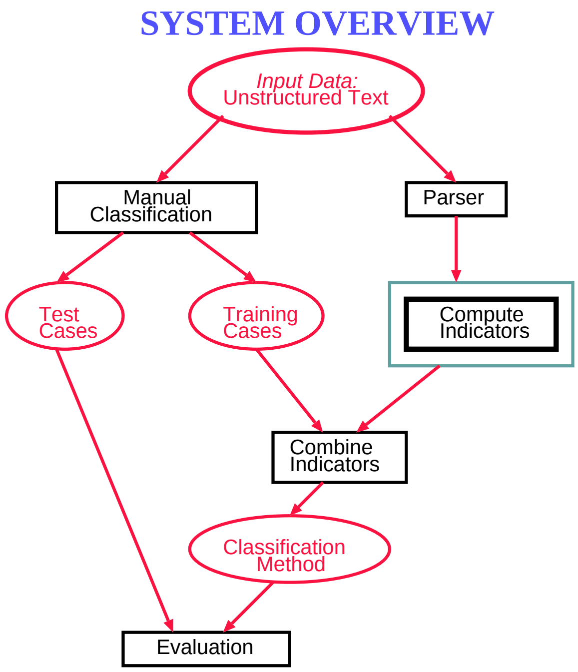


Figure 6.1: System overview.

The predictiveness of these indicators is therefore incomplete in that only the subset of verbs that adhere to the respective trends can be correctly classified. Therefore, such incomplete indicators may complement one another when placed in combination. Our goal is to take full advantage of the complete range of indicators by coordinating and combining them.

Machine learning methods can be employed to automatically generate a model that will combine indicator values. Figure 6.1 shows a system overview which facilitates this process. This diagram outlines a generic system which combines numerical indicators with machine learning for a classification problem, in natural language understanding or otherwise. Indicators are computed over an automatically parsed corpus. Then, in the "Combine Indicators" box, supervised training cases are used to automatically generate a model ("Classification Method") with supervised machine learning. This model is then evaluated over unseen supervised test cases.

There are four advantages to automating this process with machine learning:

- The cost of *manually* generating a model, which is often prohibitive, is avoided.
- Biases introduced by a human engineer may be suboptimal.
- Automated approaches are extensible to multiple natural language classification problems, across multiple domains and multiple languages.
- Resulting models often reveal linguistic insights.

corpus	3,224 medical discharge summaries	10 novels
classification problem	stativity	completedness
corpus size	1,159,891 words	846,913 words
parsed clauses	97,973	75,289
training clauses	739 (634 events)	307 (196 culminated)
testing clauses	739 (619 events)	308 (195 culminated)
verbs in test set	222	204
clauses excluded	*be*-, *have*-clauses	stative clauses
unsupervised results	N/A	stativity

Table 6.1: Two classification problems on different data sets.

In this chapter, we perform experiments with three supervised learning methods over two corpora for two classification problems. The two corpora, detailed in the previous two chapters, are summarized in Table 6.1 (available at `http://www.cs.columbia.edu/~evs/VerbData/`). Indicators computed over the corpus of medical discharge summarizes are combined to classify clauses according to stativity, and indicators computed over the corpus of novels are combined to distinguish between **culminated** and **non-culminated** clauses. For each classification problem, we perform three learning methods: log-linear regression, decision tree induction, and genetic programming. The resulting models generated by these methods are examined in detail to compare classification performance, and to detect linguistically valid insights from the models. Furthermore, we examine training performance to detect whether models have *overfit* to the training data.

We also examine an *unsupervised* method which groups verbs according to meaning. This method is generic in that it is not designed to distinguish according to any particular classification scheme. However, when performed over the corpus of novels, it is shown to produce groups that distinguish verbs according to stativity.

6.1 Related Work

The use of machine learning techniques to combine numerical indicators for classification problems is a well-established technique. This includes work with decision trees [Fayyad and Irani, 1992], log-linear regression [Santner and Duffy, 1989], and genetic programming [Koza, 1992, Tackett and Carmi, 1994]. Applications include document classification [Masand, 1994], image classification [Tackett, 1993], and stock market prediction [Allen and Karjalainen, 1995].

Machine learning has become instrumental in the development of robust natural language understanding systems. Using machine learning, models for integrating linguistic, semantic and statistical knowledge sources can be developed automatically. For example, decision tree induction has been applied to word sense disambiguation [Black, 1988, Siegel, 1994, Siegel and McKeown, 1994, Litman, 1994], determiner prediction [Knight, forthcoming], coordination parsing [Resnik, 1993], and syntactic parsing [Magerman, 1993]. An overview of psycho-linguistic issues behind learning for natural language problems in particular is given by Powers and Turk [1989]. Models resulting from machine induction can be manually inspected to discover linguistic insights [Siegel and McKeown, 1994]. The Appendix of this thesis provides a complete survey of work applying evolutionary computation (e.g. genetic algorithms) to natural language problems. However, machine learning techniques have not yet been applied to aspectual disambiguation.

Previous efforts have also applied machine induction methods to coordinate corpus-based linguistic indicators in particular, e.g., to classify adjectives according to markedness [Hatzivassiloglou and McKeown, 1995], to perform accent restoration [Yarowsky, 1994], for disambiguation problems [Yarowsky, 1994, Luk, 1995], and for the automatic identification of semantically related groups of words [Pereira et al., 1993b, Hatzivassiloglou and McKeown, 1993b, Schűtze, 1992].

6.2 Learning Methods

Given a verb and its 14 indicator values, our goal is to use all 14 values in combination to classify the verb as a state or an event. Machine learning methods can be employed to automatically generate a model that will combine these values. For example, log-linear regression creates a set of weights so that a weighted sum of the values can be used to classify the verbs. Other learning methods create more complex models. Once a function for combining indicator values has been established, previously unobserved verbs can be automatically classified according to their indicator values.

In this section, we briefly overview four learning methods. Three of these methods, log-linear regression, decision tree induction and a genetic algorithm, are *supervised*, which means they require training data, i.e. examples that have been manually classified. This corresponds to the "Combine Indicators" box in Figure 6.1. Each of these three mothods is evaluated for classification according to both stativity and completedness.

The fourth method is *unsupervised*, not requiring manually supervised training data. This method automatically clusters verbs into several groups according to their co-occurrence distributions with direct objects. This method is performed over the corpus of novels, and the resulting groups are shown to correlate with stativity.

6.2.1 Log-Linear Modeling

As suggested by Klavans and Chodorow [1992], a weighted sum of multiple indicators that results in one "overall" indicator may provide an increase in classification performance. This method embodies the intuition that each indicator correlates with the probability that a verb belongs in a certain class, but that each indicator has its own unique scale, and so must be weighted accordingly. One way to determine these weights is log-linear regression [Santner and Duffy, 1989], a popular technique for binary classification. This technique, which is more extensive than a simple weighted sum, applies an inverse logit function, and employs the iterative reweighted least squares algorithm [Baker and Nelder, 1989].

Table 6.2: Default decision tree induction parameters implemented by Splus.

Minimum partition size before first split:	5
Minimum partition size for additional splits:	10
Node selection criterion:	deviance = -2 times log-likelihood
Node purity:	deviance < .01

The Splus statistical package was used for the induction process, with parameters set to their defaults.

6.2.2 Decision Tree Induction

Another method capable of modeling non-linear relationships between indicators is a decision tree. It may be necessary to model non-linear interactions between indicators, since, for example, indicators may help classify overlapping, yet distinct, subsets of verbs.

An example tree is shown in Figure 6.4. Each internal node of a decision tree is a choice point, dividing an individual indicator into two ranges of possible values. Each leaf node is labeled with a classification (state or event). Given the set of indicator values corresponding to a verb, that verb's class is established by deterministically traversing the tree from the root to a leaf.

The most popular method of decision tree induction, which we employ here, is recursive partitioning [Quinlan, 1986, Breiman et al., 1984]. This method "grows" a decision tree by expanding it from top to bottom. At each level of the tree, an indicator and threshold is selected such that the examples are best distinguished according to aspectual class. The criterion optimized for each split is deviance, implemented as minus twice the log-likelihood. Table 6.2 shows the parameters used to control decision tree growth. The Splus statistical package was used for the induction process.

6.2.3 Genetic Programming

An alternative method that also implements non-linear combinations of indicators is to generate function trees that combine them. A popular method for generating such function trees is a genetic algorithm [Holland, 1975, Goldberg, 1989], which is modeled after population genetics and natural selection. The use of genetic algorithms to generate function trees [Cramer, 1985, Koza, 1992], often called genetic programming.

Inspired by Darwinian survival of the fittest, the GA works with a pool of initially random hypotheses (in this case, function trees), stochastically performing genetic recombination and mutation to propagate or create better individuals, which replace old or less good individuals. Recombination between function trees usually consists of selecting a subtree from each individual, and swapping them, thereby creating two new individuals [Cramer, 1985, Koza, 1992]. For random mutation, a randomly chosen subtree can be replaced by a new randomly created subtree [Koza, 1992], or a randomly chosen node can be randomly altered. Because the genetic algorithm is stochastic, each run may produce a different func-

Table 6.3: Applying genetic programming to induce a function tree that combines linguistic indicators.

Objective:	Function tree to combine linguistic indicators
Terminal set:	14, corresponding to the set of linguistic indicators.
Function set:	ADD, MULTIPLY and DIVIDE
Training cases:	739 (stativity) or 307 (completedness) verb instances.
Raw fitness:	Accuracy or negative F-measure over fitness cases, when best threshold is selected.
Parameters:	Number of generates = 50,000, population size = 500, 5-member tournament selection, steady state population [Syswerda, 1989].
Identification of best of run:	Most fit over training cases.

tion tree. Therefore, performance must be evaluated over the models produced by multiple runs.

The genetic algorithm parameters are shown in Table 6.3. The function trees are generated from a set of 17 primitives: the binary functions ADD, MULTIPLY and DIVIDE, and 14 terminals corresponding to the 14 indicators listed in Table 4.3. This set of primitives was established empirically; other primitives such as conditional functions, subtraction, and random constants failed to improve performance significantly. The polarities for several indicators were reversed according to the polarities of the weights established by log-linear regression for stativity. Runs of the genetic algorithm maintain a population size of 500, and end after 50,000 new individuals have been evaluated.

A threshold must be selected for both linear and function tree combinations of indicators. This way, overall outputs can be discriminated such that classification performance is maximized. For both methods, this threshold is established over the training set and frozen for evaluation over the test set.

6.2.4 Unsupervised Learning

Supervised learning methods require a set of manually marked examples for the training process. This marking process is time consuming, domain-dependent and requires linguistic expertise. For example, as described in Section 5.1.1, to create training and evaluation data for distinguishing according to completedness, events had to be identified with the *"What happened was"* test, and then further distinguished according to a relatively subtle semantic entailment. This process must be repeated on a corpus representing each new domain over which aspectual classification is required in order to produce supervised data. Alternative methods for clustering words have been developed that do not require manually marked examples [Hatzivassiloglou and McKeown, 1993a, Schűtze, 1992].

This section describes an *unsupervised* approach to clustering verbs developed and implemented by Vasileios Hatzivassiloglou, based on previously work for semantically clustering adjectives [Hatzivassiloglou and McKeown, 1993a, Hatzivassiloglou, 1997]. This system automatically places verbs into semantically-related groups based on the distribution of

1. `sell(27) buy(20) acquire(8)`

2. `push(28) pull(45)`

3. `raise(68) lower(24)`

4. `leave(160) enter(78)`

5. `know(164) forget(50) learn(51)`

6. `love(18) hate(16) like(112)`

7. `want(60) need(69) require(57) demand(15)`

Figure 6.2: The set of verbs manually selected for evaluating unsupervised clustering, with frequencies shown. The grouping shown here was established manually.

co-occurring direct objects. Both the number of groups and the number of verbs in each group are determined by the algorithm.

Note that this approach differs from our approach combining linguistic indicators in three significant ways. First, it requires no supervised training examples. Second, the method semantically groups words in a general sense – it is not designed or intended to group words according to any particular classification scheme such as stativity or completedness. Third, this method uses a co-occurrence relation not embodied by any of the 14 linguistic indicators presented in this thesis: the direct object. However, there are several advantages to linguistic indicators that measure the frequency of linguistic phenomena such as the *progressive* over measuring the frequencies of open class words, as described in Chapter 3.

The clustering algorithm was evaluated over the corpus of novels, which, as shown in Table 6.1, has 75,289 parsed clauses. Clauses were eliminated from this set if they had no direct object, or if the direct object was a clause, a proper noun, or a pronoun, or was misspelled. This left 14,038 distinct verb-object pairs of varying frequencies.

Because the direct object is an open-class category (noun), occurrences between any particular verb-object pair are sparse as compared to the markers measured by the linguistic indicators. For example, *make-dinner* occurs only once among the parsed clauses from the corpus of novels, but *"make"* occurs 34 times in the *progressive*. For this reason, the clustering algorithm was evaluated over a set of frequent verbs only. 56 verbs occurred more than 50 times each in the set of verb-object pairs. Of these, the 19 shown in Figure 6.2 were selected as an evaluation set because of the natural semantic groups they fall into. The groupings shown, which do not pay head to aspectual classification in particular, were established manually, but are not used by the automatic grouping algorithm.

Verbs were then automatically grouped according to their co-occurrence distributions with certain direct objects. As revealed in Section 6.3.1, the resulting groups correspond with the stative category of these verbs. For more detail on the clustering algorithm see

Linguistic Indicator	stativity	completedness
class	yes, no	yes, no
manner adverb	0.00..0.29	0.00..0.23
duration *in*-PP	0.00..0.13	0.00..0.12
continuous adverb	0.00..1.00	0.00..0.33
temporal adverb	0.00..0.40	0.00..0.97
"not" or *"never"*	0.47..1.00	0.58..1.00
duration *for*-PP	0.00..0.15	0.00..1.00
perfect	0.00..0.35	N/A
perfect (not progressive)	N/A	0.00..0.50
past/pres participle	0.00..1.00	0.33..1.00
evaluation adverb	0.00..0.46	0.00..0.95
no subject	0.00..1.00	0.00..1.00
past tense	0.00..1.00	0.00..1.00
present tense	0.00..1.00	0.00..1.00
`frequency`	1..2,131	1..13,882
not progressive	0.00..1.00	0.00..0.50

Table 6.4: Schema of inputs for supervised learning. Fourteen continuous attributes for two binary learning problems: stativity and completedness.

Hatzivassiloglou and McKeown [1993a] and Hatzivassiloglou [1997].

6.3 Results

As described in Chapters 4 and 5, supervised data over verb instances was created for two aspectual classification problems: stativity and completedness. In particular, preprocessing and manual marking resulted in training and test sets of 739 examples each for stativity, and 307 and 308 examples, respectively, for completedness. Table 6.4 illustrates the schema of inputs for supervised learning. There are a total of fourteen continuous attributes for two binary learning problems: stativity and completedness. All attributes are proportions except `frequency`, which is a positive integer. This data is available at http://www.cs.columbia.edu/~evs/VerbData/.

In this section, we examine the models produced by the four learning methods, and their classification performance. These methods are applied to combine the linguistic indicators computed over the medical discharge summaries in order to distinguish between states and events. Then, the methods are applied to indicators computed over novels in order to distinguish between culminated and non-culminated events. Details of these two data sets are given in Table 6.1.

In the final portion of this section, we examine certain conditions under which the supervised learning methods overfit to the training data, thus losing generality and performing

	overall	States		Events	
	accuracy	recall	precision	recall	precision
decision tree	93.9%	74.2%	86.4%	97.7%	95.1%
genetic programming	91.2%	47.4%	97.3%	99.7%	90.7%
log-linear	86.7%	34.2%	68.3%	96.9%	88.4%
baseline	83.8%	0.0%	100.0%	100.0%	83.8%

Table 6.5: Comparison of three learning methods, optimizing for accuracy, and a performance baseline, distinguishing states from events.

poorly on the test data.

6.3.1 States vs. Events

As shown in Table 6.5, all three supervised learning methods help improve accuracy, and each achieved a favorable recall tradeoff for distinguishing states from events. A detailed evaluation of these classification performance results is given in Chapter 4. The differences in accuracy between the three methods are each significant ($p < .01$), which highlights the importance of how linear and non-linear interactions between numerical linguistic indicators are modeled. Therefore, it is necessary to inspect these resulting models to contrast the means by which they combine linguistic indicators.

A pilot study showed no further improvement in accuracy or recall tradeoff by five additional learning algorithms: Naive Bayes [Duda and Hart, 1973], Ripper [Cohen, 1995], ID3 [Quinlan, 1986], C4.5 [Quinlan, 1993] and CART [Friedman, 1977, Breiman et al., 1984]. Furthermore, using metalearning to combine multiple learning methods hierachically (a.k.a., stacked generalization) [Chan and Stolfo, 1993, Wolpert, 1992], according to the JAM (Java Agents for Metalearning) model [Stolfo et al., 1997], produced only equivalent results. This may be due to the limited size of our supervised data.

In this section, I compare and contrast the resulting models of the three supervised learning method, as well as the unsupervised approach. In particular, we analyze what makes each model tick, and contrast the ways in which they combine indicators.

Log-Linear Modeling

Log-linear modeling successfully combined the 14 linguistic indicators, achieving an accuracy of 86.7%, as shown in Table 6.5. This is a significant improvement over the baseline accuracy of 83.8%, as measured with a binomial test. Furthermore, a stative recall of 34.2% was achieved.

The particular weighting scheme resulting from log-linear modeling for this data intrinsically implements and integrates a decision tree type rule, along with the expected log-linear weighting. This is illustrated by examining Table 6.6, which shows the weights automatically derived by log-linear regression for each of the 14 linguistic indicators. The values

Linguistic Indicator	log-linear weight	t-test P-value
manner adverb	11.04744	0.1681
duration *in*-PP	0.06209624	0.0018
continuous adverb	-0.04168417	0.8438
temporal adverb	0.02127572	0.0000
"not" or *"never"*	0.01714499	0.0000
duration *for*-PP	-0.01019155	0.8402
perfect	0.009528091	0.0054
past/pres participle	0.006981148	0.0005
evaluation adverb	0.005695407	0.1766
no subject	0.002742867	0.0000
past tense	0.002586572	0.2314
present tense	0.002409898	0.0901
`frequency`	-0.001264895	0.0000
not progressive	-0.0009369231	0.0903

Table 6.6: Weights produced by log-linear regression to distinguish between stative and event verbs. Contrast with Table 6.10.

```
(DIV (ADD (ADD (MUL (DIV DurationInPP (ADD Progressive (ADD NotNever
NotNever))) (ADD Progressive 75)) (DIV (MUL (ADD Progressive PastTense)
Progressive) NegFrequency)) NotNever) DurationInPP)
```

Figure 6.3: Example function tree designed by a genetic algorithm to distinguish between stative and event verbs, achieving 92.7% accuracy.

assigned to the *manner* adverb indicator, 11.04744, far outweighs the other 13 weights. At first glance, it may appear that this weighting scheme favors the *manner* adverb indicator in lieu of all other indicators. However, as shown in Figure 6.7, *manner* adverb indicator values are 0.0% for all verbs in the training set except the 8 indicated, all of which denote events. Therefore, the large weight assigned to the *manner* adverb indicator is only activated for those verbs, which are therefore each classified as events, regardless of the remaining 13 indicator values. For all other verbs, the remaining 13 indicator values determine the classification. Note that this rule was *not* found by decision tree induction, detailed below.

This rule cannot increase accuracy over the baseline without the remaining 13 indicators, since it does not positively identify any **states** – it only identifies **events**, which are all correctly classified by the baseline. Therefore, it is only useful in the context of the overall weighting scheme derived by log-linear regression.

Genetic Programming

Genetic programming successfully combined the 14 linguistic indicators, achieving an accuracy of 91.2%, as shown in Table 6.5. This is a significant improvement over the baseline

if frequency with *manner* adverbs is...	verbs this applies to:	frequency in training set	**then** classify:
$> 0.0\%$	*adjust*	1	as **event**
	continue	16	
	decline	2	
	decrease	4	
	improve	5	
	increase	4	
	progress	2	
	resolve	7	
0.0%	all other verbs	699	depending on other 13 indicators

Table 6.7: Linguistic rule discovered by log-linear regression.

accuracy of 83.8%, according to a binomial test. Furthermore, a stative recall of 47.4% was achieved.

Genetic programming improved classification performance by emphasizing a different set of indicators than those emphasized by log-linear modeling. Figure 6.3 shows an example function tree automatically generated by the genetic algorithm, which achieved 92.7% accuracy. Note that this classification performance was attained with a subset of only five linguistic indicators: duration *in*-PP, progressive, *"not"* or *"never"*, past tense, and `frequency`. Two of these are ranked lowest by log-linear regression: `frequency` and progressive. Furthermore, *manner* adverb, ranked highest by log-linear regression, is not incorporated in this function tree at all. This may be because this indicator only applies to a small number of verbs, as shown in Table 6.7, and because an **if**-rule such as that captured by log-linear regression is difficult to encode with a function tree that has no conditional primitive functions. Overall, we can conclude that many subsets of linguistic indicators are useful for aspectual classification if combined with the correct model.

Decision Tree Induction

Decision tree induction successfully combined the 14 linguistic indicators, achieving the highest accuracy of the three supervised learning algorithms tested, 93.9%, as shown in Table 6.5. This is a significant improvement over the baseline accuracy of 83.8%, as measured with a binomial test. Furthermore, a stative recall of 74.2% was achieved. The top portion of the tree created with recursive partitioning is shown in Figures 6.4.

To achieve this increase in classification performance, the decision tree narrowed down the training cases into relatively small partitions of verbs. Table 6.8 shows the distribution of training case verbs examined by the highlighted tree node in Figure 6.4. As seen by tracing the path from the root to the highlighted node, these are the verbs with `frequency` less than 2,013 across the corpus, and modified by *not* or *never* at least 3.48% of the time. From this subset, the highlighted node distinguishes the three verbs with `frequency` at least 314, shown in bold in Table 6.8, as **states**. This is correct for all 19 instances of these

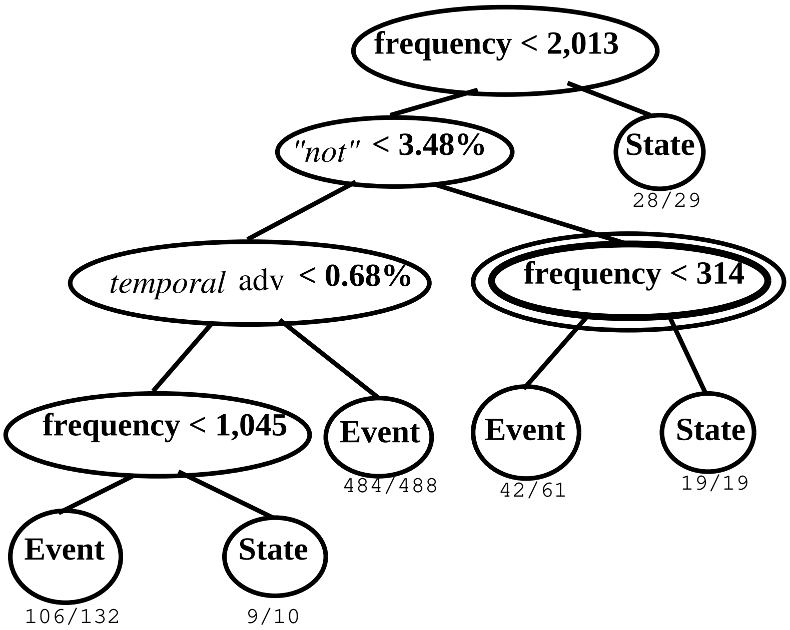


Figure 6.4: Top portion of decision tree automatically created to distinguish events from states. Leftward arcs are traversed when comparisons test true, rightward arcs when they test false. The values under each leaf indicate the number of correctly classified examples in the corresponding partition of training cases. The full tree has 59 nodes and achieves 93.9% accuracy.

Events:

get	3	talk	1	persue	1	drive	1
transfuse	2	suggest	1	provide	1	drink	1
respond	2	subside	1	pass	1	document	1
lose	2	sleep	1	notice	1	detect	1
live	2	retract	1	load	1	change	1
interfere	2	repair	1	limit	1	achieve	1
breathe	2	relieve	1	lift	1	regain	1
visualize	1	recognize	1	help	1		
tell	1	radiate	1	explain	1		

States:

feel	12	associate	3	want	1	concern	1
know	4	remember	2	support	1	account	1
appear	4	believe	2	indicate	1		
require	3	accompany	2	desire	1		

Table 6.8: Verb distribution at the decision tree node highlighted in Figure 6.4. The three stative verbs shown in bold are distinguished at this node, with no event verbs mis-classified.

```
1. *hate *like pull

2. lower raise

3. demand *know *love *want

4. buy sell

5. enter forget learn

6. acquire *need *require

7. leave push
```

Figure 6.5: Verb groupings created by an unsupervised learning algorithm developed and implemented by Hatzivassiloglou and McKeown [1993a] and Hatzivassiloglou [1997], applied over the corpus of 10 novels. Stative verbs are shown with an asterisk, and event verbs without.

three verbs, and does not misclassify any **event** verbs.

This inspection illustrates a benefit of discriminating verbs based on indicator values computed over large corpora. Most of the verbs in Table 6.8 appear in the training set a small number of times, so it would be difficult for a classification system to generate rules that apply to these individual *verbs*. Rather, since our system draws generalizations over the *indicator values* of verbs, it identifies stative verbs without misclassifying any of the event verbs shown.

Unsupervised Learning

Figure 6.5 shows the output of the unsupervised system. Seven groups were created, each with two to four verbs. The grouping algorithm used by this system is designed for data that is not as sparse with respect to the frequencies of verb-object pairs, e.g., data from a larger corpus. Thus, this partition is not representative of the full power of the approach, and a larger amount of data could improve it significantly.

However, the general, unsupervised clustering algorithm clearly discriminated event verbs from stative verbs. In Figure 6.5, stative verbs are shown with an asterisk; event verbs are shown without. Three of the groups are dominated by stative verbs, and the other four groups are composed entirely of event verbs. Each stative verb is found in a group with 70.2% states, averaged across the 7 stative verbs, and each event verb is found in a group with 82.6% events, averaged across the 12 event verbs. This is an improvement over an uninformed baseline system that randomly creates groups of 2 or more verbs each, which would achieve average precisions of 63.2% and 36.8%, respectively.

We can draw two important conclusions from this result. First, unsupervised learning

	overall accuracy	Culminated recall	Culminated precision	Non-Culminated recall	Non-Culminated precision
CART	74.0%	86.2%	76.0%	53.1%	69.0%
log-linear	70.5%	83.1%	73.6%	48.7%	62.5%
GP (4 runs)	68.6%	77.3%	74.2%	**53.6%**	57.8%
decision tree	68.5%	86.2%	70.6%	38.1%	61.4%
Baseline A	63.3%	100.0%	63.3%	0.0%	100.0%
Baseline B	49.0%	46.4%	63.3%	**53.6%**	36.7%

Table 6.9: Comparison of three learning methods, optimized for accuracy, and two performance baselines, distinguishing culminated from non-culminated events.

is a viable approach for aspectually classifying verbs according to particular classification schemes such as stativity. Second, co-occurrence distributions between the verb and direct object informs the aspectual classification of verbs. This is an additional source of information beyond the 14 linguistic indicators we combine with supervised learning.

6.3.2 Culminated vs. Non-Culminated Events

When distinguishing according to completedness, log-linear regression and CART successfully combined indicator values, improving classification accuracy over the baseline measure. As shown in Table 6.9, accuracy was 70.5% and 74.0%, respectively. A binomial test shows both accuracies to be a significant improvement over the 63.3% accuracy achieved by Baseline A. Although the accuracies attained by genetic programming and decision tree induction, 68.6% and 68.5% respectively, are also higher than that of Baseline A, based on a binomial test this is not significant. However, this may be due to our small test sample size. The relatively small differences in accuracy between the three methods also did not test as significant over our small sample size. Therefore, we cannot conclude as to their relative performance. A detailed evaluation of these classification performance results is given in Chapter 5.

The increase in the number of **non-culminated** clauses correctly classified, i.e. non-culminated recall, illustrates a greater improvement over the baseline. As detailed in Table 6.9, non-culminated recalls of 48.7%, 53.6% and 38.1% were achieved by the three learning methods, as compared to the 0.0% non-culminated recall achieved by Baseline A. Baseline A does not classify any non-culminated clauses correctly because it classifies all clauses as events. This difference in recall is more dramatic than the accuracy improvement because of the dominance of culminated clauses in the test set. Note that it is possible for an *uninformed* approach to achieve the same non-culminated recall as genetic programming, 53.6%, by arbitrarily classifying 53.6% of all clauses as **non-culminated**, and the rest as **culminated**. However, as shown in Table 6.9, the average performance of such a method (*Baseline B*) loses in comparison to genetic programming, e.g., in overall accuracy (49.0%) and non-culminated precision (36.7%).

A pilot study showed no further improvement in accuracy or recall tradeoff by four additional learning algorithms: Naive Bayes [Duda and Hart, 1973], Ripper [Cohen, 1995], ID3 [Quinlan, 1986] and C4.5 [Quinlan, 1993]. Furthermore, using metalearning to combine multiple learning methods hierachically (a.k.a., stacked generalization) [Chan and Stolfo, 1993, Wolpert, 1992], according to the JAM (Java Agents for Metalearning) model [Stolfo et al., 1997], produced only equivalent results. This may be due to the limited size of our supervised data.

All three supervised learning methods highly prioritized the *perfect* indicator. As shown in Table 5.2 (in Chapter 5), the *perfect* indicator tied with the *temporal* adverb indicator as most highly correlated with completedness, according to t-test results. And, as discussed in Section 5.4, the *perfect* indicator individually achieved the highest non-culminated recall of any indicator (52.2%) with no loss in overall accuracy. The induced decision tree (not shown) uses the *perfect* indicator as its first discriminator, log-linear regression ranked the *perfect* indicator as fourth out of fourteen (Table 6.10), and the function tree created by the genetic algorithm (Figure 6.6) includes the *perfect* indicator as one of five indicators used together to increase classification performance. Therefore, in all cases, automatic learning emphasized the *perfect* indicator, which, as discussed in Section 2.1.1, is strongly connected to completedness on a linguistic basis.

Learning methods rank differently depending on the task. The relative accuracies of the three supervised learning procedures rank in opposite orders when comparing the results in classification according to stativity (Table 6.5) to results in classification according to completedness (Table 6.9). Furthermore, when measuring classification performance as the recall of the non-dominant class (stative and non-culminated, respectively), the rankings are also conflicting when comparing results of the two classification tasks.

The same two linguistic indicators were ranked in the top two positions for both aspectual distinctions by log-linear modeling. As shown in Tables 6.10 and 6.6, which show the weights automatically derived by log-linear regression for each of the 14 linguistic indicators, the *manner* adverb and duration *in*-PP indicators are in the top two slots for both weighting schemes, corresponding to the two aspectual distinctions. These may indicate that these two indicators are universally useful for aspectual classification, at least when modeling with log-linear regression. However, the remaining rankings of linguistic indicators differ greatly between the two weighting schemes.

Genetic programming improved classification performance by emphasizing a different set of indicators than those emphasized by log-linear modeling. Figure 6.6 shows an example function tree automatically generated by the genetic algorithm, which achieved 69.2% accuracy. Note that, as for stativity, this classification performance was attained with a subset of only five linguistic indicators: no subject, `frequency`, *temporal* adverb, perfect, and not progressive. However, only two of these appeared in the example function tree shown for stativity (Figure 6.3): `frequency` and progressive.

```
(ADD (SUB (SUB (ADD (DIV NoSubject Frequency) TemporalAdv) (SUB
83 Perfect)) (DIV (DIV NoSubject Frequency) Frequency)) (SUB (ADD
NotProgressive NotProgressive) (SUB 60 Perfect)))
```

Figure 6.6: Example function tree designed by a genetic algorithm to distinguish between culminated and non-culminated verbs, achieving 69.2% accuracy and 62.8% non-culminated recall.

Linguistic Indicator	log-linear weight	t-test P-value
duration *in*-PP	-0.1207664	0.0626
manner adverb	0.03808262	0.0008
evaluation adverb	0.03212381	0.7063
perfect	-0.02304221	0.0000
temporal adverb	-0.01643347	0.0000
"not" or *"never"*	-0.01212703	0.6164
not progressive	-0.01059269	0.0031
no subject	-0.006891114	0.0241
past/pres participle	-0.004127672	0.0080
past tense	0.002484739	0.7132
present tense	0.00218274	0.0757
continuous adverb	-0.001775534	0.2563
duration *for*-PP	0.0001747421	0.1756
`frequency`	-0.0000916167	0.5652

Table 6.10: Weights produced by log-linear regression to distinguish between culminated and non-culminated verbs. Contrast with Table 6.6.

	training accuracy	testing accuracy
decision tree	96.3%	93.9%
genetic programming	93.4%	91.2%
log-linear	88.8%	86.7%
baseline	85.8%	83.8%

Table 6.11: Comparing training and test performance on three learning methods, distinguishing states from events.

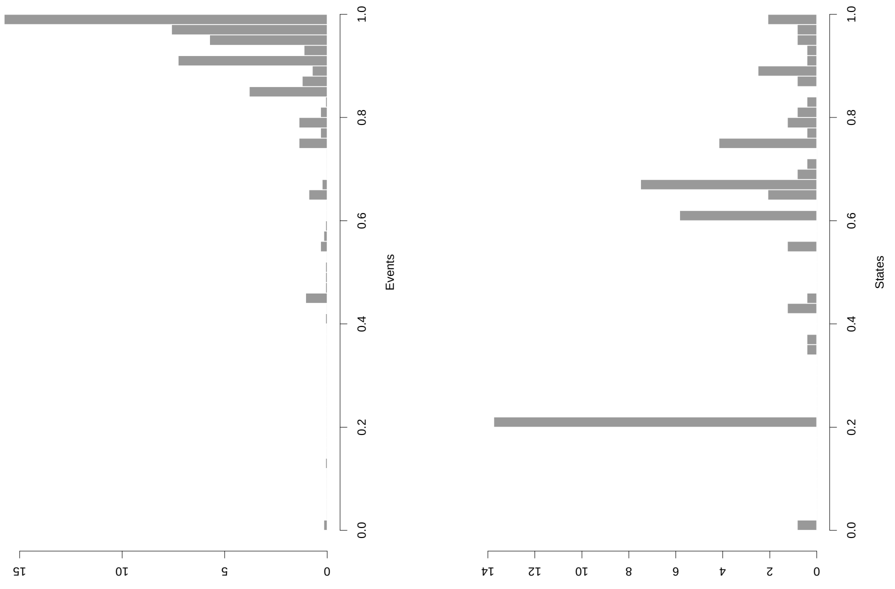


Figure 6.7: Histograms of log-linear output over the training data.

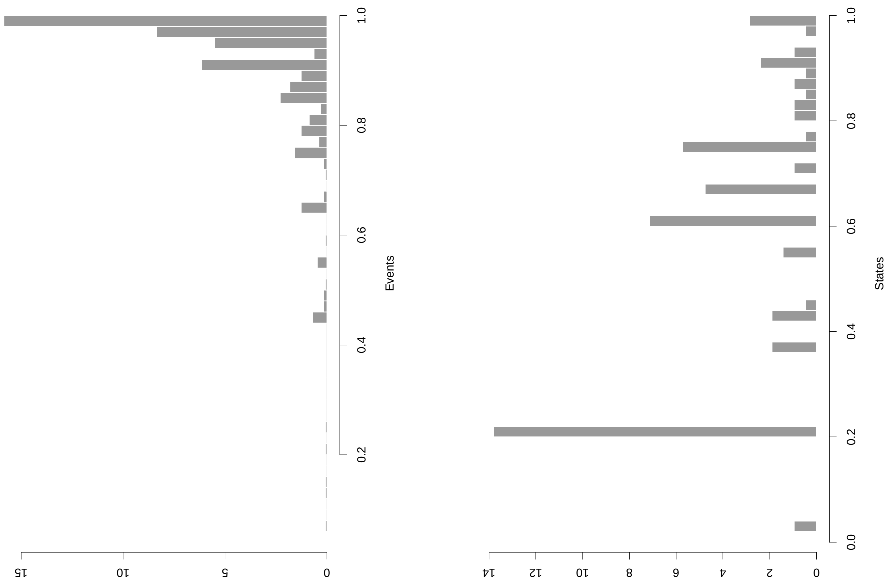


Figure 6.8: Histograms of log-linear output over the test data.

	training accuracy	testing accuracy
log-linear	79.4%	67.2%
decision tree	87.0%	59.1%
baseline	58.0%	63.3%

Table 6.12: Decision tree overlearns: comparison of training and test performance of two learning methods and a performance baseline, distinguishing culminated from non-culminated clauses, trained over a reduced set (of verbs), tested over the full test set (of verb instances).

6.3.3 Overfitting

When supervised learning methods are training, there is the risk that the resulting model's performance over the training set may not reflect its performance in general. If the model embodies general traits of all applicable data, it will *generalize* well. But if its success is due to discovering characteristics that are particular to the training data, its performance will be unpredictable over unseen data such as the test set. During training, there may a point after which increasing training performance will *hurt* testing performance [Weiss and Kulikowski, 1991]. In this section, to examine the extent of overlearning, we consider both the training and testing scores of the three supervised learning algorithms.

Overfitting was moderate for each of the three supervised learning algorithms when optimizing for classification accuracy in distinguishing between stative and event clauses. As shown in Table 6.11, each learning method's performance over the training data was about 2 points higher than that over the test data. This corresponds to a 2 point difference in baseline performance, which is due to a higher proportion of event clauses in the training data.

By examining the histograms in Figure 6.7, one can visualize the classification procedure embodied by log-linear modeling, and understand the potential for overfitting. Viewed sideways, this figure shows a histogram of the values produced by the model for each event clause in the training data, and another histogram below it (to the right, if you are not viewing sideways) corresponding to the stative clauses. For each clause, the model outputs a value by combining the indicator values that correspond to the main verb, as described in Section 6.2.1. Since output values larger than 0.5 are classified as **events**, and smaller values as **states**, it is visually clear that most values in the top histogram are classified correctly (**events**), and that a large amount in the bottom histogram are also classified correctly (**states**). This corresponds to the event and stative recalls of 96.9% and 34.2% listed in Table 6.5.

A second pair of histograms allows one to visualize the generalization performance of log-linear modeling. The log-linear model can be validated only if histograms over unseen data have similar distributions to the training histograms. Log-linear regression's improvement in classification performance over the training data is due to a scheme of combining indicator values such that the top histogram is heavy towards the right, and the bottom towards the left. If this combination scheme is idiosyncratic to the particular examples found in the training set, the distributions may be entirely different when applying the model to the test set, in which case classification performance would not generalize. However, as shown in Figure 6.8, the histograms computed over the test set appear very similar to those computed over the training set, and this is reflected by the successful generalize performance quantified in Table 6.11. Since the representation scheme embodied by log-linear modeling is limited in that it is based primarily on a linear combination, its ability to fine-tune to small subsets of training cases is limited, which alleviates its potential for overfitting.

Decision tree induction overlearned when training over a smaller set of examples. Table 6.12 shows the training and test performance of decision tree induction and log-linear modeling when training over the reduced training set introduced in Section 5.5.2. This smaller training set was created such that individual verbs do not appear multiple times, and no verb in the test set appears in the training set, leaving a total of 131 training cases. Despite these limitations, the model created by log-linear modeling achieved an accuracy 67.2% over the full set of 308 test cases, which is close to the performance achieved by training over the full training set (see Table 6.9). Although decision tree induction (no pruning) achieved higher accuracy than log-linear modeling over the training cases (87.0%), its testing score of 59.1% is below that of the baseline.

The thresholds established to discriminate the outputs of genetic programming's function trees generalize well to unseen data. When inducing these function trees with the genetic algorithm, the training set is used to form the tree, and to select a threshold which best discriminates between verbs of the different classes. There is the potential that a threshold determined over the training cases will not generalize well when evaluated over the test cases. To test this, for each of the seven function trees generated by the genetic algorithm to distinguish between states and events, the best threshold was selected over the *test* cases. For five of the function trees, there was no threshold that increased classification accuracy beyond that attained by the threshold established over the training cases. For the other two, a threshold was found that allowed for *one* more of the 739 test cases to be correctly classified.

6.4 Conclusions

Machine learning has served to illustrate the potential of the fourteen linguistic indicators by showing that they perform well in combination. This potential was not clear when evaluating indicators individually. The learning techniques showed improved classification performance with the same set of indicators for both verb classification problems. For stativity, decision tree induction achieved an accuracy of 93.9%, as compared to the uninformed baseline's accuracy of 83.8%. Furthermore, genetic programming and log-linear regression also achieved improvements over the baseline. For completedness, log-linear regression achieved an accuracy of 70.5%, as compared to the uninformed baseline's accuracy of 63.3%.

These improvements in classification performance are more dramatically illustrated by favorable tradeoffs between recall scores achieved by all three of these methods. These results are profitable for tasks that weigh the identification of the less frequent class more heavily.

Furthermore, automatic learning often derives linguistically informative insights. Inspecting the models produced by automatic learning revealed several such insights, including:

- Examining the log-linear model for classification according to stativity revealed a decision tree type of rule incorporated with the normal weighting scheme.

- Verb frequency distinguishes stative verbs at multiple levels of detail. When applied to all verbs, it identifies occurrences of *show*. Furthermore, examining an example node of the decision tree that distinguishes according to stativity revealed that verb frequency discriminates 19 stative clauses with 100.0% precision from a group of 60 training cases.

- Several proper subsets of the linguistic indicators prove independently useful for aspectual classification, when combined with an appropriate model. This is illustrated by the fact that certain models reveal combinations of small sets of indicators that improved classification performance. For example, genetic programming results for both classification tasks incorporated a subset of only five indicators each. In particular, *manner* adverb, which ranked highest by log-linear regression, is not incorporated in the example function tree induced by genetic programming. This may be because this indicator only applies to a small number of verbs, as shown in Table 6.7, and because an **if**-rule such as that captured by log-linear regression is difficult to encode with a function tree with no conditional primitives.

- Learning methods discovered that a small number of indicators are particularly useful for both classification tasks. For example, the same two indicators were weighted most heavily by log-linear regression for both tasks: duration *in*-PP and *manner* adverb.

- However, in general, learning methods emphasized different linguistic indicators for different classification tasks. For example, decision tree induction used `frequency` as the main discriminator to classify clauses according to stativity, while the *perfect* indicator was the main discriminator for classification according to completedness.

Comparing the ability of learning methods to combine linguistic indicators is difficult, since they rank differently depending on the classification task and evaluation criteria. For example, the relative accuracies of the three supervised learning procedures rank in opposite orders when comparing the results for stativity to results for completedness.

Analysis of performance over training cases reveals only a moderate level of overlearning for all three learning paradigms. Additionally, when training over a significantly reduced training set, log-linear modeling performed well, maintaining this moderate level of overfitting, although decision tree induction (no pruning) overfit this small sample to a detrimental degree.

The unsupervised grouping of verbs is a viable approach for aspectually classifying verbs according to particular classification schemes such as stativity. This result shows that co-occurrence distributions between the verb and direct object informs the aspectual classification of verbs. This is an additional source of information beyond the 14 linguistic indicators we combine with supervised learning. However, due to the sparcity intrinsic to an open class category such as the direct object, this approach was only evaluated over a small set of frequent verbs.

Chapter 7

Conclusions and Future Work

The ability to distinguish **stative** clauses from **event** clauses is a fundamental component of natural language understanding. These two high-level categories correspond to primitive distinctions in many domains, including, for example, the distinctions between *diagnoses* and *procedures* in the medical domain, and between *analyses* and *activity* in the financial domain.

Stativity is the first of three fundamental temporal distinctions that compose the *aspectual class* of a clause. Events are further distinguished along two other dimensions. First, *completedness* determines whether an event reaches a culmination or completion point in time at which a new state is introduced. Second, *atomicity* distinguishes *extended* events from non-extended (i.e. *atomic*) events. By dividing events along the second and third dimensions we derive four classes of events

Aspectual classification is a necessary prerequisite for interpreting certain adverbial adjuncts, as well as identifying temporal constraints between sentences in a discourse [Moens and Steedman, 1988, Dorr, 1992, Klavans, 1994]. It is also crucial for lexical choice and tense selection in machine translation [Moens and Steedman, 1988, Klavans and Chodorow, 1992, Klavans, 1994, Dorr, 1992]. In general, aspectual classification is a necessary component for applications that perform certain natural language interpretation, natural language generation, summarization, information retrieval and machine translation tasks.

Language interpretation models that incorporate aspect rely on the ability to automatically find the aspectual category of a clause [Hwang and Schubert, 1991, Schubert and Hwang, 1990, Dorr, 1992, Passonneau, 1988, Moens and Steedman, 1988, Hitzeman et al., 1994]. An understanding system can recognize the aspectual transformations that have effected a clause and can infer certain temporal constraints and entailments only after establishing the clause's *fundamental* aspectual category, which is a function of the main verb and several other constituents. The number of ways in which clausal constituents interactively influence fundamental aspectual class is unknown. However, syntax alone is not sufficient, and the lexical head of multiple constituents (e.g., the verb phrase and the direct object) can be influential factors [Pustejovsky, 1991].

Naturally occurring text contains vast amounts of implicit information pertaining to aspectual classification. For example, since each aspectual marker constrains the category of clauses it appears with, the frequency distributions of these markers reflect the categories of individual verbs.

We have developed a system for aspectual classification using linguistic indicators. The values for these indicators are computed automatically across corpora of text. These indicators were evaluated over the unrestricted set of verbs occurring across two corpora of text. Machine learning was used to combine multiple indicators in order to improve classification performance and to gain linguistic knowledge about indicators and their interactions.

If future work is to attain 100% accuracy for aspectual classification, both statistical and linguistic information sources must be incorporated. Doing this requires a careful balance between the two, not only in terms of implementing the ultimate classification system, but also in terms of distributing research efforts. This very theme was examined by the natural language workshop, "The Balancing Act" [Klavans and Resnik, 1997].

This thesis has performed one iteration in this ongoing process. First, we are combining statistical and linguistic approaches with linguistic indicators, which measure the frequencies of markers that are linguistically related to aspectual classification. This use of statistics solves the linguistic "bootstrapping" problem of using markers to predict fundamental aspectual class: fundamental aspect must be detected to interpret aspectual coercions, but aspectual coercions must be detected to interpret the effect of an aspectual marker. As a second way of integrating the two, we have incorporated high-level, general semantic groupings of lexical items in several ways, as summarized below in Section 7.4.

In this Chapter, I summarize the contributions of this work and describe future directions for aspectual classification.

7.1 Linguistic Indicators for Classification

Classifying verbs and clauses according to numerical indicators has several important advantages. Additionally, there are several advantages to using *linguistic* indicators in particular, which measure the frequencies of linguistic markers rather than measuring cooccurrences with open class lexical items such as the head of the direct object. The first five of the advantages listed below pertain to numerical indicators in general, and the rest to linguistic indicators in particular:

- Once a system has been trained to distinguish by indicator value, it can automatically classify all verbs appearing in a corpus, including "unseen" verbs that were not included in the supervised training sample; when measuring indicators over multiple constituents, unseen clausal templates can be classified automatically.

- While training such a system, generalizations over sets of sparse verbs can be ascertained, since generalizations are drawn over indicator values. For example, in Section 6.3.1, we examined an internal node of a decision tree that positively identifies a set of stative verbs without misclassifying any of 27 event verbs that occur only one time each in the training set.

- Indicators can be automatically measured for multiple domains and languages.
- Several machine learning methods are designed specifically for the task of combining numerical indicators, including decision trees, log-linear regression, genetic programming and neural networks.
- The tradeoff between recall and precision can be selected for particular applications by selecting an appropriate classification threshold. For example, optimizing for F-measure demonstrated a more dramatic tradeoff in recall values as compared to those attained when optimizing for accuracy.
- This framework is expandable:
  - Additional indicators can be introduced by measuring the frequencies of additional aspectual markers, as described below in Section 7.6.
  - Indicators can also be measured over clausal constituents other than the main verb to alleviate verb ambiguity and sparsity, as described below in Section 7.3.
  - Alternate machine learning methods may further improve the process of combining multiple indicators.
- The difficulty with sparsity for measuring the cooccurrence frequency between two open class lexical items (e.g., most verb-object combinations are rare) is alleviated, since many markers such as the *perfect* are prevalent across all verbs.
- The linguistic constraints associated with aspectual markers are exploited without the a-priori need to detect aspectual coercion. This solves a "bootstrapping" problem, since you need the fundamental aspectual category of a clause to detect any coercion affected by an aspectual marker, as discussed in Section 2.1.7.
- Many linguistic markers are linked with particular semantic distinctions such as aspectual categorization, whereas the semantic distinctions reflected by lexical cooccurrence are less specific, and may correspond to the least common denominator over a broad set of semantic distinctions.
- Linguistic indicators constrained with respect to one particular aspectual distinction may also correlate with other aspectual distinctions, since aspectual distinctions are likely to be mutually correlated. For example, since **points** (e.g., *hiccup*) are rare, non-extended (i.e., **atomic**) **events** are likely to be **culminations**. Therefore, an indicator that discriminates **events** according to atomicity is likely to also discriminate between **culminated** and **non-culminated events**.
- Results with linguistic indicators may be less domain-dependent, since the linguistic constraints associated with markers relate to generally-applicable semantic distinctions such as aspectual classification. On the other hand, the relationship between aspectual class and the frequencies of open class lexical items is likely to vary between domains.

When evaluated over an unrestricted set of verbs, such linguistic indicators were shown to improve aspectual classification, both individually and in combination. Of the fourteen indicators examined, seven were shown to correlate significantly with stativity, and six with completedness. Furthermore, when combined with a decision tree, the indicators achieve an accuracy of 93.9% for classification according to stativity, in comparison to 83.8% achieved by the baseline. This improvement in classification performance is even more dramatically illustrated by the stative recall of 74.2% attained, which is profitable for tasks that weigh the identification of states more heavily than events. When combined with log-linear regression, indicators achieved an accuracy of 70.5% for classification according to completedness, in comparison to 63.3% achieved by the baseline. Additionally, with genetic programming a non-culminated F-measure of 61.9 was achieved, improving over the 53.7 attained by an uninformed baseline.

We can conclude that linguistic indicators successfully exploit linguistic insights to provide a much-needed method for aspectual classification. Furthermore, these results establish lower bound performances for the automatic classification of clauses headed by an unrestricted set of verbs for two aspectual distinctions: stativity and completedness.

7.1.1 Single Indicators Help Multiple Classification Problems

The linguistic indicators evaluated in this thesis are applicable over two aspectual classification problems. One reason for this is that the two classification problems are linguistically related. As a result, some aspectual markers that constrain one aspectual distinction also correlate with another distinction. For example, the *progressive* indicator, which has been linguistically related to the distinction between extended and atomic (i.e., punctual) events, is indirectly linked to completedness. Its usefulness for predicting whether a verb is culminated or non-culminated may be due to the fact that atomic verbs are most likely to be culminated than non-culminated. As another example, some indicators are not based on aspectual constraints, e.g., *not/never* and duration *for*-PP. Such indicators may correlate with multiple aspectual distinctions because of relationships between their temporal entailments and aspectual category in general.

This interapplicability has been demonstrated empirically. Our analysis revealed correlations between stativity and five indicators that are not traditionally linked to stativity in the linguistics literature. Four were not previously hypothesized to predict any aspectual distinction: (1) Verb `frequency`, (2) occurrences modified by *"not"* or *"never"*, (3) occurrences with no deep subject, and (4) occurrences in the past or present participle. Furthermore, the *perfect* tense was not previously hypothesized to predict stativity in particular.

We have also discovered correlations between completedness and four indicators not traditionally linked to completedness on a linguistic basis. Three were not previously hypothesized to predict any aspectual distinction: (1) occurrences modified by a *manner* adverb, (2) occurrences in the past or present participle, and (3) occurrences with no deep subject. Furthermore, occurrences in the progressive voice was not previously hypothesized to predict completedness in particular.

Some indicators are particularly valuable for multiple aspectual distinctions. For example, log-linear modeling ranked the same linguistic indicators in the top two positions both for distinguishing between culminated and non-culminated events and for distinguishing between states and events: duration *in*-PP and *manner* adverb. This supports our claim that some linguistic indicators are useful for all aspectual classification problems.

Applying the same set of indicators to both aspectual distinctions has also allowed us to recognize statistical differences between these classification problems. For example, the rankings of linguistic indicators created by log-linear regression for the two classification tasks differ greatly across all but two indicators (duration *in*-PP and *manner* adverb).

7.1.2 Identifying New Linguistic Indicators

We have discovered several new linguistic indicators that inform aspectual classification. Two of these indicators measure linguistic phenomena that are constrained by aspectual class: *evaluation* adverbs, and *no deep subject*. These aspectual constraints are not traditionally identified in linguistic analyses of aspect. The linguistic motivation for these were described Chapter 2.

Furthermore, four indicators measure linguistic phenomena not traditionally conjectured to correlate with aspectual category: *not/never*, present tense, past/present participle, and verb `frequency`.

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."*

This pragmatic correlation of the *not/never* indicator, as well as that of verb `fequency`, are particularly domain-specific. However, domain-specificity reflects a limitation of our results, not of our approach. Although these correlations may not translate to other domains, they demostrate that idiosyncracies of individual domains reflect aspectual class. Statistical corpus analysis over other domains may reveal other correlations with aspectual class.

7.1.3 Evaluating Over Multiple Domains and Criteria

These results demonstrate the applicability of linguistic indicators over multiple domains. Indicators were computed and evaluated over both a corpus of medical discharge summarizes as well as a corpus of novels. On both corpora, the indicators correlated with aspectual class and improved classification performance. This shows that linguistic indicators help a classification system account for the particular distributions of aspectual categories across verbs and phrases, which varies between domains.

Linguistic indicators are beneficial for the varying demands of various applications. In addition to training over a set of verb instances, the classification performance of linguistic indicators were evaluated (for distinguishing according to completedness) by training over a set of individual verbs. Resulting classification performance was favorable as measured over three types of test sets: the "normal" set of verb instances, a set of verbs, and a set of verbs that does not overlap with the training set of verbs. This evaluation was performed for multiple indicators combined by learning techniques (as described in the next subsection, below). This complements our evaluation of the linguistic indicators in several ways:

- Fewer supervised examples must be generated.
- Non-intersecting training and test verb sets are possible.
- Both training and testing may be more domain-independent, since verb frequencies vary between domains.
- The resulting lexicon can be evaluated over word sets.

7.1.4 Advantages Over a Manual Lexicon

In many portions of this work, clauses are classified by their main verb only. Therefore, 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 the baseline method described above.

However, classifying verbs and clauses according to numerical indicators has several important advantages:

- 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 [Siegel, 1999].
- 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 a corpus, since measuring linguistic indicators for a verb is fully automatic. This is particularly important, since the meaning of rare, idiosyncratic verbs may be highly domain-dependent.
- Indicators are successful even when training examples do not include any verbs from the test set. 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 completeness. The result are given in Section 5.5.2. Accuracy remained higher than the baseline, and the recall tradeoff is felicitous.

- While training such a system, generalizations over training sets of sparse verbs can be ascertained, since generalizations are drawn over indicator values. For example, one node of the resulting decision tree trained to distinguish according to stativity identifies 19 occurrences of stative verbs without misclassifying any of 27 event verbs that occur only one time each in the training set and are sorted to that decision tree node.

- Since indicators provide scalar values for each verb, the tradeoff between recall and precision can be selected for particular applications by selecting an appropriate classification threshold. For example, optimizing for F-measure resulted in a more dramatic tradeoff in recall values as compared to those attained when optimizing for accuracy. Moreover, such scalar values could allow the development of a system that combines multiple scalar measurements in order to derive the aspectual class of a clause [Klavans and Chodorow, 1992].

- This framework is expandable, since additional indicators can be introduced by measuring the frequencies of additional aspectual markers. Therefore, future work may further improve classification performance.

- Indicators measured over multiple clausal constituents, e.g., main verb-object pairs, alleviate verb ambiguity and sparsity and improve classification performance. Incorporating constituents other than the main verb is important for verbs that are highly ambiguous, e.g., *have*. The majority aspect method could not be applied to verb-object pairs because of sparsity.

- Although supervision was required to build a model that combines indicators, the success of this model provides an existence proof that unsupervised linguistic indicators can successfully classify clauses according to aspect. It may be possible to develop other methods for combining indicators. For example, an unsupervised corpus-based technique successfully classified frequent verbs according to stativity.

- The linguistic constraints associated with aspectual markers are exploited for identifying a clause's fundamental aspectual class without the a-priori need to detect aspectual coercion. This solves a "bootstrapping" problem, since the fundamental aspectual category of a clause is needed in order to detect any coercion affected by an aspectual marker [Moens and Steedman, 1988].

7.2 Combining Linguistic Indicators with Learning

While individual indicators have predictive value, they are predictively incomplete. This incompleteness is illustrated by observing that some indicators help for only a subset of verbs, a phenomenon visualized as *outliers* on histograms. Such incompleteness is due in part to sparsity and noise when computing indicator values over a corpus of limited size, and is also a consequence of the linguistic behavior of certain indicators, as described in Section 2.1.7. Incomplete indicators can complement one another when placed in combination.

Machine learning methods can be employed to automatically generate a model that will combine indicator values, thus taking full advantage of the complete range of indicators. There are four advantages to automating this process with machine learning:

- The cost of *manually* generating a model, which is often prohibitive, is avoided.
- Biases introduced by a human engineer may be suboptimal.
- Automated approaches are extensible to multiple natural language classification problems, across multiple domains and multiple languages.
- Resulting models often reveal linguistic insights.

These methods successfully increased classification performance and provided linguistic insights when applied over relatively small training sets. Since only a small quantity of supervised data is needed, this approach is easily portable to other domains, languages and semantic distinctions.

7.2.1 Increasing Classification Performance with Learning

Machine learning has served to illustrate the potential of the fourteen linguistic indicators by showing that they perform well in combination. This potential was not clear when evaluating indicators individually. The learning techniques showed improved classification performance with the same set of indicators for both verb classification problems. For stativity, decision tree induction achieved an accuracy of 93.9%, as compared to the uninformed baseline's accuracy of 83.8%. Furthermore, genetic programming and log-linear regression also achieved improvements over the baseline. For completedness, log-linear regression achieved an accuracy of 70.5%, as compared to the uninformed baseline's accuracy of 63.3%.

These improvements in classification performance are more dramatically illustrated by favorable tradeoffs between recall scores achieved by all three of these methods. These results are profitable for tasks that weigh the identification of the less frequent class more heavily.

Comparing learning methods is difficult, since they rank differently depending on classification task and evaluation criteria. For example, when measuring classification performance as the recall of the less frequent class (stative and non-culminated, respectively), the relative performance of the three supervised learning procedures rank in opposite orders when comparing the results in classification according to stativity to results in classification according to completedness.

Analysis of performance over training cases reveals only a moderate level of overlearning for all three learning paradigms. Additionally, when training over a significantly reduced training set, log-linear modeling performed well, maintaining this moderate level of overfitting, although decision tree induction (no pruning) overfit this small sample to a detrimental degree.

7.2.2 Extracting Linguistic Insights from Learned Models

Inspecting the models produced by automatic learning provides several linguistic insights:

- Examining the log-linear model for classification according to stativity revealed a decision tree type of rule incorporated with the normal weighting scheme.
- Verb frequency distinguishes stative verbs at multiple levels of detail. When applied to all verbs, it identifies occurrences of *show*. Furthermore, examining an example node of the decision tree that distinguishes according to stativity revealed that verb frequency discriminates 19 stative clauses with 100.0% precision from a group of 60 training cases.
- Several proper subsets of the linguistic indicators prove independently useful for aspectual classification, when combined with an appropriate model. This is illustrated by the fact that certain models reveal combinations of small sets of indicators that improved classification performance. For example, genetic programming results for both classification tasks incorporated a subset of only five indicators each. In particular, *manner* adverb, which ranked highest by log-linear regression, is not incorporated in the example function tree induced by genetic programming. This may be because this indicator only applies to a small number of verbs, as shown in Table 6.7, and because an **if**-rule such as that captured by log-linear regression is difficult to encode with a function tree with no conditional primitives.
- Learning methods discovered that a small number of indicators are particularly useful for both classification tasks. For example, the same two indicators were weighted most heavily by log-linear regression for both tasks: duration *in*-PP and *manner* adverb.
- However, in general, learning methods emphasized different linguistic indicators for different classification tasks. For example, decision tree induction used `frequency` as the main discriminator to classify clauses according to stativity, while the *perfect* indicator was the main discriminator for classification according to completedness.

7.3 Measuring Indicators over Multiple Clausal Constituents

The 14 indicators measured thus far over the main verb can alternatively be measured over other clausal constituents, e.g., the direct object. In particular, the indicator values corresponding to an input clause are computed by measuring the frequency with which the clause's direct object occurs with each indicating phenomenon, e.g., the perfect tense, across the corpus of parsed clauses.

Incorporating other clausal constituents in addition to the main verb contributes to aspectual classification in two respects. First, clauses with rare or infrequent verbs pose the difficulty that there may be too few occurrences over which to measure linguistic indicators. In this case, indicators measured over other clausal features may be more reliable, since several features besides the main verb correlate with aspectual class. Second, since many verbs are aspectually ambiguous, it is often necessary to account for constituents other than the main verb. One approach to this end is to measure indicator values across multiple constituents.

Our evaluation confirms the viability of measuring linguistic indicators over multiple clausal constituents. In particular, for clauses with rare verbs, indicators measured over the direct object outperform indicators measured over the main verb. When optimizing for non-culminated F-measure with genetic programming, direct object indicators achieve a greater tradeoff in recall when distinguishing between culminated and non-culminated events. These results provide evidence that classifying clauses with infrequent verbs according to indicator values corresponding to the direct object improves over an uninformed approach, and is a more judicious approach than classifying according to indicator values computed over the main verb.

Furthermore, our results provide evidence that multiple constituents help resolve the difficulty of classifying clauses headed by aspectually ambiguous verbs. Although these indicators are measured over less instances, the classification performance achieved by indicators measured over multiple constituents is comparable to that of indicators measured over the verb only (for all three learning methods), with some evidence of a greater tradeoff in recall.

7.4 Improving Aspectual Classification with Semantic Groupings

Incorporating linguistically-motivated semantic groupings of words is a necessary complement to our statistical approach. These groupings inform aspectual classification and alleviate sparsity problems. In particular, the high-level semantic category of clausal constituents such as the main verb and direct object correlate with the aspectual category of a clause, and place open class lexical items such as verbs and nouns into a controlled number of possible semantic categories. Semantic groupings can be statistically derived [Hatzivassiloglou and McKeown, 1993b, Schűtze, 1992], or manually designed as in, e.g., Wordnet [Miller et al., 1993].

We have demonstrated the utility of semantic groupings for aspectual classification in several ways. Wordnet groupings were used to help construct the four groups of adverbs used to define and measure four of the linguistic indicators (`temporal, continuous, manner` and `evaluation`). Wordnet groupings were also used to construct the set of nominals describing temporal durations used to measure the duration *in-* and *for*-PP indicators. Four additional uses are described below, in this section.

7.4.1 Classifying *have*-Clauses

A semantic category associated with the direct object of *have* can be used to predict the aspectual class of a *have*-clause. We demonstrated a correlation between the Wordnet category of a direct object, and the stative class of the *have*-clauses that it appears in. Furthermore, a simple rule for classifying *have*-clauses based on the Wordnet category of the direct object was manually derived using linguistic observations. This classification rule using Wordnet achieved an accuracy of 79.6% as compared to the baseline of 69.9%. Furthermore, an event recall of 67.7% was achieved by the classification rule, as compared to the 0.0% event recall achieved by the baseline. This improvement in classification performance demonstrates the significance of the direct object in determining the stative class. Furthermore, we have shown that a high-level semantic categorization of the direct object can be used to predict the class of a clause.

7.4.2 Measuring Indicators over Semantic Categories

Measuring indicators over the Wordnet class of the direct object in addition to the main verb improved the classification performance of indicators. For example, when combined with a decision tree for distinguishing according to completedness, indicators measured over the main verb and direct object class achieved a more favorable recall tradeoff than those measured over the verb only, with comparable performance in accuracy.

7.4.3 Grouping Verbs Statistically

A general, high-level semantic grouping of verbs informs aspectual classification. In particular, an unsupervised grouping of verbs based on co-occurrence distributions between the verb and direct object informs the aspectual classification of verbs. Each stative verb was automatically placed in a group with 70.2% states, and each event verb is found in a group with 82.6% events. This is an improvement over an uninformed baseline method that achieves average precisions of 63.2% and 36.8%, respectively.

7.4.4 Identifying Categories of Statively Ambiguous Verbs

Six semantic categories of statively ambiguous verbs were manually derived according to linguistic principles: `communication`, `cognition`, `perception`, `location`, metaphorical and carrier verbs. These categories are postulated to include all statively ambiguous verbs. Establishing a semantic hierarchy of aspectually ambiguous verbs will help predict a verb's subcategorization frames because of the correlation between subcategorization frame and semantic class. Subcategorization frames, in term, help disambiguate verbs based on their arguments.

7.5 Incorporating Aspect into Natural Language Applications

We have illustrated several ways in which aspectual constraints must be incorporated into natural language applications. In general, language understanding systems must use the aspectual categories of input clauses to make aspectual entailments and to infer temporal constraints, and a generation system must incorporate the aspectual constraints associated with the aspectual category of each clause it constructs. In particular, we illustrated the integration of aspectual classification into summarization, information retrieval, machine translation, processing medical reports, and augmenting specialized lexical databases such as the Medical Entities Dictionary.

7.6 Future Work

In this section, I described several specific future directions for aspectual classification, linguistic indicators, and applying machine learning to natural language classification.

7.6.1 Additional Linguistic Indicators

The performance of aspectual classification may improve by incorporating additional linguistic indicators. Several aspectual constraints other than the ones measured in this thesis are known to exist, some of which present difficulties for implementation at this stage. For example, clauses with an indefinite implicit object, e.g. *"We just ate ε,"* are constrained to be **events**. However, such a clause can only be detected with the help of a subcategorization lexicon. Also, *force* and *persuade* constrain their complement to be an **event**. However, these words appeared very infrequently with clausal complements in our parsed corpora.

Additional aspectual markers include *rate* adverbs, e.g., *quickly*, which can only modify **events**, and pseudo-clefts, e.g., *"What he did was push the cart,"* which are also constrained to modify **events**. Other durative prepositional phrases have the same aspectual constraints as durative *in*-PPs, such as, *"by Thursday."* The auxiliary *can* can only appear with (intentional) **events**. Finally, certain *boundless* adverbs can only modify states and processes, e.g., *forever* and *indefinitely.*

Cooccurrences with open class lexical items, e.g., the direct object, may also provide helpful indicators; unsupervised learning was able to place frequent verbs according to cooccurrence data into groups that correlate with aspectual class.

7.6.2 Measuring Indicators Dynamically

The process of measuring linguistic indicators over verbs can be generalized in several ways. For one, as preliminarily examined in this thesis, indicators can be measured over multiple clausal constituents in addition to the main verb. Secondly, occurrences of aspectual markers could be more heavily weighted among clauses that are more similar to the input clause than occurrences in clauses with, say, only a couple constituents in common, since the former have a higher probability of having the same fundamental aspectual category as the input clause. Thirdly, interactions exist between markers within a (similar) clause, e.g., *any* progressive clause can also appear in the perfect. These interactions should be detected and accounted for.

Machine learning techniques may be able to automatically determine how best to measure linguistic indicators, if trained over enough supervised data. For example, previous work has measured indicators by applying a symbolic expression induced by genetic programming to each similar clause in a corpus [Siegel and McKeown, 1996]. This way, interactions between markers in a similar clause can be accounted for. Even more ambitiously, machine learning techniques could further generalize these methods by automatically inducing an algorithm that scans a corpus dynamically, depending on the input clause and on what it sees as it processes similar clauses. This could allow dynamic implementations of clausal similarity measures, intra-clause linguistic interactions, and weighting occurrences of markers.

7.6.3 Further Semantic Distinctions

Since our evaluation has shown the success of linguistic indicators over two aspectual distinctions, they should be tested on the remaining aspectual distinction, *atomicity*, which distinguishes between extended and atomic events. It is possible that a similar improvement in classification performance over uninformed baselines may be achieved with the same set of 14 indicators.

Additionally, other clause-level semantic distinctions may be automated with linguistic indicators. One example is detecting the correct scope of quantifiers such as *many*. A similar scoping problem also applies to aspectual disambiguation – since there can be multiple levels of aspectual transformations within a clause, the scope of each transformation must be ascertained, as in interpreting,

"It took Mary three days to play the 1 minute waltz in 30 seconds." [Moens and Steedman, 1988]

Another candidate distinction for linguistic indicators is subject- versus object-raising verbs, e.g.,

Renée **promised** *Rachel to dance the Macarena.*

Renée **persuaded** *Rachel to dance the Macarena.*

Q: Who danced the Macarena?

7.6.4 Bilingual Corpora for Training

Parallel bilingual corpora are potential sources of supervised examples for training and testing aspectual classification systems. For example, since many languages have explicit markings corresponding to completedness (as described in Section 2.3.1), the category of a clause can be determined by examining its translation.

7.6.5 Disambiguating Clausal Constituents

The ambiguity of lexical items is reflected by the Wordnet database, which has multiple entries for ambiguous words. As described in Section 3.4.1, in this work we always used only the first sense for ambiguity. This allowed us to take a first pass at implementing Wordnet senses for aspectual classification without the need to implement lexical disambiguation. However, this approach introduced difficulties dealing with ambiguous words, for example, when aspectually classifying *have*-clauses. For example, *"The patient had Medicaid"* is a **state**, but Wordnet classifies *Medicaid* as an `act` (which normally corresponds with **event**). This of course could be solved via sensitivity to capitalization. As a reverse example, *"The patient had an enema"* is an **event**, but Wordnet lists *enema* as `artifact` before `act`. As a final example, *"The patient had a urine culture"* is an **event**, but Wordnet's first sense of *culture* is `group`.

These problems would be alleviated by disambiguating terms with respect to Wordnet senses (Judith L. Klavans, personal communication), e.g., as implemented by Resnik [1995].

7.6.6 Further Linguistic Analysis

The grouping of statively ambiguous verbs shown in Section 2.1.5 represents the high-level distinctions of a semantic hierarchy of aspectually ambiguous verbs. Further linguistic analysis may provide more detailed semantic groupings that have clear relationships with subcategorization frames which in turn help predict the aspectual class of a clause, thus detecting the compositional semantics within a clause [Pustejovsky, 1991].

Iterative improvements on aspectual classification via statistical and linguistic analyses will allow researchers to converge on the appropriate balance between the two [Klavans and Resnik, 1997].

7.6.7 Advancing Machine Learning

Additional machine learning methods should be tested for combining linguistic indicators. For example, neural networks are especially suited for combining numerical inputs. Additionally, several machine learning methods could augment our current approach. For example, cross validation may help make the most out of a small supervised sample. Also, iteratively refining the model, e.g., for log-linear regression, may be an important way to eliminate indicators that do not help for a particular classification problem, and eliminate redundancy between indicators that correlate highly with one another.

Appendix A

Appendix: Applying Evolutionary Computation to Natural Language Classification

The use of machine learning paradigms to automate natural language classification is a relatively new technique. The training data for all of these approaches either comes from artificially constructed problems used to benchmark the GA, or empirically collected data which has been marked manually. For updates on this list, please contact the author.

• Mauldin uses a GA to learn scripts for information retrieval [Mauldin, 1989]. Before the GA starts, a number of sentences are retrieved via keyword search for a particular query. The accuracy of each retrieved sentence is then manually determined by the user. These manually marked sentences then serve as training data.

The fitness of a script is based on the precision and recall with which it performs over the training data. Therefore, a script is being learned for a particular query. Special genetic operators are used to yield syntactically valid new scripts.

• Masand uses genetic programming to compute confidence measures for the classifications of news articles performed by Memory Based Reasoning (MBR) (a k-nearest neighbor method) [Masand, 1994]. In particular, arithmetic expressions are evolved which have access to confidence measures output by the MBR system. The top several confidence measures are made available as terminals to the evolved arithmetic expressions. The value computed by the evolved expression is then used as an overall confidence measure.

The confidence measures given by evolved expressions are used to channel the small percentage of inaccurately classified articles to human beings for classification. Training data is readily available since thousands of news articles are manually classified every day.

• Efforts by Huijsen and Lankhorst both use a GA to induce context free grammars [Huijsen, 1993] [Lankhorst, 1994]. Lankhorst uses a bitstring representation of a grammar, with the canonical bitstring genetic operators, while Huijsen uses a higher level representation, with syntactically-specialized genetic operators.

Huijsen trains his grammars over data produced from simple target grammars, including palindromes and a very simple fragment of natural language. Lankhorst's approach is similar, but uses a slightly more complex "Micro-NL" language. In both cases, the manually created target grammars are used to *generate* the training data. The purpose of this is to benchmark grammar induction without the complexities and high cost of empirical data.

• Nettleton and Garigliano use a GA to tune parameters which guide a natural language dialogue system [Nettleton and Garigliano, 1994]. Fitness is determined by measuring how closely the system can reproduce the responses in a manually produced target dialogue. No evaluation of the general accuracy of learned parameters is performed.

• Nordin genetically induces machine code to classify Swedish words into nouns and non-nouns by their spelling [Nordin, 1994]. The machine code, which is represented as a bitstring, has no method for recursion. It has access to certain mathematical operators, which can operate on a numerical represented of an input word's spelling. No analysis of how resulting programs work has been performed.

Bibliography

F. Allen and R. Karjalainen. Using genetic algorithms to find technical trading rules. Technical report, Rodney L. White Center For Financial Research, 1995. unpublished manuscript.

J. Allen. *Natural Language Understanding*. Benjamin/Cummings, Redwood City, CA, 1995.

I. Androutsopoulos, G.D. Ritchie, and P. Thanisch. Time, tense and aspect in natural language database interfaces. *Journal of Natural Language Engineering*, 1998.

R.J. Baker and J.A. Nelder. *The GLIM System, Release 3: Generalized Linear Interactive Modeling*. Numerical Algorithms Group, Oxford, 1989.

R.I. Binnick. *Time and the Verb: A Guide to Tense and Aspect*. Oxford University Press, New York, 1991.

E. Black. An experiment in computational discrimination of english word senses. *IBM Journal of Research and Development*, 2(32), 1988.

L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, 1984.

C.E. Brodley. Applying classification algorithms in practice. 1995.

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

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

K. Church. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the 2nd Conference for Applied Natural Language Processing*, pages 136–143, 1988.

J.J. Cimino and P.D. Clayton. Coping with changing controlled vocabularies. *Journal of the American Medical Informatics Association*, 1994.

J.J. Cimino, P.D. Clayton, G. Hripcsac, and S.B. Johnson. Knowledge-based approaches to the maintenance of a large controlled medical terminology. *Journal of the American Medical Informatics Association*, 1(1), 1994.

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

N. Cramer. 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, 1985.

B.J. Dorr. 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, 1992.

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

D.R. Dowty. The effects of aspectual class on the temporal structure of discourse: semantics or pragmatics? *Linguistics and Philosophy*, 9(1):37–61, 1986.

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

W. DuMouchel, C. Friedman, Hripcsak G., Johnson S.B., and P. Clayton. Two applications of statistical modelling to natural language processing. In *Draft Proceedings, Fifth International Workshop on Artificial Intelligence and Statistics*, 1995.

U.M. Fayyad and K.B. Irani. On the handling of continuous-valued attributes in decision tree generation. *Machine Learning*, (8), 1992.

C. Friedman, P.O. Alderson, J.H.M. Austin, J.J. Cimino, and S.B. Johnson. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1(2), 1994.

C. Friedman, G. Hripcsak, W. DuMouchel, S.B. Johnson, and P.D. Clayton. Natural language processing in an operational clinical information system. *Natural Language Engineering*, 2(1), 1995.

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

D. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Publishing Company, Inc., Reading, MA, 1989.

J. Grimm. *Teutonic Mythology*. Repr. Dover, New York, 1883.

B. Grosz and C. Sidner. Attention, intention, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.

V. Hatzivassiloglou. *Automatic acquisition of lexical semantic knowledge from large corpora: The identification of semantically related words, markedness, polarity, and antonymy.* PhD thesis, Columbia University, 1997.

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

V. Hatzivassiloglou and K.R. McKeown. Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. In *Proceedings of the 31st Annual Meeting of the ACL*, pages 172–182, Columbus, Ohio, 1993b. Association for Computational Linguistics.

V. Hatzivassiloglou and K.R. McKeown. A quantitative evaluation of linguistic tests for the automatic prediction of semantic markedness. In *Proceedings of the 33rd Annual Meeting of the ACL*, pages 197–204, Boston, Massachusetts, 1995. Association for Computational Linguistics.

J. Hitzeman, M. Moens, and C. Grover. Algorithms for analysing the temporal structure of discourse. Technical report, University of Edinburgh, 1994.

J. Holland. *Adaptation in Natural and Artificial Systems.* The University of Michigan Press, Ann Arbor, MI, 1975.

G. Hripcsak, C. Friedman, P.O. Alderson, W. Dumouchel, S.B. Johnson, and P.D. Clayton. Unlocking clinical data from narrative reports. Technical report, Department of Medical Informatics, Columbia-Presbyterian Medical Center, New York, 1995.

W. Huijsen. Genetic grammatical inference, 1993. Master's thesis.

C.H. Hwang and L.K. Schubert. Interpreting temporal adverbials. Technical report, University of Rochester, 1991.

S.J. Kaplan. Cooperative responses from a portable natural language data base query system. *Artificial Intelligence*, (19), 1982.

A. Kenny. *Action, Emotion, and Will.* Routledge and Kegan Paul, London, 1963.

B. Kessler, G. Nunberg, and H. Schütze. Automatic detection of text genre. Technical report, Xerox Palo Alto Research Center, 1997.

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

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

J.L. Klavans and P. Resnik. *The Balancing Act: Combining Symbolic and Statistical Approaches to Language.* MIT Press, Cambridge, MA, 1997.

K. Knight, forthcoming.

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

J.D. Lafferty and R.L. Mercer. Automatic word classification using features of spelling, forthcoming.

M.M. Lankhorst. Breeding grammars: Grammatical inference with a genetic algorithm. Computing Science Report CS-R 9401, Dept. of Computing Science, University of Groningen, Groningen, The Netherlands, 1994.

M.M. Lankhorst and R. Moddemeijer. Automatic word categorization: an information-theoretic approach. Technical report, Dept. of Computing Science, University of Groningen, Groningen, The Netherlands, 1992.

C. Leacock, G. Towell, and E. Voorhees. Corpus-based statistical sense resolution. In *Proceedings of the ARPA Workshop on Human Language*, 1993.

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

D.J. Litman. Classifying cue phrases in text and speech using machine learning. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, Menlo Park, CA, 1994. AAAI Press.

A.K. Luk. Statistical sense disambiguation with relatively small corpora using dictionary definitions. In *Proceedings of the 33rd Annual Meeting of the ACL*, Columbus, Ohio, 1995. Association for Computational Linguistics.

D.H. Magerman. Parsing as statistical pattern recognition. Technical report, IBM, 1993.

B. Masand. Optimizing confidence of text classification by evolution of symbolic expressions. In K. Kinnear, editor, *Advances in Genetic Programming.* MIT Press, Cambridge, MA, 1994.

M.L. Mauldin. *Information Retrieval by Text Skimming.* PhD thesis, Carnegie Mellon University, 1989.

M.C. McCord. SLOT GRAMMAR: A system for simpler construction of practical natural language grammars. In R. Studer, editor, *International Symposium on Natural Language and Logic.* Springer Verlag, 1990.

G.A. Miller, R. Beckwith, C. Felbaum, D. Gross, and K. Miller. Introduction to wordnet: An on-line lexical database. Technical report, 1993.

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

M. Nagao. A framework of a mechanical translation between japanese and english by analogy principle. In A. Elithorn and R. Banerji, editors, *Artificial and Human Intelligence*. North-Hollan, 1984.

A. Nakhimovsky. Aspect, aspectual class, and the temporal structure of narrative. *Computational Linguistics*, 14(2):29–43, 1988.

D.J. Nettleton and R. Garigliano. Dialogue optimisation in the lolita natural language processor using evolutionary algorithms. Technical report, University of Durham, 1994. unpublished manuscript.

P. Nordin. A compiling genetic programming system that directly manipulates the machine code. In K. Kinnear, editor, *Advances in Genetic Programming*. MIT Press, Cambridge, MA, 1994.

M.B. Olsen and P. Resnik. Implicit object constructions and the (in)transitivity continuum. In *Proceedings of the 33rd Regional Meeting of the Chicago Linguistics Society*, 1997.

T. Parsons. *Events in the Semantics of English*. The MIT Press, Cambridge, Massachusetts, 1990.

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

R.J. Passonneau and J.L. Klavans. Use of aspectual verbs in corpus analysis of lexical aspect. Technical report, Columbia University, 1994. unpublished manuscript.

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

F. Pereira, N. Tishby, and L. Lee. Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the ACL*, pages 183–190, Columbus, Ohio, 1993b. Association for Computational Linguistics.

D.M.W. Powers and C.C.R. Turk. *Machine Learning of Natural Language*. Springer-Verlag, 1989.

J. Pustejovsky. The syntax of event structure. *Cognition*, 41(103):47–92, 1991.

J. Pustejovsky. *The Generative Lexicon*. The MIT Press, Cambridge, Massachusetts, 1995.

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

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

P. Resnik. Semantic classes and syntactic ambiguity. In *Proceedings of the ARPA Workshop on Human Language Technology*, 1993.

P. Resnik. Disambiguating noun groupings with respect to WordNet senses. In *Third Workshop on Very Large Corpora*, 1995.

P. Resnik. Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, (61), 1996.

G. Ryle. *The Concept of Mind*. Hutchinson, London, 1949.

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

H. Schűtze. Dimensions of meaning. In *Proceedings of Supercomputing*, 1992.

L.K. Schubert and C.H. Hwang. Picking reference events from tense trees: a formal, implementable theory of english tense-aspect semantics. Technical report, University of Rochester, 1990.

E.V. Siegel. Competitively evolving decision trees against fixed training cases for natural language processing. In K. Kinnear, editor, *Advances in Genetic Programming*. MIT Press, Cambridge, MA, 1994.

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

E.V. Siegel. 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, 1998.

E.V. Siegel. Corpus-based linguistic indicators for aspectual classification. In *Submitted to the the 37th Annual Meeting of the Association for Computational Linguistics*, 1999.

E.V. Siegel and A.D. Chaffee. Genetically optimizing the speed of programs evolved to play tetris. In P.J. Angeline and K. Kinnear, editors, *Advances in Genetic Programming: Volume 2*, Cambridge, MA, 1996. MIT Press.

E.V. Siegel and J.R. Koza. Genetic programming: Papers from the 1995 fall symposium. Technical Report FS-95-01, AAAI, 1995.

E.V. Siegel and K.R. McKeown. Emergent linguistic rules from inducing decision trees: disambiguating discourse clue words. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, Menlo Park, CA, 1994. AAAI Press.

E.V. Siegel and K.R. McKeown. 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, 1996. Bilkent University.

F.A. Smadja and K.R. McKeown. Using collocations for language generation. *Computational Intelligence*, 7(4):229–239, 1991.

S. Stolfo, A.L. Prodromidis, S. Tselepis, W. Lee, and W. Fan. JAM: Java agents for meta-learning over distributed databases. Technical report, Columbia University, 1997. unpublished manuscript.

M.E. Stys. Parallel science texts in english and polish: Problems of language and communication. Technical report, Warsaw University, 1991. masters thesis.

G. Syswerda. Uniform crossover in genetic algorithms. In J.D. Schaffer, editor, *Proceedings of the Third International Conference on Genetic Algorithms*. Morgan Kaufmann, 1989.

W.A. Tackett. Genetic programming for feature discovery and image discrimination. In *Proceedings of the Fifth International Conference on Genetic Algorithms*, San Mateo, CA, 1993. Morgan Kaufmann.

W.A. Tackett and A. Carmi. The donut problem: scalability, generalization and breeding policies in genetic programmiong. In K. Kinnear, editor, *Advances in Genetic Programming*. MIT Press, Cambridge, MA, 1994.

B. Taylor. Tense and continuity. *Linguistics and Philosophy*, 1(2), 1977.

C. Tenny. The aspectual interface hypothesis. 1992.

Joerg P. Ueberla. An extended clustering algorithm for statistical language models. Technical Report DRA/CIS(CSE1)/RN94/13, Forum Technology - DRA Malvern, 1994.

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

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

S.M. Weiss and C.A. Kulikowski. *Computer Systems that Learn*. Morgan Kaufmann, San Mateo, CA, 1991.

J. Wiebe, T. O'Hara, K. McKeever, and T. Őhrstrőm-Sandgren. An empirical approach to temporal reference resolution. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, Providence, RI, 1997.

D.H. Wolpert. Stacked generalization. *Neural Networks*, 5, 1992.

D. Yarowsky. Decision lists for lexical ambiguity resolution: Application to accent restoration in spanish and french. In *Proceedings of the 32nd Annual Meeting of the ACL*, San Francisco, CA, 1994. Morgan Kaufmann.